Machine Learning Explanations for the boardroom

Author:
Azimeh Gharavi

Supervisor:
Prof. Mike Chantler

Second reader:
Prof. Sven-Bodo Scholz

A thesis submitted in fulfilment of the requirements
for the degree of MSc.

in the

School of Mathematical and Computer Sciences

August 2017
Declaration of Authorship

I, Azimeh Gharavi, declare that this thesis titled, 'Machine Learning Explanations for the boardroom' and the work presented in it is my own. I confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: Azimeh Gharavi

Date: 18/08/2017
Abstract

Many machine learning algorithms produce stochastic output which will affect credibility of the them and users often resist basing their decisions on solutions from such algorithms not only because they don’t trust the results but also because they can’t defend their decision to other stakeholders. Therefore in order to increase the trust of users, firstly they need to be aware of this fact which often they don’t get informed of, and secondly they need to be given confidence of choosing and accepting one solution to go ahead with.

The assumption of the project specification provided by supervisor was that the similarity of topics was more important than the quality of the topics, this assumption steered me towards the relationships between topics rather than the quality of topics. To visualise these relationships I used hierarchical agglomerative clustering on topics and ordered not only topic clusters but also models based on how similar they are to all other models and clusters.

To best of my knowledge TopicCheck [Chuang et al., 2015] is the only scientific paper which addresses the issue of visualising stochastic results of topic modelling. Because of lack of enough resources to identify best layout I have done in-depth evolution of layouts before choosing simple tabular layout. Simplicity of layout has three advantages in terms of my project:

1. People will be more confident when they can easily understand the layout and the relationship between object on interface,
2. By using simple interface the cognitive load on user will be reduced,
3. This layout is very scale-able in terms of both increasing the number of models and topics.

I have conducted in-depth pilot experiment and focus group to evaluate my project. Outcomes of these two experiments which are of course subject to usual caveats of focus group such as small number of participants are two-fold:

1. To provide recommendations of designing user interface with aim of visualising non-deterministic solutions,
2. To gather preliminary data to highlight some open-questions for future work.
Acknowledgements

I would like to thank my supervisor Prof. Milke Chantler who gave me the opportunity to do this great project.

Secondly I would also like to thank Mahmoud & Aran.
## Contents

**Declaration of Authorship**

**Abstract**

**Acknowledgements**

**Contents**

**List of Figures**

**List of Tables**

1 Introduction
   1.1 Project motivations ................................................. 1
   1.2 Criteria for literature survey ..................................... 4

2 Literature review
   2.1 Introduction ......................................................... 5
   2.2 Topic Modelling ...................................................... 5
      2.2.1 Latent Dirichlet Allocation (LDA) and extensions of LDA .... 5
      2.2.2 Interactive topic modelling (ITM) ............................ 7
      2.2.3 Supervised topic models ...................................... 7
      2.2.4 Summary of topic modelling .................................. 8
   2.3 Topic modelling frameworks ....................................... 9
      2.3.1 Summary of frameworks ................................... 10
   2.4 Topic modelling applications and datasets ....................... 11
      2.4.1 Text mining in digital humanities .............................. 11
      2.4.2 Bio-informatics and Medical science .......................... 11
      2.4.3 Natural language processing ................................ 12
      2.4.4 Opinion mining and short text clustering using social media data ... 13
      2.4.5 Describing research portfolios ............................... 13
      2.4.6 Candidate dataset .......................................... 14
      2.4.7 Summary of datasets ...................................... 17
   2.5 Clustering algorithms ............................................... 18
      2.5.1 Introduction .................................................. 18
      2.5.2 Hierarchical clustering .................................... 19
2.5.3 Hierarchical clustering with topic modelling .................................................. 19
2.5.4 Summary of clustering algorithms ................................................................. 21
2.6 Topic modelling visualisation ............................................................................. 21
  2.6.1 Techniques - Corpus level ............................................................................. 21
    2.6.1.1 Network graphs ................................................................................. 21
    2.6.1.2 Stacked barcharts ............................................................................ 23
    2.6.1.3 Dimensionality reduction using PCA ................................................. 24
    2.6.1.4 Bubbles ......................................................................................... 24
    2.6.1.5 Heatmaps - similarity matrix .............................................................. 25
    2.6.1.6 Hierarchical clustering - dendrograms and trees ............................... 26
    2.6.1.7 Hierarchical clustering - packs and treemaps .................................. 27
    2.6.1.8 Hex maps ...................................................................................... 28
    2.6.1.9 Visualising non-determinist aspect of topic modelling ...................... 28
  2.6.2 Techniques - Topic level ................................................................................. 30
    2.6.2.1 Word list ....................................................................................... 30
    2.6.2.2 Word cloud or tag cloud ................................................................. 30
  2.6.3 Techniques - Document level ........................................................................ 31
  2.6.4 Visualisation Tools ....................................................................................... 32
  2.6.5 Summary of visualisation .............................................................................. 32
2.7 Evaluation methods .......................................................................................... 32
  2.7.1 Quantitative and Qualitative Experiment .................................................... 32
    2.7.1.1 Quantitative experiment .................................................................... 33
    2.7.1.2 Qualitative experiment ..................................................................... 33
  2.7.2 Topic modelling evaluation methods ............................................................ 34
  2.7.3 Summary of paper "Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic Labels" .................................................. 35
  2.7.4 Summary of Evaluation .............................................................................. 37

3 Requirements analysis ......................................................................................... 38
  3.1 Introduction ..................................................................................................... 38
  3.2 Project scope .................................................................................................. 38
  3.3 Research hypotheses ...................................................................................... 39
  3.4 Project requirements ...................................................................................... 39

4 Project implementation ......................................................................................... 43
  4.1 Introduction ..................................................................................................... 43
  4.2 Pre-processing data ....................................................................................... 45
    4.2.1 Handling missing / short abstracts ......................................................... 45
    4.2.2 Reducing the size of the dataset ............................................................ 46
    4.2.3 Making data ready for topic modelling with MALLET .......................... 46
  4.3 Topic modelling ............................................................................................. 47
    4.3.1 Latent Dirichlet Allocation (LDA) .......................................................... 48
    4.3.2 Import documents ............................................................................... 48
    4.3.3 Train topics ...................................................................................... 49
  4.4 Post-processing data ...................................................................................... 50
    4.4.1 Similarity matrix ................................................................................. 50
4.4.2 Sorting topic clusters and models ............................................ 53
4.5 Web interface implementation .................................................. 54
  4.5.1 Prototypes of interface ..................................................... 54
    4.5.1.1 Dendrogram ......................................................... 55
    4.5.1.2 Hexagons Layout .................................................. 56
    4.5.1.3 Tabular layout version 1 ......................................... 58
    4.5.1.4 Tabular layout - version 2 ..................................... 60
4.6 Evaluation .............................................................................. 62
  4.6.1 Designing the experiment ................................................... 63
    4.6.1.1 Main steps during experiment .................................... 63
    4.6.1.2 Consent Forms ...................................................... 63
    4.6.1.3 Presentation of project ............................................ 63
    4.6.1.4 Demonstration of interface ...................................... 63
    4.6.1.5 Questionnaire - part 1 ............................................ 66
    4.6.1.6 Group discussions .................................................. 68
    4.6.1.7 Questionnaire- part 2 ............................................. 70
  4.6.2 Recruiting participants ..................................................... 70
  4.6.3 Pilot experiment .................................................................. 70
  4.6.4 Focus group discussions ................................................... 71
4.7 Summary ................................................................................. 71

5 Results ...................................................................................... 72
  5.1 Introduction ............................................................................ 72
  5.2 Pilot experiment ..................................................................... 72
    5.2.1 Tabular layout ............................................................. 72
    5.2.2 Modifications applied on interface based on outcomes of pilot experiment .................................................. 74
    5.2.3 Hexagons layout ........................................................... 74
  5.3 Focus group .......................................................................... 75
    5.3.1 Selecting Best Model ....................................................... 75
      5.3.1.1 Process of selecting BEST model by participants .......... 76
      5.3.1.2 Comparison of selected BEST model before and after group discussion .............................................. 78
      5.3.1.3 Summary selecting BEST model ............................... 80
    5.3.2 Comparison of participants confidence on selected model before and after group discussion ......................... 80
    5.3.3 Visual features of interface .............................................. 82
    5.3.4 Layout ............................................................................ 82
    5.3.5 Font size of top words .................................................... 82
    5.3.6 Empty spaces ............................................................. 83
    5.3.7 Interaction ....................................................................... 83
  5.4 Summary of results .................................................................. 83
    5.4.1 Findings and open-questions for future work .................... 84
    5.4.2 guidelines for designing interface .................................... 85

6 Conclusion .................................................................................. 90
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A Pre-processing data</strong></td>
<td>92</td>
</tr>
<tr>
<td>A.1 Introduction</td>
<td>92</td>
</tr>
<tr>
<td>A.2 Handling missing or very short abstracts</td>
<td>92</td>
</tr>
<tr>
<td>A.3 Reducing size of dataset</td>
<td>93</td>
</tr>
<tr>
<td>A.4 Making data ready for Topic modelling</td>
<td>95</td>
</tr>
<tr>
<td><strong>B Topic Modelling with MALLET</strong></td>
<td>97</td>
</tr>
<tr>
<td>B.1 Introduction</td>
<td>97</td>
</tr>
<tr>
<td>B.2 Import data to MALLET</td>
<td>97</td>
</tr>
<tr>
<td>B.3 Train topics</td>
<td>99</td>
</tr>
<tr>
<td>B.4 Default stop words list</td>
<td>101</td>
</tr>
<tr>
<td><strong>C Post-Processing of Data</strong></td>
<td>105</td>
</tr>
<tr>
<td>C.1 Compute cosine similarity</td>
<td>105</td>
</tr>
<tr>
<td>C.1.1 SimilarityTest.java</td>
<td>106</td>
</tr>
<tr>
<td>C.1.2 ReadExcelFile.Java</td>
<td>107</td>
</tr>
<tr>
<td>C.1.3 CosineSimilarity.Java</td>
<td>108</td>
</tr>
<tr>
<td>C.1.4 PrintMatrix.Java</td>
<td>109</td>
</tr>
<tr>
<td><strong>D Interface prototypes and implementation</strong></td>
<td>110</td>
</tr>
<tr>
<td>D.0.1 Making results of Topic Modelling ready for visualisation</td>
<td>110</td>
</tr>
<tr>
<td>D.0.1.1 What files from topic modelling results I used? which information</td>
<td>110</td>
</tr>
<tr>
<td>D.0.1.2 Clustering Algorithm</td>
<td>110</td>
</tr>
<tr>
<td>D.0.2 Format of dataset used for visualisation</td>
<td>110</td>
</tr>
<tr>
<td>D.0.3 Evolution of layouts</td>
<td>111</td>
</tr>
<tr>
<td>D.0.4 Code of tabular layout</td>
<td>114</td>
</tr>
<tr>
<td>D.0.4.1 index.html</td>
<td>114</td>
</tr>
<tr>
<td>D.0.4.2 tabular.js</td>
<td>115</td>
</tr>
<tr>
<td>D.0.4.3 tooltip.js</td>
<td>120</td>
</tr>
<tr>
<td><strong>E Evaluation</strong></td>
<td>121</td>
</tr>
<tr>
<td>E.1 Advert for recruiting participants</td>
<td>121</td>
</tr>
<tr>
<td>E.2 Facilitator’s discussion guide</td>
<td>122</td>
</tr>
<tr>
<td>E.3 Presentation</td>
<td>126</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Main idea behind LDA: Each document contains a number of topics, where each topic is distribution of words. LDA un-cover these hidden topics using un-supervised probabilistic algorithms.</td>
</tr>
<tr>
<td>2.1</td>
<td>Distribution of grant funding over 16 years by Wellcome Trust</td>
</tr>
<tr>
<td>2.2</td>
<td>Distribution of grants over regions. Greater London received highest total amount of funding from Wellcome Trust over 16 years.</td>
</tr>
<tr>
<td>2.3</td>
<td>Field &quot;Sponsor&quot; has 64% missing values in the dataset. This field should be eliminated from dataset unless it gives critical information.</td>
</tr>
<tr>
<td>2.4</td>
<td>Overview of amount of fund awarded by Wellcome Trust during 16 years period.</td>
</tr>
<tr>
<td>2.5</td>
<td>Two types of hierarchical clustering algorithm: 1. agglomerative: each object is assigned to one cluster and and repeatedly combines two most similar clusters. 2. divisive: assigns all objects to one cluster and then divides it in two least similar clusters. This figure is from this website.</td>
</tr>
<tr>
<td>2.6</td>
<td>Interface of Topicheck [Chuang et al., 2015]: TopicCheck uses clustering algorithm on results of multiple runs of topic modelling in order to evaluate stability of topics. This matrix will be filled if all topics were stable across all runs.</td>
</tr>
<tr>
<td>2.7</td>
<td>Corpus view of 9 Years of UK &amp; EU Research from strategicfutures.org using Hexmaps: Each hexagon is a topic and clusters of topics are shown by using colours. This implementation uses clustering in order to distinguish major themes.</td>
</tr>
<tr>
<td>2.8</td>
<td>A section of network graph in [Blei and Lafferty, 2007]. Network graphs are a form of node-link graphs to visualise corpus view.</td>
</tr>
<tr>
<td>2.9</td>
<td>Example of network graphs used in TopicNets to visualise corpus view of results [Gretarsson et al., 2012].</td>
</tr>
<tr>
<td>2.10</td>
<td>Example of stacked barchart showing topic-document relationship [Relazioni, 2016].</td>
</tr>
<tr>
<td>2.11</td>
<td>Corpus view of LDAvis [Sievert and Shirley, 2014]. LDAvis visualise relationship between topics in two dimensional plane.</td>
</tr>
<tr>
<td>2.12</td>
<td>A section of corpus view from demo of dfr-browser. Simple circle on a grid has been used to show the topics.</td>
</tr>
<tr>
<td>2.13</td>
<td>Force directed bubble layout of topics is used by EPSRC to visualise their portfolio. This layout shows each topic as a bubble where size of the bubble is representations of how much grant is allocated for that topic. Live demonstration of this portfolio can be accessed from The EPSRC Portfolio by Research Areas.</td>
</tr>
</tbody>
</table>
2.14 Termite use heatmap to present topic-term relationship. Size of the bubble on the crossings represent strength of the relationship [Chuang et al., 2012] . ................................. 25
2.15 Argviz use heatmap to show probability of topics in current conversation. [Nguyen et al., 2013] ................................................................. 26
2.16 Heatmap of topic - topic similarity used in jsLDA .......................... 26
2.17 Dendrogram are a form of visualising hierarchical clustering of topics topics. Topics are listed at the leaves and iteratively most similar topics are joined to create a cluster. Example dendrogram from Slidshare .................................. 26
2.18 When the number of topics are too large, Polar Dendrogram can be used to utilise the available space on the screen. Example polar dendrogram from Research perspective website .................................................. 26
2.19 An example of pack clustering from LDAOverflow with Online LDA: In this layout each topic is presented as a circle. Topics in clusters can be easily distinguished they will be inside a bigger circle. ............................. 27
2.20 Treemap of topics from [Ganesan et al., 2015] ................................. 27
2.21 An example of pack clustering from a blog post. ............................. 27
2.22 Corpus view of 9 Years of UK & EU Research from strategicfutures.org using Hexmaps: Each hexagon is a topic and clusters of topics are shown by using colours. ................................................................. 28
2.23 Interface of Topicheck [Chuang et al., 2015]: TopicCheck shows stability of topics on different runs of the same model on dataset. This matrix will be filled if all topics were stable across all runs. ............................... 29
2.24 Example word lists from [Smith et al., 2017] .................................. 30
2.25 Corpus view of using word list [Chaney and Blei, 2012] .................. 30
2.26 Example word cloud (10 words) from [Smith et al., 2017] .............. 31
2.27 Example word cloud (20 words) from [Smith et al., 2017] .............. 31
2.28 Example topic cloud .................................................................. 31
2.29 Topic/document view from dfr-browser. This is an example of combining topic and document level in one view. A list of most related documents in this topic are presented at the bottom of the page ..................... 31
2.30 Examples of visualisation techniques used in [Smith et al., 2017]: Authors in this paper evaluated interpretibility of topics by asking participants to label them .......................................................... 36
2.31 Group 1 of study in [Smith et al., 2017]: were presented visualisations of topics and asked to label them and rate their confidence on this label ... 36
2.32 Group 2 of study in [Smith et al., 2017]: were presented with word list of top contributing documents to a topic and choose the best and worst label among 5 options(4 produced by group 1 and 1 is produced randomly) 36

4.1 Main steps of the project. High level explanation will be given in this chapter and detailed information about each step in corresponding chapters. 44
4.2 Pre-processing data involves 3 main steps: 1. dealing with missing/short abstracts; 2. reducing size of the dataset; 3. extracting project abstracts and titles to separate text files in order to feed in MALLET (next step) . 45
4.3 Topic modelling step: 1. Import documents to MALLET and get .mallet file; 2. Repeat 5 times with random seed set to 1, 1001, 500, 600645, 999993: a. Train 10 topics each with 10 top words; ......................... 47
4.4 Import command of MALLET is used with options to treat HTML tags and remove stop words from corpus. Stop words are words that don’t provide useful information about document, examples are punctuation (,, ?, !) or functional words (a, an, and, the). ................................. 49
4.5 Post-processing of results of topic modelling: 1. Computing topic to topic similarity matrix which is 50 x 50 2. Computing linkage table; 3. Computing agglomerative clustering using linkage table; 4. Sort topic clusters and models based on similarity. ................................. 50
4.6 Java class which takes topic-term weights in the excel format and convert it to Java matrix. Code of this class is available in Appendix C. ................................. 51
4.7 Format of excel file which contains topic-word weights for all 5 models, 50 topics in total. ................................................................. 51
4.8 Java class which takes topic-term weights and computes topic-topic similarity. Code of this class is available in Appendix C. ................................. 52
4.9 Java class which a matrix to a text file. Code of this class is available in Appendix C. ................................. 52
4.10 A section of topic-topic similarity matrix. It is a matrix of size 50 X 50. Values are in range of [0, 1] and the higher the value the greater the similarity. ................................. 53
4.11 The evolution of layouts in order to identify best layout for interface. Dendrogram, hexagons and tabular layout are 3 main prototypes of the interface which I developed using D3.js programming language. ................................. 54
4.12 Dendrogram produced using linkage table is the first prototype I developed. I used this layout to identify the best number of clusters to use in clustering algorithm. ................................. 55
4.13 High level schema rendering hexagons layout. ................................. 56
4.14 Hexagons layout is used to visualise one solution of topic modelling on Wellcome Trust dataset. Each hexagon is one topic and inside each topic top 5 words are displayed. ................................. 57
4.15 Hexagons layout: in this layout each hexagon is one topic, each island is one cluster and each colour is one model. ................................. 57
4.16 High level schema rendering hexagons layout. ................................. 58
4.17 Example implementation of transition of labels and colours on tabular layout. ................................. 58
4.18 Example of tooltip feature implemented on tabular layout. ................................. 58
4.19 Tabular web interface showing relationships between different models. Each row represent a model and each column is a cluster of topics. Topics in each column are the most similar topics across 5 models. ................................. 61
4.20 Main steps of evaluation phase of my project are designing the experiment, recruiting participant, pilot study and focus group. ................................. 62
4.21 10 Topics each showing weighted 5 top words. ................................. 63
4.22 Comparing 2 solutions of running LDA on Wellcome Trust dataset - Each solution(each row) displays 10 Topics each with weighted top 5 words. Row 1: is produced by running LDA Topic Modelling algorithm on Wellcome Trust grant funding dataset with random seed 1. Row 2: is result of running same algorithm on same dataset with same setting using random seed 1001. ................................. 64
4.23 3 solutions - 10 Topics with weighted top 5 words. Row 1: running LDA Topic Modelling algorithm on Wellcome Trust grant funding dataset with random seed 1. Row 2: result of running same algorithm on same dataset with same setting using random seed 500. ................................. 65

4.24 Tabular web interface showing relationships between different models. Each row represent a model and each column is a cluster of topics. Topics in each column are the most similar topics across 5 models. ................................. 65

4.25 Demographic questions in part 1. 3 questions are about level of Machine learning knowledge, age group and gender of participant. ................................. 66

4.26 Participant is asked to select the ”BEST” and ”WORST” models after investigating the interface ................................. 67

4.27 Questions to assess confidence of participant on their decision ................................. 67

5.1 The interface used for pilot experiment. Participant of this experiment made comments about confusing labels and also difficulty of comparing quality of topics using tooltip. These comments were addressed before focus group experiment and the interface were modified accordingly. ................................. 73

5.2 Hexagons layout: in this layout each hexagon is one topic, each island is one cluster and each colour is one model. Participant clearly liked this layout more than tabular layout ................................. 75

5.3 Process of selecting BEST model by eliminating the stable topics and examining the not stable topics in more detail. ................................. 76

5.4 Process of selecting BEST model by eliminating the WORST model. ................................. 76

5.5 Process of selecting BEST model by scoring quality of each topic. This figure is taken from participant’s notepad ................................. 77

5.6 Process of selecting BEST model by dividing models into Good and Bad models. ................................. 77

5.7 Distribution of BEST model before the group discussion: Participants were allowed to select up to 2 models. ................................. 78

5.8 Distribution of BEST model after the group discussion: Participants were allowed to select up to 2 models. ................................. 78

5.9 Two similar topics in model 3: Top 3 words for these 2 topics are same and based on this this model lost one vote after discussions. ................................. 79

5.10 Confidence on selected model before the group discussion ................................. 81

5.11 Confidence on selected model after the group discussion ................................. 81

A.1 Interface of OpenRefine which is used to handle missing / short abstracts. In this example the value ”No Data Entered” has 5066 matching rows out of 20,933 ................................. 93

B.1 Import-dir command of MALLET. This command imports add text files from specified directory and after removing stopwords produces .mallet file. ................................. 98

B.2 One of the output files of MALLET. This file lists all words in the dictionary with frequency of happening in topics. In other words it is the distribution of each word over topics. ................................. 98

B.3 train-topics command of MALLET which takes the .mallet file created in Section B.2 and finds 10 topics. ................................. 99

B.4 This file lists all words in the dictionary with frequency of happening in topics. In other words it is the sparse distribution of each word over topics. 100
B.5 –xml-topic-report command outputs a file in which for each topic lists top words and Dirichlet parameters. ........................................... 101

C.1 Structure of Java package for calculating cosine similarity of topics .... 105

D.1 Clusters Object in JSON dataset has information about 12 clusters and the member nodes of each cluster ................................. 110

D.2 Topics Object in JSON dataset has information about 50 topics for example which cluster they belong to and their top words. ............... 111

D.3 Dendrogram produced using linkage table is the first prototype I developed. I used this layout to identify the best number of clusters to use in clustering algorithm. .................................................. 111

D.4 Dendrogram produced produced by using complete distance function to generate the linkage table. Also colour codes clusters. ................. 111

D.5 Hexagons layout is used to visualise one solution of topic modelling on Wellcome Trust dataset. Each hexagon is one topic and inside each topic top 5 words are displayed. .............................................. 112

D.6 Hexagons layout: in this layout each hexagon is one topic, each island is one cluster and each colour is one model. ............................... 112

D.7 Example implementation of transition of labels and colours on tabular layout. .......................................................... 112

D.8 Example of tooltip feature implemented on tabular layout. .............. 112

D.9 10 Topics each showing weighted 5 top words. .............................. 112

D.10 3 solutions - 10 Topics with weighted top 5 words. Row 1: running LDA Topic Modelling algorithm on Wellcome Trust grant funding dataset with random seed 1. Row 2: result of running same algorithm on same dataset with same setting using random seed 500. .............................. 113

D.11 Tabular web interface showing relationships between different models. Each row represent a model and each column is a cluster of topics. Topics in each column are the most similar topics across 5 models. ................. 113
List of Tables

1.1 Example of non-deterministic results of LDA: Topic about "Children health studies" over 5 runs of LDA on the same dataset produce different words when different seed is used. Specially run # 5 adds "obesity" and "diabetes" whereas runs #1 and 4 have word "hiv". A domain expert will criticise this and will not trust the results as it is without more explanations or investigations .................................................. 3

1.2 Literature survey criteria: I will use this criteria list during literature survey ................................................................. 4

2.1 Main topic modelling frameworks: This table is an overview of widely used topic modelling framework in order to identify a suitable framework for my project ................................................................. 10

2.2 Available information in Wellcome Trust funding dataset: "project title" and "project abstract" will form main input to topic modelling, other fields will be used to provide more information about how grant funds are distributed over years, areas, organisations, and etc. ........................................ 15

3.2 Research hypotheses: I will be evaluating this hypotheses by conducting a focus group. These hypotheses are driven from project motivations ........................................... 39

3.4 Project main deliverables: these deliverables will be used in project plan and risk assessment ........................................... 40

3.6 User interface requirements: one of the main deliverables of my project is an interactive user interface. This tabel summarise the mandatory and optional features of it. This list is prepared by ideas from literature survey of topic modelling visualisation is Section 2.6 ................................ 42

4.2 Comparison of advantages and disadvantages of hexagons layout in visualising relationships between stochastic solutions ........................................... 57

4.4 Comparison of advantages and disadvantages of tabular layout in visualising relationships between stochastic solutions ........................................... 60

4.6 Main steps of evaluation phase of my project ........................................... 62

4.8 Main steps during focus group experiment ........................................... 63

5.2 All participants selected BEST model in their view by assessing quality of topics in different ways, Nobody looked into visual display and the fact that topics and models are ordered from left to right and top to bottom. ........................................... 80

5.5 Self-rated confidence levels before and after the group discussion. 50% of participant felt more confident after discussing the process of selection with other group members ........................................... 81

5.7 Research hypotheses: I evaluated these hypotheses during pilot and focus group experiment ........................................... 84
5.9 Table of recommendations on designing web interfaces where users need to select one BEST solution. These features will help different users to customise the interface based on how they process the displayed information. ................................. 89

A.1 All grants with these values for project abstract were excluded from dataset. .......................................................... 93

B.1 Format of the –output-doc-topics file. This file contains distribution of topics for each document. ................................................. 100

B.3 Default list of common English "stop words" from the text, MALLET topic modelling uses this default list. ............................. 104
Chapter 1

Introduction

1.1 Project motivations

The amount of text data being produced or digitised is expanding exponentially for example websites, blog posts, social media platforms, books, papers and etc. Using computational tool for extracting knowledge (information) from data is inevitable. these tools (machine learning algorithms) are very efficient in what they do but have two major issues which can cause scepticism among its users especially if they are not machine learning experts.

1. Non-deterministic results (re-producibility issue)

All machine learning algorithms that are dependent to randomness will output stochastic results, which means results of running the different models or different runs of same model on the same data will generate different solutions. Factor of randomness can be during initialisation of model or using stochastic search when space is very large. End users often are not informed of this fact which I believe is an ethical issue.

2. Results are difficult to understand and interpret

Results are produced by complicated mathematical equations and are not easy to understand for non-expert users.

Topic modelling\(^1\) is a prime example of non-deterministic algorithms. It is a form of unsupervised algorithm to analyse collections of un-structured text documents. It finds

\(^{1}\)Detailed information about topic modelling will be provided in Section 2.2
Figure 1.1: Main idea behind LDA: Each document contains a number of topics, where each topic is distribution of words. LDA uncovers these hidden topics using unsupervised probabilistic algorithms.

hidden topics (themes) in the collection using probabilistic algorithms. Each topic is represented by group of words that co-occur in the collection.

According to Blei [2011] simplest topic model is Latent Dirichlet Allocation (LDA). LDA assumes that each document made up of a number of topics, where each topic is a distribution of words from a fixed vocabulary. This representation of LDA is displayed in Figure 1.1([Blei, 2011]), using a generative process LDA finds groups of the words (topics) in each document with their probability.

[Blei, 2011] summarise this generative process for each document in two steps:

1. "Randomly choose a distribution over (pre-defined) topics".
2. "For each word in the document" do:
   "Randomly choose a topic from the distribution over topics in step #1".
   "Randomly choose a word from the corresponding distribution over the vocabulary."

LDA is very efficient in analysing large corpora of documents and is widely used in applications like text mining, classification and search engines but it suffers from above mentioned issues. These issues affect credibility of the topic models and users often resist basing their decisions on topic modelling analysis not only because they don’t trust the results but also because they can’t defend their decision to other stakeholders. A number of techniques can be employed to make topic models easier to understand and interpret;
1. **Improving LDA algorithm** so to produce high quality (coherent) topics in first place. Interpreting coherent topics will be easier.

2. **Dimensionality reduction** which is the easiest technique used to address this issue but it has a negative effect on generalisation of the model.

3. **Visualisation and clustering of results** of topic models. Visual representations are easier to understand for majority of people.

<table>
<thead>
<tr>
<th>Seed</th>
<th>Top 10 words of topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>study risk children health disease hiv factors treatment data population</td>
</tr>
<tr>
<td>2</td>
<td>health study children risk factors data population women care intervention</td>
</tr>
<tr>
<td>3</td>
<td>children study risk health women factors early development age maternal</td>
</tr>
<tr>
<td>4</td>
<td>health study risk children hiv data factors population intervention studies</td>
</tr>
<tr>
<td>5</td>
<td>study children risk obesity factors women diabetes age early development</td>
</tr>
</tbody>
</table>

**Table 1.1:** Example of non-deterministic results of LDA: Topic about “Children health studies” over 5 runs of LDA on the same dataset produce different words when different seed is used. Specially run # 5 adds ”obesity” and ”diabetes” where as runs #1 and 4 have word ”hiv”. A domain expert will criticise this and will not trust the results as it is without more explanations or investigations

Facilitating process of understanding and interpreting results of topic models will improve users trust and confidence on the model itself. This in turn will be useful in following situations:

1. **As a means of providing trust for decision makers.** Decision maker is anyone who needs to decide for big budgets or when this decision will have an impact on people’s lives. For example, high profile stakeholder in strategic planning, a college student when deciding about which subject to study in university, or a person deciding which area to buy a house.

2. European Union’s new General Data Protection Regulation which will become law in April 2018, states that data subjects possibly will have right to explanation for decisions that are solely made by machines. [Goodman and Flaxman, 2016].

3. **For teaching and educational purposes.**
1.2 Criteria for literature survey

Based on the motivations of the project, there are three main areas that I need to survey: topic modelling, topic modelling visualisation and topic modelling evaluation involving human participants. To organise my literature survey, I prepared following list of survey criteria. I will keep referring to these criteria throughout Chapter 2.

<table>
<thead>
<tr>
<th>#</th>
<th>Survey criteria</th>
<th>Related area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Key papers that explain topic models in general.</td>
<td>Topic modelling</td>
</tr>
<tr>
<td>2</td>
<td>Provides any form of explanations for topic modelling.</td>
<td>Topic modelling</td>
</tr>
<tr>
<td>3</td>
<td>Clustering techniques based on similarity measures.</td>
<td>Topic Modelling &amp; Visualisation</td>
</tr>
<tr>
<td>4</td>
<td>Uses any form of topic modelling visualisation to present results to users.</td>
<td>Visualisation</td>
</tr>
<tr>
<td>5</td>
<td>Papers that evaluate with people (Subjective evaluation).</td>
<td>Evaluation</td>
</tr>
<tr>
<td>6</td>
<td>Dataset: publicly available datasets, which can be used to drive decision.</td>
<td>Required technology</td>
</tr>
<tr>
<td>7</td>
<td>Open source framework in a language I am familiar with (Java, python).</td>
<td>Required technology</td>
</tr>
</tbody>
</table>

**Table 1.2:** Literature survey criteria: I will use this criteria list during literature survey
Chapter 2

Literature review

2.1 Introduction

2.2 Topic Modelling

Machine learning techniques are generally divided into two main groups of \textit{Supervised} and \textit{Unsupervised}. Topic modelling is a form of unsupervised machine learning algorithm where topics are defined using probabilistic methods. Based on MSc. project description set by my supervisor I will only focus on Latent Dirichlet Allocation (LDA) which is the most widely used topic model. Other models for example Latent Semantic Analysis (LSA) [Deerwester et al., 1990] are out of scope of this project.

2.2.1 Latent Dirichlet Allocation (LDA) and extensions of LDA

David Blei gave detailed introduction to probabilistic topic models especially Latent Dirichlet Allocation (LDA) in [Blei, 2011], he explained topic modelling as a set of algorithms which are used to discover hidden themes in unstructured corpora. LDA is the simplest of this set which was introduced by Blei himself for the first time in 2003\footnote{Original paper can be found in [Blei et al., 2003]}. LDA make 4 assumptions which are listed below: [Blei, 2011]

1. \textbf{Multiple topics in each document}: main assumption in all topic models is that each document is made up of different topics.

2. \textbf{Bag of words assumption}: the algorithm discards the order of the words and grammar in the documents and only counts how many times each word appeared in each document [Blei et al., 2003].
3. **Order of documents can change**: changing the order of the documents do not have any impact on the result.

4. **Fixed number of topics**: number of topics is selected by user and fixed during the process.

Based on above assumptions LDA finds top correlated words in entire corpus and call this set a *topic*. In topic modelling ”words” are not exclusive to one topic, means that same word can be present in different topics. This algorithm assign each word with a *weight* to each topic. One of the reasons for this is to support *polysemous* in languages.

According to [Blei, 2011] algorithm of LDA can be summerised as:

1. **Initialise parameters**: setting the number of topics, number of top words, number of iterations and hyper-parameters;

2. **Randomly initialise topic assignment for each word in each document**;

3. **Repeat for specified number of iterations**:
   (a) **Re-sample topic assignment for each word in each document** based on topic assignment of all other words and topic distribution for current document. This iterative step will generate coherent topics;

4. **Get results**

Many extensions were developed for LDA with focus on last three assumptions of LDA to provide better models. These improved models produce more accurate topics.

”**Bag of words**”:

One of the first improvements in this area was considering n-grams in sentences. In this approach words of each topic is generated considering previous words [Wallach, 2006]; Boyd-Graber, Blei, and Zhu introduced *LDAWN* - LDA with WORDNET\(^2\) - which uses word sense as a variable in LDA for disambiguating sense of the word in different context. Embedding sense of the word when words of each topic is generated is extremely useful in Natural Language Processing applications [Boyd-Graber et al., 2007]. Researchers in both papers claimed more accurate topics were generated by applying these extensions.

**Fixed number of topics**:  
Choosing the best number of topics has direct impact on success of the model, if this number is too little then topics will be too general, which in turn won’t enable user to extract useful information. On the other hand if it is too large then results will be too

\(^2\)More information about WORDNET can be found from [https://wordnet.princeton.edu/](https://wordnet.princeton.edu/)
complicated to interpret. An example of successful attempts to relax this assumptions is done by Teh et al. [2006]. The authors use Dirichlet process which is a non-parametric Bayesian approach to determine the best number of topics from document collection. Hierarchical topic model [Griffiths and Tenenbaum, 2004] address this issue by assuming that number of topics are infinite and orders them in a hierarchy. Abstract and high level topics are located near the root of tree and more solid and detailed topics are at the leaves.

**Order of the documents do not have any impact on result**

Because words and vocabulary change over time so for collections that have documents from different years and centuries this assumption is very misleading. This assumption often cause in forming two or more topics that are same in fact but use different words to represent the same topic or combine two or more different topics into one. In [Blei and Lafferty, 2006] dynamic topic model were introduced in which the algorithm, slice the document span to time slots and produce topics for time t+1 from topics of time t. A slightly different approach was taken by Jo et al. [2011], they produce time-stamped topics from collection and then generate a graph of topic evolution.

### 2.2.2 Interactive topic modelling (ITM)

User of the topics models often identify "bad" topics easily and ask questions like: "why this words are/ are not in this topic?", "why this documents are/are not included in this topic?". Interactive topic modelling in [Hu et al., 2014] try to address this questions by enabling user to provide feedback to the model and by doing this, model can keep the "good" topics and improve the "bad" topics. Basically ITM runs the topic modelling algorithm on the collection and then present the results (topics with high frequency words) to user and user can add or remove words of topics, this feedback from user is treated as new data and algorithm runs again with tree-based prior producing better topics.

Publicly available and free to use code for ITM can be found in GitHub link ³ (Criterion 5 from Section 1.2).

### 2.2.3 Supervised topic models

Topic modelling is an unsupervised algorithm but some researcher investigated forms of supervised algorithms with topic modelling which are completely out of scope of this project. examples can be found in [Li et al., 2015, Mcauliffe and Blei, 2008].

³This code is for web interface using HTML, JQuery, Ajax and JSON.
2.2.4 Summary of topic modelling

Improving topic modelling algorithms is an active area of research, but this area has already been investigated and experimented. Due to time limits on my project I will omit this area and only focus on visualisation aspects of the project. Based on project specification set by my supervisor I will use LDA as topic modelling algorithm in this project. Using LDA has following two advantages:

1. LDA with default settings is the baseline which has been used in majority of the papers in this field. Choosing this model will provide an option for comparing my results to other studies;

2. Non machine learning experts will use off the shelf tools, not customised (improved) algorithms. Because they will not have required skills and knowledge to use customised algorithms.
2.3 Topic modelling frameworks

In this section I will give an overview of main topic modelling frameworks in order to identify a suitable framework to use in my project.

**MAchine Learning for LanguagE Toolkit (MALLET)** [McCallum, 2002] is an open source, Java based framework which has a number of toolkit for clustering, document classification, topic modelling, and etc. MALLET implements a number of topic models including but not limited to LDA, Parallel LDA, Hierarchical LDA, and Labeled LDA. Extensive tutorials [Graham et al., 2012] and developer guides[^4] are available. Another main advantage of it is Incorporated data pre-processing tool.

**Gensim** [ˇReh˚uˇrek and Sojka, 2010] is another open source framework which is widely used, it is python based and implements a number of models, for example: "LSI, LDA, Random projections, Hierarchical LDA, and word2vec deep learning". Same as MALLET, help and support is available. It can process large datasets thanks to incremental algorithms and data streaming.

Next major open source framework is **Stanford topic modelling toolbox** [Ramage and Rosen, 2009]. It has been written in Scala[^5] and input and output are in the Excel format. One of the main disadvantages of this framework for my project is that I don’t have any experience of using Scala.

Some frameworks like **Yahoo LDA** [Narayanamurthy, 2011] and **MR LDA** [Zhai et al., 2012] are developed in a way that they can handle extremely large data collections very fast. Yahoo LDA is developed on Hadoop platform and written in C++, where as MR LDA implements LDA using MapReduce in Java. Zhai et al. [2012] successfully used MR LDA on a collection of 10 million blog posts.

There are various other open source packages for topic modelling in online repositories like "GitHub". Examples are implementation of "Interactive Topic Modelling (ITM)" [Hu et al., 2014] which is surveyed in Section 2.2.2. ITM is a Java based package developed originally in Mallet but moved to Apache Maven[^6]. David Blei and his team regularly release open source packages for their topic modelling publications on GitHub[^7]. Complete survey of available topic modelling frameworks is out of scope of the project. Table 2.1 summarises main features of topic modelling frameworks.

[^4]: Guides from MALLET website can be found [here](#). Video of Mallet introduction by David Mimno can be found [here](#).
[^5]: More information about Scala can be found [here](#)
[^6]: This information is taken from ITM repository on GitHub
[^7]: For more information please visit Blei Lab on [here](#).
Table 2.1: Main topic modelling frameworks: This table is an overview of widely used topic modelling framework in order to identify a suitable framework for my project

2.3.1 Summary of frameworks

Based on my survey criteria from Section 1.2 all frameworks mentioned in Section 2.3 meets the criteria of being open source. Among them MALLET is the most suitable framework for my project, as it implements variety of topic models, it is written in Java and extensive help and support is available online if required. Gensim will serve as backup plan which I will use in case of any major issues with MALLET.
2.4 Topic modelling applications and datasets

In this section I will list applications of topic modelling and datasets that has been used in literature and discuss their suitability to my project. This section by no means covers all available literature and only mentions a few examples for each application. Main purpose of this section is comparing different possible datasets in order to identify candidate datasets for my project.

2.4.1 Text mining in digital humanities

Digital humanities were pioneers in adopting and using topic modelling, for a simple reason, they have huge collections of unstructured text which humans are unable to read all of them to extract useful information. The first use cases of topic modelling in this domain focused on classification of very large archives to find important or interesting documents in areas of interest, for example, Block [2006] used topic modelling on a dataset of nearly 82,000 articles from Pennsylvania Gazette during 1728 - 1800. Blevins [2010] has done a similar work on daily diary of Martha Ballard over 27 years and Jockers and Mimno [2013] extracted themes (topics) 3,200 novels from 19th century. Majority of datasets in digital humanities are publicly available but these dataset do not meet all the equipments of my survey, to be precise they can not be used in strategic planning or any kind of decision making process.

2.4.2 Bio-informatics and Medical science

Pan et al. [2010] used an extension of LDA called LDA-RF (latent Dirichlet allocation-random forest) on protein data to predict protein interaction. They extracted "36,630 unique positive protein-protein pairs" from "human protein references database", this dataset and other versions can be requested from Human protein reference database website. [Zhao et al., 2014] proposed using topic modelling to cluster large sets of biological and medical data. They used 3 datasets to evaluate their proposed clustering method which are listed below: 1. "Salmonella PFGE genotyping data from CDC\textsuperscript{8}", 2. "lung cancer micro-array dataset", this dataset is publicly available from Gene Expression Omnibus website, and 3. "breast cancer micro-array dataset", this dataset is also publicly available and can be downloaded from Computational Cancer Biology website, authors of [Zhao et al., 2014] reported improvements in clustering results on all 3 datasets.

\textsuperscript{8}https://www.cdc.gov/pulsenet/
Park et al. [2014] proposed an extension to LDA which is called DMTM (Disease-Medicine Topic Model). This model incorporates Disease-Medicine relationship into model, in order to improve the quality of produced topics on patient records corpus. They evaluated their model on a dataset of 189,086 medical records from "South Korean healthcare data warehouse" and concluded that topics generated by DMTM are easier to identify. [Newman et al., 2009b] used LDA and ATM (Author Topic Modelling) on large set of medical articles to produce labels and subject headings for them. They retrieved articles using queries on MeSH database from US National library of Medicine.

Most of the datasets in this domain are publicly available and can be used to drive decisions from them, but they have a number of disadvantages for my project which are listed below:

1. They are very domain specific and often require in depth domain knowledge to interpret them,
2. It will be very time consuming to pre-process the data for using in my system and because of time constraints on MSc. projects it won’t be feasible to use them in my project,
3. I don’t have access to experts in this field in order to evaluate my system.

2.4.3 Natural language processing

A number of extensions of LDA are developed in order to incorporate features of spoken languages to LDA. In Section 2.2.1 I introduced [Wallach, 2006] which adds n-grams and [Boyd-Graber et al., 2007] which embeds sense of the words using WORDNET\(^9\) to improve produced topics. Wallach [2006] used 2 datasets to evaluate it’s model, 1. 150 random news postings from 20 NewsGroup dataset (this datasets has around 20,000 news articles from 20 news group. It is very popular in text clustering and topic modelling literature and is available from UCI machine learning repository\(^10\)), 2. 150 random abstracts from "Psychological Review Abstracts", Wallach mentions that second dataset is from topic modelling toolbox website, but I was not able to find this dataset on this website. Boyd-Graber et al. also evaluated using 2 datasets, 1. A dataset of labelled text documents called SEMCOR, and 2. A collection from "British National Corpus (BNC)". SEMCOR dataset and many other text based datasets are publicly available from Rada Mihalcea’ homepage and "British National Corpus (BNC)". SEMCOR is collection of

\(^9\)More information about WORDNET can be found from https://wordnet.princeton.edu/

\(^10\)https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups
100 million of English language samples and is available from British National Corpus website.

This paragraph covers 2 more recent papers in this field. [Liu et al., 2015, Nguyen et al., 2015] uses word embeddings with LDA to improve topic coherence in clustering tasks. Nguyen et al. uses 3 publicly available datasets, 1. 20 NewsGroup dataset, 2. TagMyNews dataset which has around 32,000 short English text extracted from RSS feeds of online newspapers, this dataset can be downloaded from TagMyNews Datasets website, 3. Sanders Twitter corpus, this dataset is free to use and contains 5,513 tweets is classified into 4 topics by hand, this dataset can be accessed from Twitter Sentiment Corpus website. Liu et al. use another popular public datasets which is 2 million English documents from Wikipedia in April 2010, more information about this dataset and download links can be obtained from Wesbury Lab Website.

Also all the datasets mentioned in this section have advantages for example: free, public datasets, and they are not domain specific so everyone will be able to evaluate them, they can’t be used for decision making situations. People will not base their decision on historic, social media, short texts from books and newspapers or Wikipeida data.

2.4.4 Opinion mining and short text clustering using social media data

[Hong and Davison, 2010b] collected 1,992,758 tweets from 514,130 users, [Alvarez-Melis and Saveski, 2016] pooled 161,607 tweets, using Tweeter API’s, then both performed LDA (Latent Dirichlet Allocation) and ATM (Author-Topic Model) models. They didn’t publish their collected data. One of the datasets used in [Nguyen et al., 2015] is Sanders Twitter sentiment corpus. As explained in Section 2.4.3 this dataset is free to use and contains 5,513 tweets which are classified into 4 topics by hand, this dataset can be accessed from Twitter Sentiment Corpus website. None of these datasets is not suitable for my project as I need data that can be used as a base for making decisions where it will affect self or others.

2.4.5 Describing research portfolios

Topic modelling has been successfully used to describe research and investment portfolios in recent years. Topic modelling can help in providing overview of main strengths and gaps in science, hot topics, and how topics and areas of research changed over time. In this section I will explain a number of most relevant works to my project.

11http://dev.twitter.com/
NIH is the largest source of funding in biomedical research in the world. It consists of 27 centres and in total funds about $32.3 billion\textsuperscript{12} each year. Information about awarded grants by NIH can be obtained from RePORTER website. NIH datasets has been used by researchers in different ways. For example, while Herr II et al. [2009] focused on all grants awarded in a specific year, Park et al. [2016] used a dataset of grants in specific theme (cancer projects) over 20 years. Herr II et al. used a dataset of 60,568 projects funded by NIH in 2007, and Park et al. used labelled LDA and logistic regression on a dataset of grants over longer period of time (a dataset of 149,901 projects about cancer funded by NIH over 20 years). NSF is the science and engineering counterpart of NIH. It funds $7.5 billion worth of research in the fields of science and engineering every year. TopicNets [Rehůřek and Sojka, 2010] as part of it’s evaluation uses a dataset of 50,000 NSF grant applications.

In all above mentioned cases main input to topic modelling was produced using ”project title” and ”project abstract” of each grant. One major advantage of grant datasets is they contain financial information for each grant, making them suitable candidate to drive decisions from them.

2.4.6 Candidate dataset

After checking a number of datasets I came to the conclusion that a grant funding dataset is the best option for my project, as it meets all the requirements set in Section 1.2. In this section I will investigate 2 candidate datasets from UK funding agencies and then describe my chosen dataset in more details.

**RCUK (Research Councils UK)** is the managing body of seven research councils in UK and all information about awarded grants by RCUK is publicly available from RCUK gateway to research under the Open Government Licence v2.0. **Wellcome Trust** is another funding agency in UK. Wellcome trust supports research in the areas of public health and bio-medicine. Wellcome trust’s dataset can be obtained from Wellcome Trust website.

Both candidates satisfy my survey criteria (publicly available and decision making possibility). RCUK at the time writing this report provide information about 72,000 projects, for the scope and timeline of MSc. dissertation project this dataset proves too large so I will use Wellcome trust’s dataset which consists of 20,933 projects. RCUK dataset will serve as backup (contingency) dataset for my project. In following paragraphs I will introduce my chosen dataset in more detail.

\textsuperscript{12}This information was obtained from NIH website
My chosen dataset is the dataset of all grants awarded by Wellcome Trust from 1 October 2000 to 30 September 2016. Wellcome trust’s website lists 5 main research areas that they fund: 1.”biomedical science”, 2.”population health”, 3.”product development and applied research”, 4.”humanities and social science”, 5.”public engagement and creative industries”. This dataset is in the form of excel spreadsheet and includes information about 20,933 grants in total. Table 2.2 shows available information and an example entry from this dataset. Main input to topic modelling tool will be ”project title” and ”project abstract”.

<table>
<thead>
<tr>
<th><strong>Field label</strong></th>
<th><strong>Example data</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial year</td>
<td>2015/16</td>
</tr>
<tr>
<td>Grant</td>
<td>202922</td>
</tr>
<tr>
<td>Grant Reference</td>
<td>202922/Z/16/Z</td>
</tr>
<tr>
<td>Applicant Surname</td>
<td>Chen</td>
</tr>
<tr>
<td>Lead Applicant</td>
<td>Prof Zhengming Chen</td>
</tr>
<tr>
<td>Other applicant(s)</td>
<td>Prof Robert Clarke, Dr Yi-Ping Chen, Prof Liming Li (Lee), Prof Sir Richard Peto, Prof Sir Rory Collins, Dr Robin Walters, Dr Guo Yu, Dr Iona Millwood</td>
</tr>
<tr>
<td>Sponsor(s)</td>
<td>-</td>
</tr>
<tr>
<td>Grant Type</td>
<td>Biomedical Resources Grant</td>
</tr>
<tr>
<td>Organisation</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Region</td>
<td>South East</td>
</tr>
<tr>
<td>Country</td>
<td>UNITED KINGDOM</td>
</tr>
<tr>
<td>Project Title</td>
<td>China Kadoorie Biobank (CKB) prospective study of 0.5 million adults</td>
</tr>
<tr>
<td>Project Abstract</td>
<td>CKB is a blood-based prospective study of 512,000 adults, recruited during 2004-8 from 10 diverse regions of China, with extensive data collected at baseline and subsequent resurveys using questionnaires, physical measurements, and stored biological samples. By 1.1.2014, 25,000 deaths and 1.5M coded disease events had been recorded among participants, through linkages with death and disease registries and ...</td>
</tr>
<tr>
<td>Start Date</td>
<td>01/09/2016</td>
</tr>
<tr>
<td>End Date</td>
<td>01/09/2018</td>
</tr>
<tr>
<td>Amount awarded (£)</td>
<td>1737405</td>
</tr>
<tr>
<td>Date of Award</td>
<td>16/06/2016</td>
</tr>
<tr>
<td>Panel</td>
<td>Biomedical Resources and Multi-User Equipment Committee</td>
</tr>
</tbody>
</table>

**Table 2.2**: Available information in Wellcome Trust funding dataset: ”project title” and ”project abstract” will form main input to topic modelling, other fields will be used to provide more information about how grant funds are distributed over years, areas, organisations, and etc.
Some of the fields in this dataset have missing data and some fields are redundant so pre-processing of this dataset is required. To gain better view of dataset I converted it to .arff format and then used Weka v.3.6 to analyse the dataset. Weka is a free machine learning software which is developed by University of Waikato\textsuperscript{13}. Figures 2.1, 2.2, 2.3 and 2.4 gives an overview of this dataset. A list of issues discovered are listed below:

1. "Grant" and "Applicant Surname" are redundant and can be removed without loosing any information.

2. "Sponsor(s)" has 13,486 missing values so this filed needs major work if I want to keep it. Due to time limits I will remove this field from dataset.

3. Typing errors are needs to be address, one major error of this type is adding full stop (.) at the end of some fields.

4. In project abstract which will form input to topic model there are 5,066 entries with "no data entered" value. One possible way of resolving this issue is using both "project title" and "project abstract" to produce topics.

\textbf{Figure 2.3:} Field "Sponsor" has 64% missing values in the dataset. This fields should be eliminated from dataset unless it gives critical information.\textsuperscript{13} More information can be found in \url{http://www.cs.waikato.ac.nz/ml/weka/}
2.4.7 Summary of datasets

In this section I surveyed areas that topic modelling has been used in order to identify suitable dataset for my project. Datasets used in area like humanities, medical science, natural language processing and opinion mining (using social media datasets) are not suitable candidates for my project as they often are historic data or data that cannot be used in decision making process.

Finally grant datasets which has been used in research portfolios satisfy requirements for my dataset from Section 1.2. Among possible datasets I decided to choose dataset of all grants awarded by Wellcome Trust from 1 October 2000 to 30 September 2016. I provided detailed description of this dataset in Section 2.4.6.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>150</td>
</tr>
<tr>
<td>Maximum</td>
<td>45000000</td>
</tr>
<tr>
<td>Mean</td>
<td>337474.723</td>
</tr>
<tr>
<td>StdDev</td>
<td>989836.051</td>
</tr>
</tbody>
</table>

Figure 2.4: Overview of amount of fund awarded by Wellcome Trust during 16 years period.
2.5 Clustering algorithms

Based on assumptions of project specification provided by my supervisor, similarity of topics are more important than the quality of the topics, this assumption steered me towards the relationships between topics rather than the quality of them. Similarity is closely relates to clustering and because of this reason I surveyed the clustering algorithms in this section.

2.5.1 Introduction

Clustering is an unsupervised algorithm which tries to find groups / structures in unlabelled data. A cluster is a collection of things that are similar. This explanation is very similar to explanation provided for Topic Modelling in Section 2.2, in other words topic modelling is a form of clustering algorithm. 4 main types of clustering algorithms are:

1. **Exclusive:**
   
   Each object belongs to only one cluster;

2. **Overlapping:**
   
   Each object can be a member of more than one clusters, the degree of membership or weight of object in each cluster is defined using fuzzy algorithms;

3. **Hierarchical:**
   
   As name implies it is a hierarchy of clusters where algorithm starts by assigning each object to a cluster then merges similar clusters successively until the condition is met;

4. **Probabilistic:**
   
   It clusters objects using parametric distributions for example Gaussian.

Based on project specification the most appropriate clustering algorithm is the hierarchical method. In Section 2.5.2 this algorithm is explained in more detail.
2.5.2 Hierarchical clustering

As name implies it is a hierarchy of clusters where algorithm starts by assigning each object to a cluster then merges similar clusters successively until the condition is met. Two types of hierarchical clustering, are agglomerative (bottom-up) and divisive (top-down). Figure 2.5 shows the difference between these two types in one glance.

Divisive or top-down method starts by assigning all objects to one cluster then repeatedly divides each cluster to two least similar groups until each object is assigned to one cluster. Agglomerative is the opposite of divisive method in which each object is assigned to one cluster and two most similar clusters are joined successively until there is only one cluster.

Clustering algorithms take a similarity/difference matrix as input and then compute linkage table to form clusters. There are several approaches to measure similarity between objects for example: KL-divergence [Newman et al., 2009a], average Log Odds Ratio Chaney and Blei [2012], cosine similarity [He et al., 2009].

Linkage table is produced by calculating distance between clusters, this distance function can be single linkage (shortest distance between two members of each cluster), complete linkage (greatest distance between clusters) or average linkage (average of distance between each member of cluster 1 to every member in cluster 2).

2.5.3 Hierarchical clustering with topic modelling

Hierarchical Agglomerative Clustering(HAC) is used in different ways with topic modelling:

1. **To improve the algorithm of LDA:**

   HLDA (Hierarchical LDA) [Griffiths and Tenenbaum, 2004] is one of the first successful attempts to embed HAC to algorithm of LDA. This improved model
enables user to drill down into topics to find more detailed information. Having a fixed number of topic will make some topics to be very general. [Balagopalan, 2012] uses HAC to increase re-producibility of topics in LDA. He clusters results of different runs of LDA (with different random seed) then feeds this clusters to LDA in initialisation phase. [Lancichinetti et al., 2015] also uses clustering as part of algorithm to improve re-producibility and accuracy of topics.

2. In order to analyse the results:

One of the prime examples of using clustering to analysis results of topic modelling is developed by strategicfutures.org. This website presents a data-driven analysis of research funding over 9 years. Each topic is presented as a hexagon and clusters of topics is distinguished by colours. This example clusters results of only one solution of topic modelling. To best of my knowledge [Chuang et al., 2015] is the only scientific paper which uses clustering to display solutions in an interactive interface to evaluate stability of topics over models.

Figures 2.6 and 2.7: Interface of TopicCheck [Chuang et al., 2015]: TopicCheck uses clustering algorithm on results of multiple runs of topic modelling in order to evaluate stability of topics. This matrix will be filled if all topics were stable across all runs.

Figures 2.6 and 2.7: Corpus view of 9 Years of UK & EU Research from strategicfutures.org using Hexmaps: Each hexagon is a topic and clusters of topics are shown by using colours. This implementation uses clustering in order to distinguish major themes.
2.5.4 Summary of clustering algorithms

Based on project specification, my intuition is that using hierarchical agglomerative clustering on combined topics from multiple runs of the model will work better to group similar topics over different runs. Then I can use these clusters to visualise similarity of topics over different models which in turn will help user to select the "BEST" model among available one to use.

2.6 Topic modelling visualisation

In this section I will explain techniques and tools used to visualise topic modelling. Techniques are different layouts and visual representations on the interface and tools are the languages or library which has been used to produce such layouts.

Techniques used to visualise topic models can be divided into 3 group of visualising corpus level, visualising topic level, and visualising document level.

2.6.1 Techniques - Corpus level

This view show all topics in the corpus to give an overall view. I briefly explain main representations of corpus. For each techniques I will show an example figure.

2.6.1.1 Network graphs

In this form of visualisation topics are shown as node-link graphs. Nodes of the graph and words of the topics and links represent words that occur together. Some implementations change the weight of the link according the strength of the connection. Network graphs has been implemented in both static and interactive forms to show relationships between topics. Following paragraphs mention examples of each form before enumerating advantages and disadvantages of using this techniques in my project.
Blei and Lafferty [2007] demonstrated correlated topic model using a static network graph. Each node of the graph shows 5 top words and font of the words are representation of their contribution in that topic. A section of the topic network graph from [Blei and Lafferty, 2007] are shown in Figure 2.8. [Smith et al., 2017] also used static network graph as one of the visualisation techniques in their user study to evaluate interpretability of topics by humans.

TopicNets [Gretarsson et al., 2012] is an interactive web-based implementation of this technique. Main objective of TopicNet is knowledge discovery and to achieve this it combines three research areas of text analysis, topic modelling and information visualisation. The authors reported successful implementations and use cases of TopicNets (1. 50,000 NSF grant projects, 2. 10,000 publications of a research centre, 3. single documents of PhD thesis). Examples of network graph layouts used in TopicNets can be seen in Figure 2.8. Techniques used in TopicNet [Gretarsson et al., 2012] has 2 main disadvantages for my project:

1. This system is intended for knowledge discovery, also having more knowledge will be useful in "trust issue", they are completely different issues.

2. The visualisation techniques used in TopicNets (network graphs) is very complicated. People are more reluctant to use systems that they find complicated, this again in turn will degrade their trust.

Positive points from TopicNets [Gretarsson et al., 2012] that I can use in my project:

1. Different levels of visualisations: corpus level, topic level, document level.

2. Search and filter options.


2.6.1.2 Stacked barcharts

Stacked barchart as the name implies stacks topics in each document on top of each other. Because topic modelling assumes that each document is made up of \( n \) topics, so topic-document relationship can be displayed by stacked barcharts. This form of visualisation will work better when the number of topics are small and can only represent document-topics relationship not topic-topic. This layout will not add any value to my interface so I will not implement it in my project.

**Figure 2.9:** Example of network graphs used in TopicNets to visualise corpus view of results [Gretarsson et al., 2012].

**Figure 2.10:** Example of stacked barchart showing topic-document relationship [Relazioni, 2016].
2.6.1.3 Dimensionality reduction using PCA

PCA (Principal Component Analysis) is basically reducing dimensions of the model to two and presenting the correlation between topics on a two axes grid. LDAvis [Sievert and Shirley, 2014] is a prime example of interface which use PCA to present corpus (global) view of topics. It uses circles to plot topics in 2 dimensional plane. This layout can show the relationship between topics but some topics will overlap and for non expert users it is difficult to understand the mathematical reason behind it, because of this reason I will not use this layout.

2.6.1.4 Bubbles

One of the simplest ways used to visualise corpus view is by presenting topics as a circles on a grid. In a demonstration of dfr-browser, same size circles on a grid used to show topics in PMLA (Publications of the Modern Language Association of America), inside each circle a list of 6 top words in the topic is displayed.
EPSRC (The Engineering and Physical Sciences Research Council) use force directed bubble layout to visualise their portfolio, example screenshot of this layout from EPSRC website can be seen in Figure 2.13. This layout is visually appealing and intuitive but because it uses force directed layouts use a random number to start the process so it won't be obvious to show that change in the layout is because of the change in results of topic model not random function on force layout.

2.6.1.5 Heatmaps - similarity matrix

Heatmap is a form of showing aggregate data in a rectangular tiles. Colour of the tiles are shaded to represent the similarity of each topic to other topics. Termite [Chuang et al., 2012] is a tool for checking quality of topics in large text collections. It uses simple tabular view in the form of heatmap to show relationships between topics and the terms representing them, see Figure 2.14. Using tabular views has one big advantage which is its simplicity, user can easily compare topics and identify themes. Argviz [Nguyen et al., 2013] is an interactive framework for analysing topics in multiparty conversations. Argviz uses heatmap as way of showing probability of topics in current turn of conversation, see Figure 2.15. David Mimno uses heatmap to show the similarity between topics on a grid of circles. Size of the circles scaled to show the similarity, Figure 2.16.
2.6.1.6 Hierarchical clustering - dendrograms and trees

Dendrogram is a form of hierarchical tree layout which shows correlation between topics. All topics are listed at the bottom of the dendrogram and then clusters of topics are formed by iterative joining 2 most correlated topics. Main input to produce dendrograms is topic-topic similarity matrix (heatmap). Dendrograms can be oriented in different directions, for example: top to bottom (Figure 2.17) or polar (Figure 2.18).

**Figure 2.15:** Argyiz use heatmap to show probability of topics in current conversation. [Nguyen et al., 2013]

**Figure 2.16:** Heatmap of topic - topic similarity used in jsLDA

**Figure 2.17:** Dendrogram are a form of visualising hierarchical clustering of topics topics. Topics are listed at the leaves and iteratively most similar topics are joined to create a cluster. Example dendrogram from Slidshare

**Figure 2.18:** When the number of topics are too large, Polar Dendrogram can be used to utilise the available space on the screen. Example polar dendrogram from Research perspective website
2.6.1.7 Hierarchical clustering - packs and treemaps

Pack layout shows clusters (groups) of topics using information from similarity (correlation) matrix. By using colour pallets and other visual formatting techniques, clusters can be identified easily. An example of this layout can be seen in Figure 2.19. By adding zoom functionality to this layout, users can change the granularity of topics, which will help them in the process of understanding the corpus and eventually making informed decisions.

LDAExplorer [Ganesan et al., 2015] implements treemap clustering to display topics. Treemap is collection of rectangles in which each rectangle is a topic. It can be implemented as a flat representation of topics [Ganesan et al., 2015] (see Figure 2.20) or can be adopted to show the hierarchy of topics (see Figure 2.21).

Simple shapes on a fixed frame is the easiest to understand and remember the relationships they are showing. So either treemap or pack layout is one of the candidate techniques to use in my interface.
2.6.1.8 Hex maps

One of the most intuitive and visually appealing techniques of displaying corpus view is hex map which is used in strategicfutures.org. This website presents a data-driven analysis of research funding over 9 years. Each topic is presented as a hexagon and clusters of topics is distinguished by colours. Inside each hexagon a word cloud of top 5 words for each topic is displayed. The only negative point of this layout in terms of my project is that, minor changes in the topics won’t be easy to identify. I need to experiment with this layout to identify possible ways of highlighting non-determinist aspect of topic modelling results.

![Figure 2.22: Corpus view of 9 Years of UK & EU Research from strategicfutures.org using Hexmaps: Each hexagon is a topic and clusters of topics are shown by using colours.](image)

2.6.1.9 Visualising non-determinist aspect of topic modelling

As I mentioned in section 1.1 topic models are non-deterministic. They can produce different results on the same collection with different models or even different runs of the same model. Because of 2 main reasons: using random seed to initialise the model and stochastic search to produce results. To best of my knowledge [Chuang et al., 2015] is the only scientific paper which discuss this issue of topic models and provide an interactive interface to evaluate it, this paper satisfy criteria 2, 3, and 4 of my survey from Section 1.2 so I will discuss it in more details. This tool is called TopicCheck and is based on 4 guidelines of reproducible codes from [Krippendorff, 2004]14. TopicCheck is an agglomerative clustering technique, inputs to the algorithm are: topic models, topic similarity matrix and criteria to evaluate. It generates a matrix of topic groups (each group has maximum one topic from each model), this topic groups then checked iteratively to group 2 most similar groups until no match is found (based on the matching criteria) [Chuang et al., 2015]

---

14This paper is only cited for information and I did not read it.
In the above paragraph "list of topic models" can be m different topic models or m runs of the same model. Running same model on the same data will generate different results because it uses different seed to initialise the process. TopicCheck use explanatory visualisation for assessing stability issue with topic models and some aspects of it will be useful in my project which I will list them below.

1. Web based interface is deployed using JavaScript.

2. Use tabular layout to present the results. People like tabular and simple layouts because it is easier on eyes and they can easily see the relationships between topics. User feels more confident to use the tool if it does not look very complicated.

3. The authors evaluated TopicCheck on 3 datasets each for different aspect of topic models. first 2 tests are not relevant to my project as they don’t use human participants, so I will not discuss them, main points of 3rd test are listed here:
   - Data: More than 200,000 hours of news broadcast. - Main task was to identify topics quality using 50,000 user ratings
   - Word intrusion test, and quantitative test: for each topical group they calculated "coherence score" and concluded that the higher this score the better the quality of topical group.

Figure 2.23 shows an example of 50 runs of same topic model on the same dataset, each column is one run and each row is one topic. Topics at the top of the matrix are stable but towards the bottom of the the matrix instability is very obvious. At the right side of the graph top words in each topic are listed.

Figure 2.23: Interface of Topicheck [Chuang et al., 2015]: TopicCheck shows stability of topics on different runs of the same model on dataset.

This matrix will be filled if all topics were stable across all runs.

---

15 They used Amazon Mechanical Turk to find participants, for more information please visit [here](#)
2.6.2 Techniques - Topic level

Main purpose of topic view is to give more information about topic, for example what are the top frequency words are, or/and what are the top documents in the topic are. using word list and word clouds are the most widely used techniques which are often accompanied with simple charts, e.g. barchart, piechart.

2.6.2.1 Word list

Simplest way of visualising and presenting topics is displaying N high probability words for each topic. All topic models represent each topic as a set of words that appear together. To represent weight of each word some people change the font of the word accordingly or show barchart of the distribution of words next to each word.

For example one of the techniques used in [Smith et al., 2017] was presenting topics as vertical list of top words with barcharts next to them to show their weights (Figure 2.24). Also word lists are mainly used to visualise topic level, Chaney and Blei [2012] used it in corpus view where topics are presented as a vertical barchart and top words of the topic are listed inside each bar(Figure 2.25).

This technique is easy to understand and even non-expert users can make sense of what the topic is about. I will use this visualisation on topic view of my interface.

2.6.2.2 Word cloud or tag cloud

Word clouds, also known as tag clouds present the words of each topic which font of word is an indication of either frequency or rank of the word in topic. The number of words can vary from as little as 5 to as much as 100 words. Examples of word clouds are shown in Figures 2.26, 2.27, and 2.28. I will add word clouds to my user interface. I will choose number of words in each cloud during implementation of interface.
2.6.3 Techniques - Document level

This view lists top documents in the current topic. It is main purpose is allowing user to check the documents in detail if he / she wants to get more information about topic. This view is often combined with topic view. In situations where trust is an issue users should be able to see main documents in the topic. As a results of this reason I will incorporate this view in my interface.

Figure 2.29: Topic/document view from dfr-browser. This is an example of combining topic and document level in one view. A list of most related documents in this topic are presented at the bottom of the page.
2.6.4 Visualisation Tools

I will not survey existing tools and APIs for visualisations of topic models and only explain my chosen tool which is D3.js programming language. My reasons for making this decision are listed here:

1. Main objective of this project is not choosing best tool for visualisation.

2. D3.js is a powerful java script library which is been used to visualise data from variety of data sources. It is open source and can be used with other web technologies like HTML, CSV and SVG to produce intuitive visualisations. more information about D3.js can be found in Data-Driven Documents.

3. I am familiar with D3.js, as I used this language as part of my previous course\textsuperscript{16}. As a result the risk and overhead of learning new programming language will be minimised.

2.6.5 Summary of visualisation

A visual web interface is one of the main deliverables of my project and in this section I reviewed existing visualisation techniques and introduced my chosen tool (D3.js) to implement these techniques. Main features of the interface I will be developing are:

1. Web-based interactive interface;

2. In order to identify most suitable layout I will prototype a number of layouts for example tabular layout with simple shapes to present the corpus level. This will help users to identify the changes in the topics in different runs of the model;

3. Multiple views, corpus view, topic and document view.

4. Intuitive design, by adding interactivity, tooltips, transitions, and search and filter options.

2.7 Evaluation methods

2.7.1 Quantitative and Qualitative Experiment

There are 2 types of experiments than can be employed in research projects, quantitative and qualitative experiments. Simplest way to differentiate these methods are type of data that they collect and generate. Quantitative experiment generates numerical data which statistical data analysis techniques can be applied on them. On the other hand qualitative research aims to is exploratory and tries to gain insights to ‘how’ and ‘why” about a behaviour [Carr, 1994].

\textsuperscript{16}Data analytic and visualisation
2.7.1.1 Quantitative experiment

As name implies this type of experiment collects numeric data, then by applying statistical tests drive statements or facts about the data. These tests can be simple descriptive statistics such as mean or median, or can be complicated inferential tests such as correlation (Chi-square, pearson correlation), regression (simple/multiple) or comparison of means (ANOVA, t-test). By performing such tests researcher can calculate p-value to see whether the findings are "statistically significant".

Automated processes such as questionnaire, surveys or coding are usually employed to collect this type of data. As a result size of the sample is much larger than qualitative research. In recent years crowd-sourcing platforms became very popular among researchers which enable them to reach more participants in a short time. Amazon Mechanical Turk is an example of such platforms.

2.7.1.2 Qualitative experiment

Quantitative experiment tells "What" but it’s unable to find "Why", "how". Goal of qualitative research is to find trends and insights by documenting opinions and behaviour of participants. These insights then can be used in two ways: 1. highlighting some open questions that researcher would follow up in terms of quantitative analysis, 2. provide recommendations for decision making or designing products. Because this type of experiment can not be automated it is time consuming sample size is usually very small.

Methods of collecting this type of data in research projects are open-ended questionnaire, interviews or observing behaviour.

1. Individual interview:

Interview can be unstructured such as informal conversation about any topic, this method is not suitable for evaluating a project but rather will help in finding ideas to start up a project. On the other side in structured interviews all participants will be asked to answer same set of questions.

2. Focus group:

[Morgan, 1997] defines focus group as a group interview that relies on interaction among members of the group. This group discussions are guided by a moderator which usually is the researcher.

(a) Main advantage of focus group is that because participants are exposed to other members responses it can provide extensive in-depth information about the topics in relatively short time:
(b) On the other hand disadvantage of this type of research is due to sample size it is hard to generalise the finding, also some participants answeres might be biased if there is a very dominant participant in the group. For example participant might hide his/her true opinion in presence of his/her boss.

2.7.2 Topic modelling evaluation methods

Most widely used topic modelling evaluation methods and main papers in this area are surveyed in this section. I will finish evaluation methods with in depth survey of one very relevant user study in evaluating interpretibility of topic models in Section 2.7.3. Topic modelling is unsupervised machine learning algorithm, so it does not require validation data. This feature is often considered as one of the advantages of topic modelling but on the other hand evaluation of it’s results prove to be a challenge. Evaluation methods of topic modelling results are divided in 2 main groups: Statistical methods, and evaluating by humans.

Statistical methods or quantitative experiment:
Also statistical methods, (for example calculating cosine similarity [Aletras and Stevenson, 2014] to measure similarity of topics to each other and marginal probability of test data (perplexity) [Blei et al., 2003, Wallach et al., 2009]) provide useful insights about whether topic models are successful in producing topics or not, they are out of scope of this project and I Will not consider them in my literature survey.

Evaluating by human or qualitative experiment:
Evaluation methods by humans which also called ”human in the middle” try to measure interpretability of inferred topics. Word intrusion and topic intrusion tests was proposed by Chang et al. [2009] to evaluate topic models. In word intrusion participants were asked to identify the intruder among 6 words (5 top words of each topic + 1 word with low probability in that topic). For coherent topics participants should be able to easily identify intruder. For example, in the set of dog, cat, horse, apple, pig, cow [Chang et al., 2009] intruder is apple because other 5 words make sense as a group. On the other hand if participants find it difficult to identify intruder (e.g. choosing intruder at random), it means that generated topic is not coherent. Topic intrusion test [Chang et al., 2009] is performed to check quality of found topics in each document by model. Each participant was presented with title and a few lines of one document and 4 sets of words (topics). 3 of the high probability topics along with 1 very low probability topic from same document. Participants were asked to identify the topic that does not make sense in presented document.
Both word and topic intrusion tests aim to measure whether results of topic models make sense to humans. This is different from my objective for this project. I aim to see whether people will trust the non-deterministic results of topic models in decision making process. The closest user study I could find measures people’s understanding of topics by asking them to label them [Smith et al., 2017], I will explain this paper in more detail in next section (Section 2.7.3).

2.7.3 Summary of paper ”Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic Labels”

Smith et al. evaluate interoperability of topics by users which is to some extend relative to objectives of my project. Their experiment is based on the idea that descriptive label will improve understandability of results of the topic model. They presented same topics in four different visualisation forms to one group of people and asked them to label these topics, then second group is asked to assess not only these labels (manually produced by group 1) but also labels that automatically generated from Wikipedia articles [Smith et al., 2017]. Main points from this paper are listed below [Smith et al., 2017]:

1. 50 topics generated using MALLET implementation of LDA [Yao et al., 2009]17.

2. 4 visualisation techniques used: word list, word list with bar chart to represent the probability of words in topic, word cloud and network graph (node-link diagram). With each visualisation they experimented with 3 cardinality, (number of words to display in each topic). Please see Figure 2.30 for examples of the visualisation techniques used in this experiment.

3. Between subjects evaluation method were used on Amazon Mechanical Turk users. 589 users in group 1 produced 3,212 manual labels for topics, then group 2 were presented with top 10 documents of the topic and 5 labels (4 manually generated by users of group 1 and 1 automatically generated from Wikipedia articles) and were asked to select the best and worst title for each topic.

4. 3 type of information were recorded, time it took to label each topic by users in group 1, group 1 was also asked to rate their confidence on labels they gave for each topic. This information was recorded in a 5 scale likert chart, and finally ratings of the topic labels provided by group 2 users.

5. If generated topics are not good topics then user’s ability in giving good label to that topic will be affected. Because of this issue they used ”topic coherence”

---

17This paper is only cited for information and I did not read it.
to analyse the results of the user study. "Topic coherence" is basically a way of measuring "how much sense a topic makes to user" [Smith et al., 2017].

6. They concluded their experiment with following statements [Smith et al., 2017];

"The four visualisation techniques lead to similar quality labels as rated by end users. However, users label more quickly with the simple word list, yet tend to incorporate phrases and more generic terminology when using the more complex network graph. Additionally, users feel more confident labelling coherent topics, and manual labels far outperform the automatically generated labels against which they were evaluated.”

**Figure 2.30:** Examples of visualisation techniques used in [Smith et al., 2017]: Authors in this paper evaluated interpretability of topics by asking participants to label them.

**Figure 2.31:** Group 1 of study in [Smith et al., 2017]: were presented visualisations of topics and asked to label them and rate their confidence on this label

**Figure 2.32:** Group 2 of study in [Smith et al., 2017]: were presented with word list of top contributing documents to a topic and choose the best and worst label among 5 options(4 produced by group 1 and 1 is produced randomly)
This paper satisfy criteria 3 and 4 of my survey from Section 1.2 and will help me in designing evaluation plans for my project. I identified 3 good points from this paper that I can implement in my project:

1. Using simple visualisation techniques will improve people’s confidence and understanding of the topics.

2. Two phase evaluation can be beneficial for my project as well, where in phase 1 participants can be asked to make decisions on a number funding applications and phase 2 participants can be asked to validate and rate decisions of group 1.

3. Asking participants to rate their confidence will give me useful information.

2.7.4 Summary of Evaluation

In this section I provided brief overview of main evaluation methods for topic modelling results (statistical methods and human in the loop methods) but this methods mainly measure quality and coherence of the produced topics. Based on my project objectives I will need to carryout an experiment to evaluate “which” and “how” people select a model from available models. The first question can be answered using quantitative experiment but the second one requires qualitative data. Due to limited time and resources I have for evaluation of my project I will conduct a focus group but in the meantime will ask participants of focus group to fill small questionnaire. The objective of my experiment will be to provide design recommendations for developing web interfaces and open ended questions to form a base for future work.
Chapter 3

Requirements analysis

3.1 Introduction

In this chapter I will define scope and research hypotheses then set out requirements of the project, and finally I explain candidate experiment designs to evaluate hypotheses.

3.2 Project scope

In this project I will develop a simple web based interface to present non-deterministic results of topic modelling algorithm. In order to reduce cognitive load on people this interface will be very simple so they can focus on the main purpose of the interface which is a tool to help them select the best solution for their purpose. This simplicity is crucial in my project unlike usual trend among computer scientists that make interfaces very complex looking certainly in terms of visualisations.

Target user group is anyone who is responsible in making decisions which will affect self or others. Decisions can vary from high level strategic planning to non-strategic decisions (e.g. What to study in university, getting information about areas of the town to decide which area is good to buy a house).

Outcome of this project will be a user experiment to evaluate whether non-deterministic nature of topic modelling will affect the trust of user, whether a group of people be able to agree on one single solution and finally recommendations for designing such interfaces. These recommendations will be driven by documenting how people choose the best solution, what is the process they take to do so.

This project has fixed deadline (18 August 2017), and should be delivered by that date regardless of state of the implementation and progress. Because of this constraint, improving algorithm of the topic models is out of scope of this project and will not be experimented.
3.3 Research hypotheses

Based on project motivations from Section 1.1, I will design and implement my user interface in order to evaluate following hypotheses.

<table>
<thead>
<tr>
<th>#</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exposing non-deterministic (stochastic) nature of topic modelling will reduce confidence of people on produced results in terms of both using them for planning purposes and explaining them to third parties.</td>
</tr>
<tr>
<td>2</td>
<td>The visible consensus (general agreement) of other people in a group will increase person’s confidence on produced results in terms of both using them for planning purposes and explaining them to third parties.</td>
</tr>
<tr>
<td>3</td>
<td>Solution which would be least different from all other solutions would be the one that people will pick.</td>
</tr>
</tbody>
</table>

**Table 3.2**: Research hypotheses: I will be evaluating this hypotheses by conducting a focus group. These hypotheses are driven from project motivations

3.4 Project requirements

I divided requirements of the project into project deliverables and interface requirements. Project deliverables list main requirements in terms of the project for example report, poster and a web based user interface (Table 3.4).

Interface requirements list mandatory and optional features of the user interface which are listed in Table 3.6.

<table>
<thead>
<tr>
<th>#</th>
<th>Project deliverables</th>
<th>Mandatory?</th>
<th>Notes and Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Project report</td>
<td>Yes</td>
<td>Deadline 18 August 2017</td>
</tr>
<tr>
<td>2</td>
<td>Project poster</td>
<td>Yes</td>
<td>Deadline 25 August 2017</td>
</tr>
</tbody>
</table>
3. A web based interface which user can use it to make informed decisions

Yes

This user interface will be the only source of presenting information to user so he / she can use if to make decisions. Requirements of this interface are listed in Table 3.6.

4. An Recommendations for designing interfaces which will be used to help people make informed decision when non-deterministic algorithms are used.

Yes

This will be delivered by observing and documenting how participants of focus group experiment select best solution.

Table 3.4: Project main deliverables: these deliverables will be used in project plan and risk assessment.

In Table 3.6 requirements for user interface listed. Requirements are clearly divided to mandatory and optional features. In Chapter 4 I will explain that I will employ incremental build approach in my project, so I will be implementing mandatory features in prototype 1 before attempting to add / remove nice-to-have optional feature. As I explained in Section 3.2 web interface needs to be as simple as possible in terms of visualisations in order to reduce cognitive load on people so they focus on the task rather than features of interface.
<table>
<thead>
<tr>
<th>#</th>
<th>Interface requirements</th>
<th>Mandatory?</th>
<th>Notes and Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><strong>Corpus view:</strong> This view will give an overall view of the corpus, how topics relate together. This interface will be the first layout presented to user in user study so will have a huge impact on user’s view about system. It needs to be simple enough for non-expert users to use and understand it’s content, but at the same time it needs to be intuitive and provide information that user needs in order to make an informed decision.</td>
<td>Yes</td>
<td>This interface will be evaluated directly.</td>
</tr>
<tr>
<td>2.</td>
<td>A layout to show topics change on different runs of the model. Candidate solutions are:</td>
<td>Yes</td>
<td>This layout will be main way of testing my hypothesis.</td>
</tr>
<tr>
<td></td>
<td>1. using same methods as TopicCheck [Chuang et al., 2015]. In Section 2.6.1.9 I explained that TopicCheck use a simple matrix to show stability of the topics generated on different runs of the model.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Showing corpus view of different runs on a tabular layout in the screen.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| 3. | **Topic and document view:**  
User will be able to see more detailed information about selected topic. For example: word list (top N words in the topic), word clouds, link to top documents in the topic. | No | The evaluation of my project will be in the form of a focus group so participant’s won’t be able to interact with interface themselves. The interface will be demonstrated by moderator. |
|---|---|---|---|
| 4. | **Intuitive design:**  
Transitions,  
Tooltips,  
colours. | No | This feature will not be evaluated. Main reason for having this feature is keeping user engaged and interested to the interface. But to reduce the cognitive load on participants I will avoid adding this features as much as possible. |
| 5. | **Keyword search:**  
To enable users to search the corpus for keywords. Non machine learning expert users will trust the model more if they can find the can compare results of the topic modelling with keyword search. | No | |
| 6. | Incorporating design requirements for **visually impaired** users. | No | I will not add this feature in evaluation design. |

Table 3.6: User interface requirements: one of the main deliverables of my project is an interactive user interface. This tabel summarise the mandatory and optional features of it. This list is prepared by ideas from literature survey of topic modelling visualisation is Section 2.6
Chapter 4

Project implementation

4.1 Introduction

I used incremental build approach, where main requirements and optional feature will be divided into sections and prototypes of the system will be developed and tested before adding next level of features.

This chapter will provide a very high level process of steps I took to complete this project. Figure 4.1 illustrates these steps which are:

1. Pre-processing data: takes an excel dataset, cleans data and makes it ready to use in MALLET topic modelling framework. Section 4.2 briefly explains this step and detailed information of this step is available in Appendix A;

2. Topic modelling: takes a collection of text files and produces 50 topics. Section 4.3 briefly explains this step and detailed information of this step is available in Appendix B;

3. Post-processing data: using output files from MALLET, calculates cosine similarity matrix and linkage table, then clusters 50 topics, order these clusters and then outputs a JSON file which will be used in implementing interface; Section 4.4 briefly explains this step and detailed information of this step is available in Appendix C;

4. Web interface implementation: first explains all prototypes developed for web interface before explaining chosen layout for evaluation; Section 4.5 briefly explains this step and detailed information of this step is available in Appendix D;

5. Evaluation: explains design of evaluation, pilot experiment and focus group. Section 4.6 briefly explains this step and detailed information of this step is available in Appendix E;
Figure 4.1: Main steps of the project. High level explanation will be given in this chapter and detailed information about each step in corresponding chapters.
4.2 Pre-processing data

My chosen dataset is the dataset of all grants awarded by Wellcome Trust from 1 October 2000 to 30 September 2016. Wellcome Trust is a funding agency in UK which supports research in the areas of public health and bio-medicine. This dataset can be obtained from Wellcome Trust website. This dataset is in the format of a excel file with 20,933 rows where each row contains information about one project. It requires pre-processing data not only to deal with missing/short abstracts and HTML tags, but also to make data ready for using in MALLET. Figure 4.2 shows main steps involved in pre-processing data for my project. More detailed information about this step can be found in Appendix A.

![Figure 4.2: Pre-processing data involves 3 main steps: 1. dealing with missing/short abstracts; 2. reducing size of the dataset; 3. extracting project abstracts and titles to separate text files in order to feed in MALLET (next step)](image)

4.2.1 Handling missing / short abstracts

One of the major issues of this dataset is missing values in project abstract. A possible solution for this issue is using project title as input to algorithm for these grants. Generally topic modelling algorithms assume that each document is a mixture of topics and each topic is probabilistic distribution over words. LDA discovers these distribution by applying statistical techniques to documents [Blei, 2011]. Therefore applying LDA on project titles which are very short texts will cause data sparsity issue [Hong and Davison, 2010a]. Data sparsity issue means algorithm don’t see enough word co-occurrences patterns to see how they are related.

Because of above mentioned issue I decided to exclude all grants that have missing abstracts. This step can be performed using any text editor but because of scale of dataset I used a data wrangling tool called ”OpenRefine”. Using regular expressions with OpenRefine reduced the time needed to clean this dataset. OpenRefine is a free and open source online tool, information about this tool can be found from OpenRefine website.
I excluded grants that had values like "No Data Entered", "Summary not available", "To be submitted" and blank from dataset. Table A.1 in Appendix A lists all values of project abstract that I excluded from dataset. At the end of this step my dataset was reduced from 20,933 to 15,504 grants.

### 4.2.2 Reducing the size of the dataset

With reduced dataset from Section 4.2.1 after running topic modelling algorithm I had to set number of topics to at least 30 to get intuitive topics\(^1\). This number is too much for the limited time and resources I have for evaluation. After consulting with my supervisor, I decided to reduce the size of the dataset.

Last field of the dataset is called "panel" which is Wellcome Trust’s internal deciding panels. There are 141 panels in original dataset. I used this field to reduce dataset. I needed to select a few panels that has the largest number of grants which generate around 10 topics. In order to choose which panels to use I wrote a python script which counts the number of grants for each panel. See Listing A.1 in Appendix A for details of this script.

I excluded panels that meet at least one of the following criteria from dataset, at the end of this step the number of panels were reduced from 141 to 31 with 9,901 grants.

1. **Number of grants is too little**: If the number of grants is less than 50 then that panel will be deleted;

2. **Panels that only fund PHD studentship application**: PHD grants are often don’t have full abstract and are often very general description of project which don’t create intuitive topics;

3. **Panels like "Medicine in Society" or "History of medicine"**: Majority of Documents in this panel only contribute to 1 or 2 topics (very general topics - these topics are always appear in every run of the model).

### 4.2.3 Making data ready for topic modelling with MALLET

MALLET, topic modelling framework, takes individual text document as input. I extracted project title and project abstract for each grant to a text file. This process is completed by writing a python script. This script is listed in Listing A.2 in Appendix A and does following steps:

1. **Replace HTML characters and tags**: Because I obtained dataset from Web, majority of grants had HTML tags which was treated here;

\(^1\)This steps will be explained in Section 4.3
2. **Convert all plurals to singular**: Topic modelling algorithm will consider "book" and "books" as two different words. This is a very important step because otherwise size of the vocabulary will be very large and generated topics will not be of good quality. Some topic modelling frameworks for example Gensim [Řehůřek and Sojka, 2010] have commands which does this step however current version of MALLET lacks this feature [McCallum, 2002]. Package WordNetLemmatizer converts plural words to their singular form;

3. For each grant extracts project title and project abstract to a text file.

At the end of this step I had 9,901 individual text file which I fed to MALLET.

### 4.3 Topic modelling

My chosen framework for topic modelling is MAchine Learning for Language Toolkit (MALLET) [McCallum, 2002]. MALLET is an open source, Java based framework which has a number of toolkit e.g. clustering, document classification, topic modelling, and etc. Code of MALLET also is available from Github. More detailed information about this step can be found in Appendix B.

MALLET toolkit implements a number of topic model algorithms. Based on project description topic modelling algorithm used for this project is Latent Dirichlet Allocation (LDA) [Blei, 2011]. MALLET is command based, I used 2 commands:

1. **import-dir**: is used to import text files to MALLET;

2. **train-topics**: this command is used 5 times with different random seeds to discover 10 topics in each run. Each on these runs will be considered as one model.

![Figure 4.3: Topic modelling step](image)

1. Import documents to MALLET and get .mallet file;
2. Repeat 5 times with random seed set to 1, 1001, 500, 600645, 999993:
   a. Train 10 topics each with 10 top words;
4.3.1 Latent Dirichlet Allocation (LDA)

I used LDA to produce topics which is the default topic modelling algorithm in MALLET. LDA is explained in Section 2.2.1 which briefly has following steps:

1. Initialise parameters: setting the number of topics, number of top words, number of iterations and hyper-parameters;

2. Randomly initialise topic assignment for each word in each document:

3. Repeat for specified number of iterations:

   (a) Re-sample topic assignment for each word in each document based on topic assignment of all other words and topic distribution for current document. This iterative step will generate coherent topics;

4. Get results

4.3.2 Import documents

Command import-dir is used to import text files to MALLET with following setting. 3 main options applied to this command are:

1. --skip-html: Also Python script in Section 4.2.3 delas with HTML tags, I also applied this option to import command. This option removes any text inside ¡...¿, as in HTML or SGML;

2. --remove-stopwords: remove a default list of common English "stop words" from the text. Stop words are words that don’t provide useful information about document, examples are punctuation (., ?, !) or functional words (a, an, and, the). This list can be found in Table B.3 of Appendix B.

3. --extra-stopwords FILENAME: to reduce the size of vocabulary and get better topics I used extra stop words on this dataset. This list contains two types of words:

   (a) General words related to dataset for example "wellcome", "trust", "fund";

   (b) All words that appears less than 8 times in the dataset. To create this list I ran the --train-topics command with default list of stop words. This command will be explained in Section 4.3.3. One output of this command is a file that lists all the words in the dictionary with frequency of it in topics. After this step my dictionary reduced from 56,367 to 9,911 words.
Figure 4.4: Import command of MALLET is used with options to treat HTML tags and remove stop words from corpus. Stop words are words that don’t provide useful information about document, examples are punctuation (, . ? !) or functional words (a, an, and, the).

4.3.3 Train topics

Main motivation of my project is to expose non-deterministic nature of topic modelling algorithm. Non-deterministic means results (solutions) of topic modelling changes by changing parameters or when a different algorithm from same family is used. I decided to keep the algorithm same and only change one parameter. My chosen parameter is random seed which corresponds to step 2 of LDA algorithm in Section 4.3.1. Each of these solutions is called a model throughout this report. I used train-topics command of MALLET for 5 times which produced 10 topics in each model. Setting used with this command are:

1. Input mallet file: it the output of import command from Section 4.3.2;

2. Number of topics: 10 topics;

3. Number of top words in each topic: 10 words;

4. Number of iterations: 1000 times; Algorithm will re-sample 1000 times the assignment of each word to topics;

5. Optimise interval: 20; MALLET website explains this option as: “This option turns on hyperparameter optimization, which allows the model to better fit the data by allowing some topics to be more prominent than others.”

6. Random seed: One of the steps of algorithm is to randomly assigns each word in dictionary to a topic. This setting is the only difference between different model,

7. Output files: main outputs of topic modelling algorithm topics, distribution of words over topics and distribution of documents over topics. MALLET has options to get more output files which I enabled and used them in my project, for example
I enabled \texttt{-topic-word-weights-file} option and used the output file to create topic-term weights which is used to calculate topic-topic similarity matrix in Section 4.4.1.

4.4 Post-processing data

In order to make output from topic modelling usable for visualisation I performed 4 steps on them to get desired output in the format of a JSON file. Code and detailed information about this step is available in Appendix C.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{post-processing.png}
\caption{Post-processing of results of topic modelling:
1. Computing topic to topic similarity matrix which is $50 \times 50$
2. Computing linkage table;
3. Computing agglomerative clustering using linkage table;
4. Sort topic clusters and models based on similarity.}
\end{figure}

4.4.1 Similarity matrix

There are several approaches to measure similarity between topics, majority of these approaches use compare topics using probability distribution of words in topics. For example, KL-divergence [Newman et al., 2009a], average Log Odds Ratio Chaney and Blei [2012], cosine similarity [He et al., 2009].

I calculated cosine similarity of topic to topic, to do so I wrote a Java package which does 3 main steps:
1. Convert Topic-Term weights to Java Matrix:
The class is called ReadExcelFile.java and takes topic-term weights in excel format and converts them to a Java Matrix. Input excel file is obtained from combining information from _topic-word-weights-file_ for all 5 models.

In Section 4.3.3 I explained that I trained 5 models each with 10 topics. One of the outputs of train-topics command is a file that contains weights of each word (term) of dictionary in each topic. I combined the information from this file in 5 models and created an excel file which has following format:

![Table](image)

**Figure 4.7:** Format of excel file which contains topic-word weights for all 5 models, 50 topics in total.
2. Compute cosine similarity:
The class is called `CosineSimilarity.java` and takes topic-term weights matrix and by applying following equation.

\[
similarity(A, B) = \frac{A \cdot B}{|A| \cdot |B|}
\] (4.1)

where:
\[
A \cdot B = \sum A_i \cdot B_i
\]
\[
|A| = \sqrt{\sum A_i^2}
\]
\[
|B| = \sqrt{\sum B_i^2}
\]
for \(i = [0..n-1]\),
where \(n\) = number of terms in topic-term matrix.
\(n\) in this project equals to 9,911

3. Prints topic-topic similarity matrix to a text file.
The output of this step is a matrix of size 50 X 50 (total number of topics in 5 models are 50).
Figure 4.10: A section of topic-topic similarity matrix. It is a matrix of size 50 X 50. Values are in range of [0, 1] and the higher the value the greater the similarity.

4.4.2 Sorting topic clusters and models

Clusters of topics are sorted based on similarity of all topics in cluster to average distance from topics of all remaining clusters. For example First cluster in the JSON dataset is the least distance to average of remaining 11 clusters, same process then identifies the second least different cluster among remaining 11 clusters and so on.
4.5 Web interface implementation

Based on project specifications I developed a web based interactive interface. This interface is web-based so I will be able to easily adopt it in order to conduct quantitative analysis to validate finding from focus group using crowd sourcing platforms. It is one of the required deliverable of this project and is relatively simple in terms of visualisation for three reasons:

1. In order to reduce cognitive load on participants so they can focus on main objective of this project which is to pick best model among multiple valid solutions of topics modelling;

2. The evaluation of this project will be a focus group so participants won’t be able to interact with interface directly;

3. In this project I’m laying the ground work by using D3.js\(^2\) programming language to develop the interface. Interfaces developed in D3.js easily can be made interactive and the intention is to incorporate this design rules and do more work in future;

4.5.1 Prototypes of interface

As I mentioned in Section 2.5.3 to best of my knowledge TopicCheck [Chuang et al., 2015] is the only scientific paper which addresses the issue of visualising stochastic results of topic modelling. Because of lack of enough resources to identify best layout for the task of the project I prototype a number of layouts and listed advantages and disadvantages of each layout in this section before selecting the best layout to implement.

\(^2\)D3 is a powerful java script library which is been used to visualise data from variety of data sources. It is open source and can be used with other web technologies like HTML, CSV and SVG to produce intuitive visualisations. more information about D3.js can be found in Data-Driven Documents.
4.5.1.1 Dendrogram

Dendrogram is a form of hierarchical tree layout which shows correlation between topics. All topics are listed at the bottom of the dendrogram and then clusters of topics are formed by iterative joining 2 most correlated topics. Main input to produce dendrograms is topic-topic similarity matrix.

Because this layout was the first prototype to see whether or not topics over models form clusters I needed to quickly form dendrogram so I used a freely available web server to do so. This website is available from DendroUPGMA.

**Advantages:** of this layout is that it is simple to understand and shows the clusters very well.

**Disadvantage:**

1. It is unable to show relationships between models, for example if user wants to see all topics of model 1 is not easily possible.
2. It is not scale-able, this dendrogram only shows 5 models with 10 topics in each. Increasing the number of models will make the layout very long which then will make it very difficult to analyse.

![Dendrogram](image)
4.5.1.2 Hexagons Layout

One of the most intuitive and visually appealing layouts is hexagons which was my second prototype. This layout is produced by applying hierarchical clustering algorithm on 50 topics. Clustering algorithm\(^3\) check 2 conditions:

1. The size of each cluster is less than or equal to number of models,
2. There is only one topic from each model in each cluster.

This layout works really well for showing one solution but is not very successful in displaying relationship between different solutions. Figure 4.14 shows using this layout on one solution of my project. Similarly Figure 4.15 shows five models. In this prototype each hexagon represent one topic, each island is one cluster and each colour is one model. Inside each hexagon top 3 words of that topic is displayed.

\[\text{Figure 4.13: High level schema rendering hexagons layout.}\]

---

\(^3\)The code used from clustering is donated by Mr. Pierre Le Bras. This code is then modified to accommodate different models. Unfortunately I can not publish his code due to his current research projects.
Figure 4.14: Hexagons layout is used to visualise one solution of topic modelling on Wellcome Trust dataset. Each hexagon is one topic and inside each topic top 5 words are displayed.

Figure 4.15: Hexagons layout: in this layout each hexagon is one topic, each island is one cluster and each colour is one model.

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy to spot the WORST model, in this case it will be the singleton</td>
<td>Hard to see relationship between models.</td>
</tr>
<tr>
<td>Intuitive and visually appealing</td>
<td>scale-ability issue with increased number of models</td>
</tr>
<tr>
<td></td>
<td>At first glance each island can be mistaken for one model where as it is one cluster of topics.</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of advantages and disadvantages of hexagons layout in visualising relationships between stochastic solutions
4.5.1.3 Tabular layout version 1

Version 1 is the layout I used for pilot experiment. In this layout I implemented both mandatory requirements for interface also added features of interactivity and intuitive design requirement from Table 3.6.

Figure 4.16: High level schema rendering hexagons layout.

Figure 4.17 illustrates an example transition feature implemented in version 1 of tabular layout. Here labels and colours of topic clusters are delayed for 2 seconds.

Example of tooltip feature implemented in version 1 of tabular layout. By hoovering over any topic top 10 words of the topic is displayed where the first word in the list is the word with highest probability.

Because the code is written in D3.js it will be very easy to modify this code or add more intuitive features to the layout should users ask for such features. This is because D3 is slightly different programming model in terms of General Update Pattern. GUP will
make everything more flexible. For example it will be very easy to put a button which when clicked will hide a column and re-size the layout to fit on the screen and get more details about topics.

Listing 4.1 highlights General Update Pattern.

function GUP_tiles()
{
    //GUP = General Update Pattern to render small multiples
    //GUP: BIND DATA
    var selection = svg
    .selectAll("rect")
    .data(dataset);

    //GUP Update 1
    selection
    .attr("border", tabColour)
    .style("stroke-width", border);

    //GUP: Enter selection
    //i.e. create necessary DOM elements if they don't already exist
    selection.enter()
    .append("g")

    //GUP Update2 (this selection = update1 + enter)
    //i.e. update all selected visual elements on page
    selection
    d3.select(this).selectAll("rect")
    .data(d2.nodes)
    .enter()
    .append("rect")
    .on("click", tileClick)
    .transition()
    .delay(function(d) { return (( i )) / dataset.length * 5000; })
    .duration(4000)
    .attr("width", tileWidth)
    .style("stroke", bordercolor);

    //GUP exit selection
    selection
    .exit()
    .remove();
};

Listing 4.1: General Update Pattern of D3.js GUP will make everything more flexible.
### Advantages

- Relationships between topics and models is very easy to understand
- Very scale-able in terms of adding more models and/or more topics in each model
- Simplicity of layout reduce cognitive load on user

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationships between topics and models is very easy to understand</td>
<td>Too simple so user might loose interest in interacting with it</td>
</tr>
<tr>
<td>Very scale-able in terms of adding more models and/or more topics in each model</td>
<td></td>
</tr>
<tr>
<td>Simplicity of layout reduce cognitive load on user</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Comparison of advantages and disadvantages of tabular layout in visualising relationships between stochastic solutions

Despite hexagons, tabular layout is very successful in visualising relationships between models and topics which makes it my chosen layout for final interface development. All codes used for this layout is included in Appendix D.

#### 4.5.1.4 Tabular layout - version 2

Based on outcomes of pilot experiment I changed tabular layout slightly. 3 main changes are:

1. Confusing labels of topic clusters: labels are not ordered. For focus group I renamed these labels.

2. Using tooltip to display top 10 words of each topic makes comparison of topics difficult. This feature is changed completely for focus group. The tooltip function is disabled and only top 3 words of each topic is displayed inside each square. Font size of top words are scaled to weight of that word in the topic.

3. Labels of topics are confusing. This labels are replaced by top 3 words of each topic.
Figure 4.19: Tabular web interface showing relationships between different models. Each row represent a model and each column is a cluster of topics. Topics in each column are the most similar topics across 5 models.

Each row represent a model and each column is a cluster of topics. Topics in each column are the most similar topics across 5 models. Columns are ordered from left to right based on similarity of all topics in column to average distance from topics of all remaining columns. For example First column is the least distance to average of remaining 11 columns, second column is the least different column to remaining 10 columns and so on. Similarly rows (models) are sorted from top to bottom. Top 3 words of each topic is displayed, the font of each word is scaled to it’s weight in the topic.
4.6 Evaluation

I evaluated my project in two stages:

1. Pilot experiment with one participant: aim of this experiment was to identify that my experiment fits the purpose and to detect floor or ceiling effect, floor effect is when the task is too simple for everyone and ceiling effect means it is very hard for everyone;

2. Focus group discussions: in Section 2.7.1.2 I explained focus group and it’s advantages and disadvantages. Main trade off in focus group is time versus low number of participants.

![Figure 4.20: Main steps of evaluation phase of my project are designing the experiment, recruiting participant, pilot study and focus group](image)

The process that I took to evaluate my project are listed in Table 4.6 and explained in following sections.

<table>
<thead>
<tr>
<th></th>
<th>Designing the Experiment</th>
<th>2.</th>
<th>Recruiting participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.</td>
<td>Pilot experiment</td>
<td>4.</td>
<td>Apply changes from pilot</td>
</tr>
<tr>
<td>5.</td>
<td>Conduct focus group</td>
<td>6.</td>
<td>Transcript recording</td>
</tr>
<tr>
<td>7.</td>
<td>Analysis transcripts and report</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6: Main steps of evaluation phase of my project
4.6.1 Designing the experiment

4.6.1.1 Main steps during experiment

<table>
<thead>
<tr>
<th>#</th>
<th>Step</th>
<th>Estimated time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consent Forms</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Presentation of project</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Demonstration of Interface</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Questionnaire part 1</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>Group discussion</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>Questionnaire part 2</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 4.8: Main steps during focus group experiment

4.6.1.2 Consent Forms

Each participant will be asked to read and sign a consent form. This form explains the project, what will happen, what sort of data will be collected and how data will be treated after the experiment. Then participant is asked to sign the form twice to agree not only voluntary consent but also voice recording the session.

4.6.1.3 Presentation of project

The aim of this presentation is to introduce my project and give minimum background information about dataset and topic modelling algorithm to participants. Slides of this presentation can be found in Appendix E.

4.6.1.4 Demonstration of interface

Demonstration is in 4 steps as below:

Step 1: Show only 1 model

I will present this interface to participants and explain topics and how the words are weighted. This interface is the typical interface shown to users. (Only one solution)

Figure 4.21: 10 Topics each showing weighted 5 top words.
**Step 2 : Show results of 2 models with slight differences**

Using same algorithm on the same dataset, same number of topics and words per topic. By only changing the RANDOM SEED, algorithm will result in slightly different results. Some of the Topics will stay stable for example Topic 2 in Figure 4.22, but change in Topic 4 is more obvious.

![Figure 4.22: Comparing 2 solutions of running LDA on Wellcome Trust dataset - Each solution(each row) displays 10 Topics each with weighted top 5 words.](image)

Row 1: is produced by running LDA Topic Modelling algorithm on Wellcome Trust grant funding dataset with random seed 1.

Row 2: is result of running same algorithm on same dataset with same setting using random seed 1001

**Step 3 : Show 2 models with obvious differences**

In previous step participants were presented 2 solutions with slight changes on some of the topics, in this step I will show them another solution with more drastic differences. As it can be seen from Figure 4.23:

1. some topics stay stable for example Topic 2.
2. some topics change slightly for example Topic 2 in which 3 words out of top 5 are different between these 2 rows.
3. 2 Topics disappeared completely (Topics 7 and 8). These topics are divided into other topics.
4. Because the number of topics are fixed (10 in my experiment), topic modelling algorithm found 2 new topics to replace missing ones (Topics 10 and 11).
Figure 4.23: 3 solutions - 10 Topics with weighted top 5 words.
Row 1: running LDA Topic Modelling algorithm on Wellcome Trust grant funding dataset with random seed 1.
Row 2: result of running same algorithm on same dataset with same setting using random seed 500.

**Step 4 : Show 5 models in one interface**
This interface shows 5 solutions among many available models. Participants will be asked to look carefully and try to make a note of which model they think is the WORST and which model is the BEST among these 5 models.

Figure 4.24: Tabular web interface showing relationships between different models. Each row represent a model and each column is a cluster of topics. Topics in each column are the most similar topics across 5 models.
4.6.1.5 Questionnaire - part 1

This questionnaire is designed using Google Forms⁴ and has 7 questions in total. It includes 3 types of questions:

1. Demographic questions for example age group and gender.

2. About the interface which has questions like selecting the BEST / WORST model. This question is part of assessment of hypothesis 3 of Table 3.2. This hypothesis is about whether a group of people will agree on one best solution.

3. Participant’s confidence on selected model. This question is part of assessment of hypothesis 2 of Table 3.2. Which is the visible consensus (general agreement) of other people in a group will increase person’s confidence on produced results in terms of both using them for planning purposes and explaining them to third parties.

1. Demographic questions
To keep this step very short I only included only 3 questions which can be seen on Figure 4.25.

1. participant’s level of knowledge about Machine Learning algorithms - 5 point likert scale.

2. Participant’s age group. He/She can refuse to answer to this question.

3. Participant’s gender. Again participant can select “Prefer not to say”.

⁴Google forms is a free platform, visit Google Forms for more information.
2. Selecting WORST and BEST model

Next 2 questions will ask user to select the WORST and BEST model among available 5 options. Please see Figure 4.26.

WORST model here is the model that participant will never select as BEST model. The one that he/she will discard at first. BEST model here is the one that participant will select.

Participant can only select 1 model as WORST but can select up to 2 models for BEST. The later is because on the interface two groups of 2 models are formed. Please see Figure 4.24.

Confidence of user on selected model

Next 2 questions are about participant’s confidence on selected model.

1. One of these 2 questions asks participants to rate their confidence in explaining their chosen BEST model to their line manager. Line manager here is a person who is in higher level in organisational ranking.

2. Last question of this part is to see weather participants are more confident to use this model in situations that they can share responsibility of their decision with other people.
4.6.1.6 Group discussions

Anonymity
Despite being recorded, I assure you that the discussion will be anonymous. The recording will be kept safely in a password protected laptop until they are transcribed word for word, then they will be destroyed.

Legal requirement about Wellcome Trust dataset
This analysis uses openly available grant funding information made publicly available by the Wellcome Trust. Wellcome have not been consulted in or have reviewed the subsequent analysis and interpretation of the data undertaken.”

Ground rules
1. Only one person speaks at a time;
2. There are no right or wrong answers;
3. I’m just as interested in negative comments as positive comments, sometimes negative comments are the most helpful;
4. When you do have something to say, please do so. There are many of you in the group and it is important that I obtain the views of each of you;
5. You do not have to agree with the views of other people in the group.

Questions
For each question I list explanation of it, reason for including it and also possible probes that I will use to get more information or involve all participants in the discussions. These questions will be used to start the conversations or guide participants in case they find it difficult to engage in the discussions.

Question 1 : Which model you chose as BEST model if you only had access to this interface?

• Explanation: Here BEST model is the one that participant is most likely to choose by just examining the interface;

• Reason: This is the opening question, participants will be able to answer this question easily as they already answered this question in part 1 of questionnaire. Please see Section 4.6.1.5 for details of question.

• Probes: this question don’t need any probes.
Question 2: How did you choose this model?

• Explanation: I will explain that they might feel their reasons are very basic and obvious, but they are important to me.

• Reason: based on answers to this questions I will see whether participants can agree on one BEST model or not. Also this question will assess their ability to explain their decision to others.

• Probes:
  1. "Would you explain further?"
  2. "Would you give an example?"
  3. "I don’t understand."
  4. "Does anyone else have some thoughts on that?"

Question 3: What are the advantages of this interface when selecting the BEST model?

• Explanation: I will display interface on screen for them and ask them to tell each other at least 1 advantage of the interface.

• Reason: Answers to this question will form guidelines for developing interfaces that aim in showing non-deterministic nature of machine learning algorithms.

• Probes:
  1. "Can you explain further?";
  2. "Does anyone else have some thoughts on that?";
  3. "X (X here is the name of the person who is not involved in discussion), do you agree with this?"

Question 4: What are the disadvantages of this interface?

• Explanation: For some people finding what they don’t like is easier than what they like. Same as question 3, I will display interface on screen for them and ask them to tell each other at least 1 advantage of the interface.

• Reason: Answers to this question will form guidelines for developing interfaces that aim in showing non-deterministic nature of machine learning algorithms.

• Probes:
  1. "Can you say more about that?";
  2. "Does anyone else have some thoughts on that?"
Question 5: 5. What suggestions do you have to improve this interface?

- **Explanation:** What extra features participants like to have.

- **Reason:** Answers to this question will form ”Nice to have” features for designing interfaces.

- **Probes:**
  1. "Can you explain how this feature will help you to select the BEST model?"
  2. "Does anyone else have some thoughts on that?"

4.6.1.7 Questionnaire- part 2

At the end of experiment participants will be asked to fill another set of questionnaire to assess weather group discussion changed their views and confidence or not. 4 questions out of 5 are same as part 1:

1. Selecting WORST model;
2. Selecting BEST model;
3. Rating self-confidence in explaining chosen models to self’s line manager;
4. Whether participant will use this interface in decision making situation;
5. Further comments and suggestions.

4.6.2 Recruiting participants

Because my evaluation is a focus group so the number of participants are low. Between 5 to 10 participants are the advised number for focus groups. I advertised my experiment over social media 2 weeks in advance of intended data for focus group. I received 10 responses within 2 days of advertisement. I selected one participant for pilot experiment and 6 for focus group.

4.6.3 Pilot experiment

After designing the experiment for focus group, I conducted a pilot study with only one participant one week prior to actual experiment. Aims of this study are listed below:

1. Is the allocated time for each step is realistic?;
2. To assess whether or not designed experiment fits the purpose ;
3. Can a non-machine learning expert understand the purpose of the study and questions?
4.6.4 Focus group discussions

The experiment took place in Library of Heriot-Watt University and all 6 participants attended the session. The whole session took around 70 minutes and all participants but one were actively involved in discussions which I tried to engage that participant by asking questions directly.

4.7 Summary

The hypothesis was that the solution which would be least different from all other solutions would be the one that people will pick. Considering this hypothesis after pre-processing data and running LDA on cleaned dataset multiple times (differently random-seeded) I ran hierarchical agglomerative clustering on the topics. Then used these clusters to visualise similarity of models. In order to identify most appropriate layout for interface I prototype multiple layouts and chose tabular layout because I needed an interface which is as simple as possible to reduce cognitive load on participant. The evaluation of my project was 2 fold: 1- pilot experiment and 2- focus group. Results of both of these experiments are available from Chapter 5.
Chapter 5

Results

5.1 Introduction

Results of pilot and focus group experiments are reported here in different sections.

5.2 Pilot experiment

This pilot experiment is conducted by one participant. It took 45 minutes to complete all tasks which confirms the appropriate length of the experiment. I had to spend more time explaining Topic Modelling with examples and use-cases, at the end participant were able to understand Topic Modelling and how it goes from document to topics and why results can change between different runs. Also participant was not a machine learning expert \(^1\), she was able to relate the use of Random Seed with the ”initial guess” used in some engineering software which she used during her studies.

5.2.1 Tabular layout

When interface with 5 models were displayed she requested to take control of laptop and investigate the interface herself. This interface is shown in Figure 5.1. I listed my observations from pilot here.

1. Starting from left side (dark blue tiles \(^2\)) she scrolled on each tile vertically reading out top 4-5 words to see how the words change in topics of each cluster. She spent approximately 5 minutes doing this.

2. In several occasions she said *10 words are just too much to make a mental note, 5 or even 4 is just enough.*

\(^1\)rated her level of knowledge 2 out of 5 in first question of part 1 questionnaire, for more details please see Section 4.6.1.5

\(^2\)Each small square represents one topic
3. She asked: *Why the labels of the groups are not ordered?* This question was about labels of topic clusters, I explained that I ordered them based on how similar are they to each other. Then she said: *It is confusing like this, can’t you just rename the labels?*

4. Next question was about label of each tile, she asked: *For example, why labels of squares in Topic 9 (she meant Topic cluster 9 - pink colour) are T4-M1, T3-M2 and so on?* I had difficulty explaining the reason without going too much in detail which was in turn confusing for her, at the end I just asked her to ignore those labels.

5. She easily selected last row of the interface as WORST model because of singleton topic.

6. To select BEST model she checked top words of Topics 9, 1, 6 and 5 of remaining four models. While she was doing this she said *I want to see the flow of words between topics.*

   (a) Also she said she wouldn’t role out any model based on visual representation, but finally she rolled out 3rd and 4th rows because of empty space.

   (b) Among remaining two models she chose the first row as BEST model just because it came first.

7. **Advantages of the interface:**

   (a) Using very distinctive colours for each topic cluster make it easy to navigate.

8. **Disadvantages of the interface:**

   She said:
(a) *There is no setting that I can change and modify, I like software that I can change settings and see how results change for myself.* This request relies on interactivity of layout.

(b) *It is very simple and is just like you made it with excel.*

At the end of experiment I showed her interface with hex layout and asked whether she prefers this one and how she would go and select BEST model? This information is documented in Section 5.2.3

### 5.2.2 Modifications applied on interface based on outcomes of pilot experiment

3 main outcomes of pilot in terms of visualisation features of interface were:

1. Confusing labels of topic clusters: labels are not ordered. For focus group I renamed these labels.

2. Using tooltip to display top 10 words of each topic makes comparison of topics difficult. This feature is changed completely for focus group. The tooltip function is disabled and only top 3 words of each topic is displayed inside each square. Font size of top words are scaled to weight of that word in the topic.

3. Labels of topics are confusing. This labels are replaced by top 3 words of each topic.

### 5.2.3 Hexagons layout

I only evaluated hexagons layout in pilot experiment due to her interest in more complex layouts. Limited time on focus group was the reason that I avoided to include this layout in experiment.

Hex layout are explained in Section 4.5.1.2 and a screen capture of the interface is displayed in Figure 5.2.

1. She liked this interface more and in her opinion it is more professional and intuitive.

2. She explained the process she would take to select the BEST model as below:

   (a) Role out the singleton - Model 3 in Figure 5.2,

   (b) Clusters of 3 topics: after excluding model 3, each cluster has 2 different models so move on to next size;

   (c) Clusters of size 4: are identical, so move on to clusters of size 5;

   (d) Between clusters of size 5 select the one that appears more in the centre. Model 1 here.
Chapter 5. Results

Model 1 - red colour: 3 times
Model 2 - green colour: 1 time
Model 4 - purple colour: 2 times
Model 5 - blue colour: 1 time.

Figure 5.2: Hexagons layout: in this layout each hexagon is one topic, each island is one cluster and each colour is one model. Participant clearly liked this layout more than tabular layout.

5.3 Focus group

5.3.1 Selecting Best Model

One of my research hypothesis is about whether or not a group of people will be able to select ONE BEST Model among available solutions by examining visual interface. To assess this question I asked participants to fill a small questionnaire before group discussion, then they explained how they chose it and finally at the end of the experiment they were asked to fill another questionnaire to assess whether group discussion changed their views or not.

Documenting "how" participants selected the BEST model and what was important for them gave me interesting insights which raised open-questions for future research. This findings are discussed in Section 5.4.

Also participants didn’t selected agreed on one BEST solution they all assessed quality of topics rather than relationships between them. They checked quality of topics by reading the top words of each topic to make sense of what it is about. They took different approaches to do so, Section 5.3.1.1 explains these approaches.
5.3.1.1 Process of selecting BEST model by participants

All participants carefully assessed quality of topics, they read the top words of each topic to make sense of what the topic is about but they used different methods to do so. These methods are explained in following paragraphs.

1. **Column-Wise by eliminating the stable topics**

   In this method participant started from left to right of the layout and for each column checked how stable the topic is in different models. This process is illustrated in Figure 5.3.1.1. For participant’s who used this method most important data was the ones that made differences between models in other words they believed topics that are stable are not important in decision making.

2. **Row-Wise by eliminating the WORST model**

   In this method participant started by finding the WORST model which is the one that has most differences to other models. All participants who used this model first eliminated Model 5 as it has one singleton topic (Topic 11).
3. Scoring quality of topics
One participant draw a grid similar to interface and scored quality of the each topic. This method also works column-wise but instead of eliminating stable topics, it give scores to each topic in all models. The better the topic the higher the score. He couldn’t finished scoring all 50 topics in 5 minutes and only decided based on topics he scored.

4. Dividing models based on column labels
One participant divided 5 models into two groups based on title of topics (columns). She named them as "GOOD" and "BAD" models.

1. **Good models**: Model 1 and 2, these two models have all topics from Topic 0 to Topic 9.

2. **Bad models**: Model 3, 4 and 5 which have at least one of Topic 10 or Topic 11.

She thought that second group (**Bad models**) had to work more because it had to find Topic 10 and Topic 11 where as first group (**Good models**) didn’t need to do that. This process is illustrated in Figure 5.3.1.1.
5.3.1.2 Comparison of selected BEST model before and after group discussion

After presentation and demonstration of the interface, participants had 5 minutes to select the BEST model in their view. Figure 5.7 shows which models were selected by participants before group discussion and Figure 5.8 shows how this distribution changed after group discussion.

**Figure 5.7:** Distribution of BEST model before the group discussion: Participants were allowed to select up to 2 models.

**Figure 5.8:** Distribution of BEST model after the group discussion: Participants were allowed to select up to 2 models.
There are 3 differences between these two figures.

1. **Model 1 lost 1 vote:** participant’s were allowed to choose up to two BEST models to see whether group discussion make them more confident on their selection or not. One of the two participant’s who chose 2 models before discussion only selected one model at the end of the experiment.

2. **Model 3 lost 1 vote:** main reason for this change was one participant noticed that this model has two similar topics, please see Figure 5.3.1.2. I explained to them that this interface only shows top 3 words for each topic and if they had the ability to check the rest of the words in the topic they would see the difference between these two topics. Only one person were convinced of this which kept his vote for model 3.

   This duplicate topics raises the issue about importance of selecting best number of top words to show in each topic.

3. **Model 2 gained 1 vote:** participant who discovered duplicate topics in model 3 changed his vote to model 2 instead of model 3.

![Figure 5.9](image)

**Figure 5.9:** Two similar topics in model 3: Top 3 words for these 2 topics are same and based on this this model lost one vote after discussions.

Other two models didn’t changed because during group discussions these two models mostly were ignored. Because model 5 was the WORST model in most participant view and model 1 was neither very good nor very bad.
5.3.1.3 Summary selecting BEST model

Table below summarises reasons and processes of selecting the best model.

<table>
<thead>
<tr>
<th>#</th>
<th>BEST</th>
<th>Reason</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3 &amp; 4</td>
<td>Quality of topics</td>
<td>Row-wise, started by eliminating the WORST model</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>words in topics are not repeated</td>
<td>Row-wise</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>All topics in Model 4 made sense</td>
<td>Draw a grid and gave each cell a mark based on quality of topics in his view</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Quality of topics</td>
<td>Worked column-wise and started by eliminating the stable topics.</td>
</tr>
<tr>
<td>5</td>
<td>1 &amp; 2</td>
<td>Quality of topics</td>
<td>Divided 5 models in to 2 groups based on title of topics (Topic0, Topic1, ... Topic11).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Group 1: Model 1 and 2;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Group 2: Model 3, 4 and 5.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>She thought that Group 2 had to work more because it had to find Topic 10 and Topic 11 where as Group 1 didn’t need to do that.</td>
</tr>
<tr>
<td>6</td>
<td>Model 2</td>
<td>Quality of topics and words in topics are not repeated</td>
<td>Column-wise, looked for repetitive words</td>
</tr>
</tbody>
</table>

Table 5.2: All participants selected BEST model in their view by assessing quality of topics in different ways, Nobody looked into visual display and the fact that topics and models are ordered from left to right and top to bottom.

5.3.2 Comparison of participants confidence on selected model before and after group discussion

I asked each participant to rate their confidence on their selected model before and after group discussion. I used 5 scale likert format for this question where 1 means "Not confident at all" and 5 is "Completely confident". The hypotheses that I evaluated with this question was that:
Chapter 5. Results

The visible consensus (general agreement) of other people in a group will increase person’s confidence on produced results in terms of both using them for planning purposes and explaining them to third parties.

Table 5.5 summarises individual responses for all participant, from, Figures 5.10, 5.11 it is obvious that confidence level is increased after group discussion. This increase can be explained for two reasons.

1. Because it was a focus group knowing how other people think and what model they chose increased people’s confidence,

2. Or alternatively group members agree with the group decision rather than individual decisions.

50% of participant felt more confident after discussing the process of selection with other group members this is a great finding but because of small sample size it requires follow-up quantitative experiments to validate the finding in terms of significance testing.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Confidence before</th>
<th>Confidence after</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 5.5: Self-rated confidence levels before and after the group discussion. 50% of participant felt more confident after discussing the process of selection with other group members.
5.3.3 Visual features of interface

5.3.4 Layout

Due to limited time on focus group session I did not present the hexagon layout to participants. But during pilot experiment I observed following points which needs more in-depth investigations to confirm.

1. Users prefer more complex layouts. Maybe because they look more professional and complicated.

2. On simple layouts users carefully read top words of each topic and investigate the quality of topics rather than deciding purely based on visual features.

3. If layout is more complicated for example hexagon layout, user spends more time analysing the layout visually. In pilot experiment, participant explained that she will first eliminate the singleton, then gradually check the clusters of size 3 and 4, finally she will choose the topic than appears most in the centre of hexagons as BEST model. She didn’t checked top words for topics and purely based her decision on layout.

5.3.5 Font size of top words

In the interface font size of each word in each topic is scaled to weight of that word in the topic. This weight can also be explained as normalised frequency of each words in the topic. One participant thought the font size shows the ability or confidence of the algorithm to find that word, I explained the process algorithm takes to find topics with an example.

There were disagreement between participant about whether scaling the font sizes is useful or not. Some participant’s argued that having different font sizes is confusing and if I was not there to explain what it is they wouldn’t noticed that it’s weight of the word. On the other hand others believed that it is a useful feature. One participant explained this with an example from interface to others.

*I think the font is important because you can have two different modelling with same words but one of them will pick “gene” for example in bigger font, so in that model “gene” is quite important. So I think the differentiation can be useful.*
5.3.6 Empty spaces

Empty spaces raised lots of questions for participants. Having empty spaces is unavoidable in tabular layouts when clustering algorithm is applied to results. The clustering algorithm will check following conditions:

1. Number of topics in each cluster is less than or equal to number of models (here 5 models).
2. Each cluster can have at most one topic from each model.

Based on this observations I believe having an information box (for example using a tooltip) to explain them will be a useful feature to implement.

5.3.7 Interaction

Only one participant wanted to check the interaction features on interface, he wanted to know what happens when I click on topic or hover over it.

5.4 Summary of results

I ran two experiments to evaluate my project: pilot experiment and a focus group, I divided my findings from these two experiments in two sections:

1. Answers to my research hypotheses which then I will drive open-questions from them for future research on this subject. All of these finding are subject to caveats of focus group experiment in terms of small sample size and also effect of dominant participants in the group. Ideally all of these findings will be experimented in more depth in order to see whether they generalise to the population of not.

2. Recommendations table for designing web interfaces to visualise stochastic solutions of machine learning algorithms. Such interfaces will serve as a tool in decision making situations where user needs to select a solution to use.
5.4.1 Findings and open-questions for future work

I specified my research hypotheses in Section 3.3 which I have listed here for ease of reading this report.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exposing non-deterministic (stochastic) nature of topic modelling will reduce confidence of people on produced results in terms of both using them for planning purposes and explaining them to third parties.</td>
</tr>
<tr>
<td>2</td>
<td>The visible consensus (general agreement) of other people in a group will increase person’s confidence on produced results in terms of both using them for planning purposes and explaining them to third parties.</td>
</tr>
<tr>
<td>3</td>
<td>Solution which is least different from all other solutions in terms of similarity between topics will be the one that people will pick.</td>
</tr>
</tbody>
</table>

Table 5.7: Research hypotheses: I evaluated these hypotheses during pilot and focus group experiment.

1. Due to nature of focus group and experiment design I was not able to evaluate first hypothesis but nearly all participants were confused at first about this fact and none of them were aware of this issue.

2. In terms of second hypothesis 50% of participants felt more confident after discussing the process of selection with other group members. This increase can be explained for two reasons.

   (a) Because it was a focus group knowing how other people think and what model they chose increased people’s confidence,

   (b) Or alternatively group members agree with the group decision rather than individual decisions.

3. I derived third hypothesis from assumption of the project specification provided by my supervisor which was the similarity of topics is more important than the quality of the topics. Based on this assumption I based my hypothesis and interface design on relationships between topics. But results of both experiments suggest "quality of topics where more important to people than relationships". All participants assessed quality of topics in order to pick best model.

One topic for future work will be instead of sorting topics by commonality of topics and clustering of topics use a different technique for example explicitly assess topic quality and design interface based on topic quality. Palmetto is an API which calculates the semantic topic quality can be used to perform a feasibility study on this open-question. Palmetto for non-commercial uses is licensed under LGPL v3.0 License
### 5.4.2 guidelines for designing interface

The table below contains design guidelines from focus group and pilot experiment. These are only guidelines and needs to be tailored to users who will use the interface. These guidelines are divided into 3 groups: layout, extra information and general visual features.

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>Observations and Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Layout</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Option to toggle layouts between tabular and hexagon layout. See Section 4.5.1.2 for more information about hexagon layout.</td>
<td>Due to limited time in focus group Hexagon layout was only presented in pilot experiment. She clearly liked Hexagon layout better than tabular layout. I explained to her that displayed models are only 5 out of many models and in real application user would need to examine more than 5 models so this will make Hexagon layout confusing. She suggested to have option to stitch between layouts.</td>
</tr>
<tr>
<td><img src="image.png" alt="Diagram" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Extra information</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Hide/show a column or a row</td>
<td>Specially participants who used column-wise method to select the BEST model requested this feature. Having this option will make it easier for them to focus on topics that make difference between models. By scaling remaining topics/models so it shows more information will be an added bonus. For example increasing the number of displayed top words for remaining topics.</td>
</tr>
<tr>
<td>3</td>
<td>Ability to highlight all occurrences of a word</td>
<td>One participant suggested to have all occurrences of each word with same colour. I believe having too many colours in the layout will be confusing instead an option to highlight a word by clicking or hovering on it will be a useful feature.</td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>4</td>
<td>Display similarity between two adjacent topics in each row.</td>
<td>Participant suggested display this information in tooltip when hovered over separating line between two topics. However I believe similarity between topics of each column will be more useful than each row. By having column-wise similarity user will be able to eliminate stable topics with more confidence, and then spend more time on the rest of the topics. For column-wise similarity the opacity of background colour can be adjusted to this measure. By doing this if all tiles of a column had same opacity then that column is a stable topic. On the other hand if opacity changes a lot then that topic needs more detailed investigation.</td>
</tr>
</tbody>
</table>
**Chapter 5. Results**

Information about top documents contributing to each topic.

Regarding duplicate topics in model 3 (which has 2 topics with same top 3 words), I asked participants whether they like to see detailed information about these 2 topics? for example when clicked have top 10 words displayed or even have an option to open top 5 documents that contributed to each of these topics. An example of this information which is implemented in dfbrrowser. is displayed here.

Answer to my question was a definite "NO", One participant explained this answer and also explained when he would like to see more information.

"For me if I see a repeated topic I just discard that model. In my view an algorithm is not a good one when it can't decide that these are same topics so they shouldn't be divided. But when I'm choosing between two best ones I like to see more information. I like to be able to click on the topics and read documents."

**General visual features**

| Labelling rows (models) or a line/space between rows | All participants mentioned lack of model labels as a disadvantage of interface. One participant clearly stated that this was the "most confusing" aspect of interface for her. |
**Colour coding:**
Using colours to differentiate either models or topics. Enabling user to choose the colour coding method will be a useful feature to have in the interface. This feature can be implemented using radio buttons.

Participant’s who worked column-wise to select BEST model preferred to have different colours for each column, similar to my interface. But people who took the row-wise approach preferred to have different colours for each model. Please see Section 5.3.1.1 for more information on column-wise and row-wise methods for selecting the BEST model.

![Diagram 1](image1.png)

Colour coding each topic (each column), this method works best for people who select the BEST model by eliminating stable topic.

![Diagram 2](image2.png)

This way of colour coding the interface will work best for people who work on each model and try to eliminate the WORST model first.
**Font size:**
Enabling user to choose how to display the weight of the word in the topic, 3 available options are:

1. Scaling font size of each word to its weight in the topic. This option is implemented in the layout.

2. Having all words with same font size and displaying the frequency of each word next to it.

3. Displaying top words using barchart where the height of the bar is scaled to weight / frequency of the word in the topic.

As explained in Section 5.3.5 some participants were confused about different font sizes. They preferred to have all words in same font and either have the frequency of the word next to it or display top words using barcharts. Using barcharts were also implemented by [Chuang et al., 2015]. Figures below shows mock-ups of these two methods.

![Mock-up of barchart method](image1)

![Mock-up of layout method](image2)

**Using bold font for title of the columns (topics)**

All participants except one wanted to have titles of columns in bold and in one colour. The person who didn’t agree with rest of the group explained that he doesn’t want his attention at first glance.

**Legend / information box for interface**

All participants agreed that this interface needs a legend or information box which explains what is it about, what are the rows and columns. I recommend to include a small information box to the interface with minimum information which has links to more detailed information.

| Table 5.9: Table of recommendations on designing web interfaces where users need to select one BEST solution. These features will help different users to customise the interface based on how they process the displayed information. |
Chapter 6

Conclusion

Many machine learning algorithms produce stochastic output which will affect credibility of the them and users often resist basing their decisions on solutions from such algorithms not only because they don’t trust the results but also because they can’t defend their decision to other stakeholders. Therefore in order to increase the trust of users, firstly they need to be aware of this fact which often they don’t get informed of, and secondly they need to be given confidence of choosing one solution.

Based on project specification that assumed similarity of topics was more important than the quality of the topics I designed my interface and experiments considering relationships between topics rather than the quality of topics. To visualise these relationships I used hierarchical agglomerative clustering on topics and ordered not only topic clusters but also models based on how similar they are to all other models and clusters.

Because of lack of enough resources to identify best layout I have done in-depth evolution of layouts before choosing simple tabular layout. Simplicity of layout has three advantages: firstly people will be more confident when they can easily understand the layout and the relationship between object on interface, secondly the cognitive load on user will be reduced, and thirdly it is scale-able in terms of both increasing the number of models and topics.

I have conducted in-depth pilot experiment and focus group to evaluate my project. Outcomes of these two experiments which are subject to usual caveats of focus group such as small number of participants are two-fold: first design recommendations for visualising non-deterministic solutions, and second gathering preliminary data to highlight some open-questions for future work.
Focus group experiment proved that the visible consensus of group members will increase person's confidence on produced results in terms of both using them for planning purposes and explaining them to third parties. 50% of participant felt more confident at the end of focus group. This increase can be because knowing how other people think and what model they chose increase one’s confidence, or alternatively group members agree with the group decision rather than individual decisions.

Surprising finding was that all participants of both pilot and focus group experiments assessed quality of topics rather than their relationships. This finding opens one question for future work which is instead of sorting topics by commonality of topics and clustering of topics use a different technique for example explicitly assess topic quality and design interface based on topic quality. Palmetto is an API which calculate the semantic topic quality can be used to perform a feasibility study on this open-question.

I also derived a design recommendation for web interfaces intended for visualising stochastic results of machine learning systems. Main points are:

1. Users preferred some level of complexity of layout in terms of visualisation, and extra features and information which was fundamental to comparing the topics and assessing quality of them. For example hide/show option for both columns and rows.

2. In more complex layouts for example hexagons participant were tempted to base his/her decision on visualisation features.

3. Quality of topics were associated with the top words displayed on the screen which highlights the importance of selecting best number of top words to display.
Appendix A

Pre-processing data

A.1 Introduction

In Section 2.4.6 I introduced my chosen dataset which is the dataset of all grants awarded by Wellcome Trust from 1 October 2000 to 30 September 2016. Wellcome Trust is a funding agency in UK which supports research in the areas of public health and biomedicine. This dataset can be obtained from Wellcome Trust website. This dataset is in the form of excel spreadsheet and includes information about 20,933 grants in total. Main input to topic modelling tool will be "project title" and "project abstract". Some of the fields in this dataset have missing data and some fields are redundant so pre-processing of this dataset is required.

A.2 Handling missing or very short abstracts

As Mentioned in Section 2.4.6 one of the major issues of this dataset is missing values in project abstract. One possible solution for this issue is using project title as input to algorithm for these grants.

Generally topic modelling algorithms assume that each document is a mixture of topics and each topic is probabilistic distribution over words. LDA discovers these distribution by applying statistical techniques to documents [Blei, 2011]. Therefore applying LDA on project titles which are very short texts will cause data sparsity issue [Hong and Davison, 2010a]. Data sparsity issue means algorithm don’t see enough word co-occurrences patterns to see how they are related.

Because of above mentioned issue I decided to exclude all grants that have missing abstracts. This step can be performed using any text editor but because of scale of dataset I used a data wrangling tool called "OpenRefine". Using regular expressions with OpenRefine reduced the time needed to clean this dataset. OpenRefine is a free and open source online tool, information about this tool can be found from OpenRefine website.
Table A.1 lists all values of project abstract that I excluded from dataset. Figure A.1 shows that there are 5,066 grants with "No Data Entered" in their project abstract field. At the end of this step my dataset was reduced from 20,933 to 15,504 grants.

<table>
<thead>
<tr>
<th>#</th>
<th>Value</th>
<th>#</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No Data Entered</td>
<td>10</td>
<td>PhD Studentship proposal</td>
</tr>
<tr>
<td>2</td>
<td>Not Applicable</td>
<td>11</td>
<td>Research Resources Scoping award</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>12</td>
<td>Research Resources in Medical History award</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>13</td>
<td>No Summary provided for the Web</td>
</tr>
<tr>
<td>5</td>
<td>Not available</td>
<td>14</td>
<td>Value In People Award</td>
</tr>
<tr>
<td>6</td>
<td>No image</td>
<td>15</td>
<td>Awaiting Revised Summary</td>
</tr>
<tr>
<td>7</td>
<td>blank</td>
<td>16</td>
<td>Currently not available</td>
</tr>
<tr>
<td>8</td>
<td>Summary not available</td>
<td>17</td>
<td>To be submitted later</td>
</tr>
<tr>
<td>9</td>
<td>VIPA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.1: All grants with these values for project abstract were excluded from dataset.

Figure A.1: Interface of OpenRefine which is used to handle missing / short abstracts. In this example the value "No Data Entered" has 5066 matching rows out of 20,933.

A.3 Reducing size of dataset

With reduced dataset from Section A.2 after running topic modelling algorithm I had to set number of topics to at least 30 to get intuitive topics\(^1\). This number is too much for the limited time and resources I have for evaluation. After consulting with my supervisor, I used "panel" field to reduce the dataset. Last field of the dataset is called

\(^1\)This steps will be explained in Section B.3 of Appendix B
"panel" which is Wellcome Trust’s internal deciding panels. There are 141 panels in original dataset. I decided to choose grants from panel or panels that produce around 10-15 topics. Preferably a few panels that has the largest number of grants. In order to choose which panels to use I wrote a python script which counts the number of grants for each panel. See Listing A.1 for details of this script.

Listing A.1: Python script used to count the number of grants in each panel. In order to produce intuitive topics I have to reduce the size of dataset. To do so I will choose a few panels with highest number of grants.

I excluded panels that meet at least one of the following criteria from dataset.
1. **Number of grants is too little**: If the number of grants is less than 50 then that panel will be deleted;

2. **Panels that only fund PHD studentship application**: PHD grants are often don’t have full abstract and are often very general description of project which don’t create intuitive topics;

3. **Panels like ”Medicine in Society” or ”History of medicine”**: Majority of Documents in this panel only contribute to 1 or 2 topics (very general topics - these topics are always appear in every run of the model).

At the end of this step the number of panels were reduced from 141 to 31 with 9,901 grants.

### A.4 Making data ready for Topic modelling

1. Extracting project title and abstract
2. explain the python script
3. link to appendixes where the scripts are..

MALLET, topic modelling framework, takes individual text document as input. I extracted project title and project abstract for each grant to a text file. This process is completed by writing a python script. This script is listed in Listing A.2 and does following steps:

1. **Replace HTML characters and tags**: Because I obtained dataset from Web, majority of grants had HTML tags which was treated here;

2. **Convert all plurals to singular**: Topic modelling algorithm will consider ”book” and ”books” as two different words. Some topic modelling frameworks for example Gensim [Řehůřek and Sojka, 2010] have commands which does this step however current version of MALLET lacks this feature [McCallum, 2002]. Package WordNetLemmatizer converts plural words to their singular form;

3. For each grant extracts project title and project abstract to a text file.

At the end of this step I had 9,901 individual text file which I fed to MALLET.

#This script takes an excel file as input and generates separate text files
#from content of 2 columns of each row
#author: Azimeh Gharavi 10/05/2017

from __future__ import unicode_literals
import xlrd
import os.path
from pattern.text.en import singularize
from nltk import word_tokenize
from nltk.stem.wordnet import WordNetLemmatizer

#replace html characters and tags..
def ReplaceString(stringToCheck):
    replaceWithSpace = ["&ldquo;", "&rdquo;", "&nbsp;","<br>"","<\br>"","</p>"","<p>"","&curren;"
    replaceWithHyphen = ["­","—","–"]
    for item in replaceWithSpace:
        stringToCheck = stringToCheck.replace(item," ")
    for item in replaceWithHyphen:
        stringToCheck = stringToCheck.replace(item,"-")
    return stringToCheck

#convert all plurals to singular form
def convertToSingular(stringToCheck):
    lmtzr = WordNetLemmatizer()
    singularized_tokens=[]
    tokenized_text = word_tokenize(stringToCheck)
    for word in tokenized_text:
        singularized_tokens.append(lmtzr.lemmatize(word))
    singularized_text = ' '.join(singularized_tokens)
    return singularized_text

# Open the workbook and select the first worksheet
wb = xlrd.open_workbook('D:\MSC_Project\Step 2 - convert to JSON\python\wellcom-test-with20rows.xlsx')
sh = wb.sheet_by_index(0)
#path to save txt files
save_path = 'D:/MSC_Project/text/

# Iterate through each row in worksheet and fetch values
for rownum in range(1, sh.nrows, 1):
    row_values = sh.row_values(rownum)
    #each txt file will have the "Grant Reference" as it's name
    #replaces "/" with "-" as filename cannot have "/
    file_name = row_values[2].replace("/","-")
    #getting content of tct file from Project_title and Project_abstract
    p_title = ReplaceString(row_values[10])
    p_title = convertToSingular(p_title)
    p_abstract = ReplaceString(row_values[11])
    p_abstract = convertToSingular(p_abstract)
    text_content = p_title + " " + p_abstract
    completeName = os.path.join(save_path, file_name + "\"".txt")
    file1 = open(completeName, "w")
    file1.write(text_content.encode("utf-8"))
    file1.close()
Appendix B

Topic Modelling with MALLET

B.1 Introduction

Topic modelling algorithm used for this project is Latent Dirichlet Allocation (LDA) [Blei, 2011]. One of the downsides of topic modelling LDA is that the number of topics is fixed and selected by user. As I explained in Section 2.2.1, choosing the best number of topics has direct impact on success of the model, if this number is too little then topics will be too general, which in turn won’t enable user to extract useful information. On the other hand if it is too large then results will be too complicated to interpret. Teh et al. [2006] attempts to automatically select best number of topics by incorporating Dirichlet process to the algorithm. Hierarchical topic model [Griffiths and Tenenbaum, 2004] address this issue by assuming that number of topics are infinite and orders them in a hierarchy. Abstract and high level topics are located near the root of tree and more solid and detailed topics are at the leaves. These extensions to LDA improved the accuracy of the algorithm, however among non-machine learning experts try-and-error is the most widely used approach to determine the number of topics. I took the same approach and considering limited time and resources I have for evaluation of project, decided to have 10 topics.

B.2 Import data to MALLET

Command import-dir is used to import text files to MALLET with following setting.

1. import-dir: loads the contents of a directory into mallet instances;

2. –input FILE-NAME: path to text files directory;

3. –output FILE-NAME: path to output .mallet file;

4. –skip-html: removes any text inside ¡...¿, as in HTML or SGML.;
5. \texttt{--keep-sequence}: output data will be a FeatureSequence rather than a FeatureVector. Current version of topic modelling in MALLET only supports FeatureSequence;

6. \texttt{--remove-stopwords}: remove a default list of common English "stop words" from the text.

Stop words are words that don’t provide useful information about document, examples are punctuation (, . ? ! ) or functional words (a, an, and, the). This list can be found in Table B.3 of Appendix B.

7. \texttt{--extra-stopwords FILE-NAME}: to reduce the size of vocabulary and get better topics I used extra stop words on this dataset. This list contains two types of words:

(a) General words related to dataset for example "wellcome", "trust", "fund";
(b) All words that appears less than 8 times in the dataset. To create this list I ran the \texttt{--train-topics} command with default list of stop words. This command will be explained in Section B.3. One output of this command is a file that lists all the words in the dictionary with frequency of it in topics. See Figure B.2. After this step my dictionary reduced from 56,367 to 9,911 words.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{import-dir.png}
\caption{Import-dir command of MALLET. This command imports add text files from specified directory and after removing stop-words produces .mallet file.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{freq.png}
\caption{One of the output files of MALLET. This file lists all words in the dictionary with frequency of happening in topics. In other words it is the distribution of each word over topics.}
\end{figure}
B.3 Train topics

`train-topics` is the command that is used to find topics. Figure B.3 shows model settings, output files and also results of running the command.

1. `train-topics`: trains a topic model from .mallet file;
2. `--input FILENAME`: this file is the output of `import-dir` command;
3. `--num-topics 10`
4. `--num-top-words 10`: each topic is the distribution over words in the dictionary, this command specifies the number of top words in each topic;
5. `--num-iterations 1000`: number of iterations that algorithm re-samples assigned topics for each word, See Section 2.2.1 for more details;
6. `--random-seed 1`: one of the steps of algorithm is to randomly assigns each word in dictionary to a topic. This setting is the the only difference between different model,
7. `--optimize-interval 20`: Mallet website explains this option as: “This option turns on hyperparameter optimization, which allows the model to better fit the data by allowing some topics to be more prominent than others. Optimization every 10 iterations is reasonable”;
8. `--output-state FILE-NAME`: Its a compressed file which contains all words in the corpus and their assigned topics.
Appendix B. Topic Modelling with MALLET

9. **–output-topic-keys FILE-NAME**: This file lists 10 topics each with 10 words;

10. **–output-doc-topics FILE-NAME**: distribution of topics in each document;

<table>
<thead>
<tr>
<th>#</th>
<th>File name</th>
<th>Weight in Topic 0</th>
<th>Weight in Topic 1</th>
<th>...</th>
<th>Weight in Topic 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>file0.txt</td>
<td>4.7061919575E-4</td>
<td>5.487427601E-4</td>
<td>...</td>
<td>0.3654651447</td>
</tr>
<tr>
<td>1</td>
<td>file1.txt</td>
<td>4.7061919571E-4</td>
<td>0.5203634343</td>
<td>...</td>
<td>5.748853170E-4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Table B.1**: Format of the –output-doc-topics file. This file contains distribution of topics for each document.

11. **–word-topic-counts-file FILE-NAME**: for each word in the dictionary writes a sparse frequency of in in topics. This file is used in Section B.2 to create the extra stop words list;

<table>
<thead>
<tr>
<th>word</th>
<th>Topic</th>
<th>Frequency of &quot;costs&quot; in topic 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>costs</td>
<td>14: 4</td>
<td>16: 3 19: 1 8: 1</td>
</tr>
<tr>
<td>day</td>
<td>14:138</td>
<td>7:105 19:67 0:31 1:30 11:16 12:14</td>
</tr>
<tr>
<td>cutline</td>
<td>1:22 7: 13 6:8 16:5 8:4</td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>1:3087 9:2798 14:534 6:372 16:227 15</td>
<td></td>
</tr>
<tr>
<td>programme</td>
<td>9:945 1:245 7:214 15:170 19:144 0:9</td>
<td></td>
</tr>
<tr>
<td>welcome</td>
<td>9:861 14:79 19:71 7:67 12:43 1:1</td>
<td></td>
</tr>
</tbody>
</table>

**Figure B.4**: This file lists all words in the dictionary with frequency of happening in topics. In other words it is the sparse distribution of each word over topics.

12. **–topic-word-weights-file FILE-NAME**: un-normalised weights of all words in the dictionary in each topic;

13. **–output-topic-docs**: lists most prominent documents in each topic; which to write measures of topic quality, in XML format.

14. **–xml-topic-report**: lists top words for each topic with Dirichlet parameters in XML format. Figure B.5 shows part of this file.
Appendix B. Topic Modelling with MALLET

10. \texttt{–xml-topic-report} command outputs a file in which for each topic lists top words and Dirichlet parameters.

15. \texttt{–diagnostics-file FILE-NAME}: measures of topic quality, in XML format.

16. MALLET LDA, 10 topics: confirms topic modelling algorithm and number of topics;

17. \{10\}_i, \{20\}_i, \ldots \}: number of iterations;

18. 0, 1, 2, \ldots \}: topic number;

19. model, study, brain,\ldots \}: top 10 words in topic 0 after 40 iterations.

B.4 Default stop words list

Stop words are words that do not give any useful information about document, examples are punctuation (,. ? !) or functional words (a, an, and, the).

\begin{center}
\begin{tabular}{|c|c|c|c|c|}
\hline
a & able & about & above & according \\
\hline
accordingly & across & actually & after & afterwards \\
\hline
again & against & all & allow & allows \\
\hline
almost & alone & along & already & also \\
\hline
although & always & am & among & amongst \\
\hline
an & and & another & any & anybody \\
\hline
anyhow & anyone & anything & anyway & anyways \\
\hline
anywhere & apart & appear & appreciate & appropriate \\
\hline
are & around & as & aside & ask \\
\hline
asking & associated & at & available & away \\
\hline
awfully & b & be & become & because \\
\hline
became & becomes & becoming & been & before \\
\hline
beforehand & behind & being & believe & below \\
\hline
beside & besides & best & better & between \\
\hline
beyond & both & brief & but & but \\
\hline
\end{tabular}
\end{center}
<table>
<thead>
<tr>
<th>by</th>
<th>c</th>
<th>came</th>
<th>can</th>
<th>cannot</th>
</tr>
</thead>
<tbody>
<tr>
<td>cant</td>
<td>cause</td>
<td>causes</td>
<td>certain</td>
<td>certainly</td>
</tr>
<tr>
<td>changes</td>
<td>clearly</td>
<td>co</td>
<td>com</td>
<td>come</td>
</tr>
<tr>
<td>comes</td>
<td>concerning</td>
<td>consequently</td>
<td>consider</td>
<td>considering</td>
</tr>
<tr>
<td>contain</td>
<td>containing</td>
<td>contains</td>
<td>corresponding</td>
<td>could</td>
</tr>
<tr>
<td>course</td>
<td>currently</td>
<td>d</td>
<td>definitely</td>
<td>described</td>
</tr>
<tr>
<td>despite</td>
<td>did</td>
<td>different</td>
<td>do</td>
<td>does</td>
</tr>
<tr>
<td>doing</td>
<td>done</td>
<td>down</td>
<td>downwards</td>
<td>during</td>
</tr>
<tr>
<td>e</td>
<td>each</td>
<td>edu</td>
<td>eg</td>
<td>eight</td>
</tr>
<tr>
<td>either</td>
<td>else</td>
<td>elsewhere</td>
<td>enough</td>
<td>entirely</td>
</tr>
<tr>
<td>especially</td>
<td>et</td>
<td>etc</td>
<td>even</td>
<td>ever</td>
</tr>
<tr>
<td>every</td>
<td>everyone</td>
<td>everybody</td>
<td>everything</td>
<td>everywhere</td>
</tr>
<tr>
<td>ex</td>
<td>exactly</td>
<td>example</td>
<td>except</td>
<td>f</td>
</tr>
<tr>
<td>far</td>
<td>few</td>
<td>fifth</td>
<td>first</td>
<td>five</td>
</tr>
<tr>
<td>followed</td>
<td>following</td>
<td>follows</td>
<td>for</td>
<td>former</td>
</tr>
<tr>
<td>forth</td>
<td>four</td>
<td>from</td>
<td>further</td>
<td>furthermore</td>
</tr>
<tr>
<td>g</td>
<td>get</td>
<td>gets</td>
<td>getting</td>
<td>given</td>
</tr>
<tr>
<td>gives</td>
<td>go</td>
<td>goes</td>
<td>going</td>
<td>gone</td>
</tr>
<tr>
<td>gotten</td>
<td>greetings</td>
<td>h</td>
<td>had</td>
<td>happens</td>
</tr>
<tr>
<td>hardly</td>
<td>has</td>
<td>have</td>
<td>having</td>
<td>he</td>
</tr>
<tr>
<td>hello</td>
<td>help</td>
<td>hence</td>
<td>her</td>
<td>here</td>
</tr>
<tr>
<td>hereafter</td>
<td>hereby</td>
<td>herein</td>
<td>hereupon</td>
<td>hers</td>
</tr>
<tr>
<td>herself</td>
<td>hi</td>
<td>him</td>
<td>himself</td>
<td>his</td>
</tr>
<tr>
<td>hither</td>
<td>hopefully</td>
<td>how</td>
<td>howbeit</td>
<td>however</td>
</tr>
<tr>
<td>i</td>
<td>ie</td>
<td>if</td>
<td>ignored</td>
<td>immediate</td>
</tr>
<tr>
<td>in</td>
<td>insasmuch</td>
<td>inc</td>
<td>indeed</td>
<td>indicate</td>
</tr>
<tr>
<td>indicated</td>
<td>indicates</td>
<td>inner</td>
<td>insofar</td>
<td>instead</td>
</tr>
<tr>
<td>into</td>
<td>inward</td>
<td>is</td>
<td>it</td>
<td>its</td>
</tr>
<tr>
<td>itself</td>
<td>j</td>
<td>just</td>
<td>k</td>
<td>keep</td>
</tr>
<tr>
<td>keeps</td>
<td>kept</td>
<td>know</td>
<td>knows</td>
<td>known</td>
</tr>
<tr>
<td>l</td>
<td>last</td>
<td>lately</td>
<td>later</td>
<td>latter</td>
</tr>
<tr>
<td>latterly</td>
<td>least</td>
<td>less</td>
<td>least</td>
<td>less</td>
</tr>
<tr>
<td>lest</td>
<td>let</td>
<td>like</td>
<td>liked</td>
<td>likely</td>
</tr>
<tr>
<td>little</td>
<td>look</td>
<td>looking</td>
<td>looks</td>
<td>ltd</td>
</tr>
<tr>
<td>m</td>
<td>mainly</td>
<td>many</td>
<td>may</td>
<td>maybe</td>
</tr>
<tr>
<td>me</td>
<td>mean</td>
<td>meanwhile</td>
<td>merely</td>
<td>might</td>
</tr>
<tr>
<td>more</td>
<td>moreover</td>
<td>most</td>
<td>mostly</td>
<td>much</td>
</tr>
<tr>
<td>must</td>
<td>much</td>
<td>my</td>
<td>myself</td>
<td>n</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>name</td>
<td>namely</td>
<td>nd</td>
<td>near</td>
<td>nearly</td>
</tr>
<tr>
<td>necessary</td>
<td>need</td>
<td>needs</td>
<td>neither</td>
<td>never</td>
</tr>
<tr>
<td>nevertheless</td>
<td>new</td>
<td>next</td>
<td>nine</td>
<td>no</td>
</tr>
<tr>
<td>nobody</td>
<td>non</td>
<td>none</td>
<td>noone</td>
<td>nor</td>
</tr>
<tr>
<td>normally</td>
<td>not</td>
<td>nothing</td>
<td>novel</td>
<td>now</td>
</tr>
<tr>
<td>nowhere</td>
<td>o</td>
<td>obviously</td>
<td>of</td>
<td>off</td>
</tr>
<tr>
<td>often</td>
<td>oh</td>
<td>ok</td>
<td>okay</td>
<td>old</td>
</tr>
<tr>
<td>on</td>
<td>once</td>
<td>one</td>
<td>ones</td>
<td>only</td>
</tr>
<tr>
<td>onto</td>
<td>or</td>
<td>other</td>
<td>others</td>
<td>otherwise</td>
</tr>
<tr>
<td>ought</td>
<td>our</td>
<td>ours</td>
<td>ourselves</td>
<td>out</td>
</tr>
<tr>
<td>outside</td>
<td>over</td>
<td>overall</td>
<td>own</td>
<td>p</td>
</tr>
<tr>
<td>particular</td>
<td>particularly</td>
<td>per</td>
<td>perhaps</td>
<td>placed</td>
</tr>
<tr>
<td>please</td>
<td>plus</td>
<td>possible</td>
<td>presumably</td>
<td>probably</td>
</tr>
<tr>
<td>provides</td>
<td>q</td>
<td>que</td>
<td>quite</td>
<td>qv</td>
</tr>
<tr>
<td>r</td>
<td>rather</td>
<td>rd</td>
<td>re</td>
<td>really</td>
</tr>
<tr>
<td>reasonably</td>
<td>regarding</td>
<td>regardless</td>
<td>regards</td>
<td>relatively</td>
</tr>
<tr>
<td>respectively</td>
<td>right</td>
<td>s</td>
<td>siad</td>
<td>same</td>
</tr>
<tr>
<td>saw</td>
<td>say</td>
<td>saying</td>
<td>says</td>
<td>second</td>
</tr>
<tr>
<td>secondly</td>
<td>see</td>
<td>seeing</td>
<td>seem</td>
<td>seemed</td>
</tr>
<tr>
<td>seeming</td>
<td>seems</td>
<td>seen</td>
<td>self</td>
<td>selves</td>
</tr>
<tr>
<td>sensible</td>
<td>sent</td>
<td>serious</td>
<td>seriously</td>
<td>seven</td>
</tr>
<tr>
<td>several</td>
<td>shall</td>
<td>she</td>
<td>should</td>
<td>since</td>
</tr>
<tr>
<td>six</td>
<td>so</td>
<td>some</td>
<td>somebody</td>
<td>somehow</td>
</tr>
<tr>
<td>someone</td>
<td>something</td>
<td>sometime</td>
<td>sometimes</td>
<td>somewhat</td>
</tr>
<tr>
<td>somewhere</td>
<td>soon</td>
<td>sorry</td>
<td>specified</td>
<td>specify</td>
</tr>
<tr>
<td>specifying</td>
<td>still</td>
<td>sub</td>
<td>such</td>
<td>sup</td>
</tr>
<tr>
<td>sure</td>
<td>t</td>
<td>take</td>
<td>taken</td>
<td>tell</td>
</tr>
<tr>
<td>tends</td>
<td>th</td>
<td>than</td>
<td>thank</td>
<td>thanks</td>
</tr>
<tr>
<td>thanx</td>
<td>that</td>
<td>thats</td>
<td>the</td>
<td>their</td>
</tr>
<tr>
<td>theirs</td>
<td>them</td>
<td>themselves</td>
<td>then</td>
<td>thence</td>
</tr>
<tr>
<td>there</td>
<td>thereafter</td>
<td>thereby</td>
<td>therefore</td>
<td>therein</td>
</tr>
<tr>
<td>theres</td>
<td>thereupon</td>
<td>these</td>
<td>they</td>
<td>think</td>
</tr>
<tr>
<td>third</td>
<td>this</td>
<td>thorough</td>
<td>thoroughly</td>
<td>those</td>
</tr>
<tr>
<td>though</td>
<td>three</td>
<td>through</td>
<td>throughout</td>
<td>thru</td>
</tr>
<tr>
<td>thus</td>
<td>to</td>
<td>together</td>
<td>too</td>
<td>took</td>
</tr>
<tr>
<td>toward</td>
<td>towards</td>
<td>tried</td>
<td>tries</td>
<td>truly</td>
</tr>
<tr>
<td>try</td>
<td>trying</td>
<td>twice</td>
<td>two</td>
<td>u</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>un</td>
<td>under</td>
<td>unfortunately</td>
<td>unless</td>
<td>unlikley</td>
</tr>
<tr>
<td>until</td>
<td>unto</td>
<td>up</td>
<td>upon</td>
<td>us</td>
</tr>
<tr>
<td>use</td>
<td>used</td>
<td>usefull</td>
<td>uses</td>
<td>using</td>
</tr>
<tr>
<td>usually</td>
<td>uucp</td>
<td>v</td>
<td>value</td>
<td>various</td>
</tr>
<tr>
<td>very</td>
<td>via</td>
<td>viz</td>
<td>vs</td>
<td>w</td>
</tr>
<tr>
<td>want</td>
<td>wants</td>
<td>was</td>
<td>way</td>
<td>we</td>
</tr>
<tr>
<td>welcome</td>
<td>well</td>
<td>went</td>
<td>were</td>
<td>what</td>
</tr>
<tr>
<td>whatever</td>
<td>when</td>
<td>whence</td>
<td>whenever</td>
<td>where</td>
</tr>
<tr>
<td>whereafter</td>
<td>whereas</td>
<td>whereby</td>
<td>wherein</td>
<td>whereupon</td>
</tr>
<tr>
<td>wherever</td>
<td>whether</td>
<td>which</td>
<td>while</td>
<td>whither</td>
</tr>
<tr>
<td>who</td>
<td>whoever</td>
<td>whole</td>
<td>whom</td>
<td>whose</td>
</tr>
<tr>
<td>why</td>
<td>will</td>
<td>willing</td>
<td>wish</td>
<td>with</td>
</tr>
<tr>
<td>within</td>
<td>without</td>
<td>wonder</td>
<td>would</td>
<td>x</td>
</tr>
<tr>
<td>y</td>
<td>yes</td>
<td>yet</td>
<td>you</td>
<td>your</td>
</tr>
<tr>
<td>yours</td>
<td>yourself</td>
<td>yourselves</td>
<td>z</td>
<td>zero</td>
</tr>
</tbody>
</table>

Table B.3: Default list of common English "stop words" from the text. MALLET topic modelling uses this default list.
Appendix C

Post-Processing of Data

This Appendix contains detailed information about steps taken to post process the results of topic modelling algorithm.

C.1 Compute cosine similarity

This step is written in java and tested it by implementing JUnit testing to compute the similarity matrix. The structure of the package is:

I order to make code usable in future I divided the task into 3 section:

1. ReadExcelFile.Java : takes an excel file which contains topic-term weights of all 50 topics;

2. CosineSimilarity.java : Takes matrix from step 1 and returns a Matrix of size 50 x 50; (number of topics = 50);


Figure C.1: Structure of Java package for calculating cosine similarity of topics
C.1.1 SimilarityTest.java

```java
package TopicSimilarity;
import java.io.File;
import java.io.PrintWriter;
import org.junit.Before;
import org.junit.Test;
import Jama.Matrix;
import TopicSimilarity.PrintMatrix;

public class SimilarityTest {

@Before
public void setUp() throws Exception {
}

@Test
//test cosine with input matrix from excel file
public void testCosine() throws Exception{
//define instances
ReadExcelFile readTest = new ReadExcelFile();
CosineSimilarity cosineSim = new CosineSimilarity();
PrintMatrix pr = new PrintMatrix();

//1. read excel file and compute the topic-term matrix
Matrix termDocMatrix = new Matrix(readTest.readXLSXFile());

//2. computes cosine similarity
Matrix similarity = cosineSim.transform(termDocMatrix);

//3. prints the similarity matrix to a text file
String[] topicNames = pr.getTopicNames(termDocMatrix);
//define output text file
File file = new File("files/cosineSim04072017.txt");
PrintWriter writer = new PrintWriter(file);
pr.prettyPrintMatrix("Cosine Similarity From Excel file", similarity,
                    topicNames, writer);
}
}
```

Listing C.1: JUnit testing of TopicSimilarity package. This unit calls 3 java classes: ReadExcelFile.java CosineSimilarity.java PrintMatrix.java ; label
C.1.2 ReadExcelFile.Java

For information about JAMA Matrix please visit JAMA : A Java Matrix Package

```java
package TopicSimilarity;
import java.io.FileInputStream;
import java.io.IOException;
import java.io.InputStream;
import java.util.Iterator;
import org.apache.poi.ss.usermodel.Cell;
import org.apache.poi.ss.usermodel.Row;
import org.apache.poi.xssf.usermodel.XSSFCell;
import org.apache.poi.xssf.usermodel.XSSFRow;
import org.apache.poi.xssf.usermodel.XSSFSheet;
import org.apache.poi.xssf.usermodel.XSSFWorkbook;

// This class reads data from an excel file and stores values into 2D array @author Azimeh Gharavi

public class ReadExcelFile {
    public static double[][] readXLSXFile() throws IOException {
        //set the path to the excel file and initialise workbook and sheet
        InputStream ExcelFileToRead = new FileInputStream("files/WTW-10Topics.xlsx");
        XSSFWorkbook wb = new XSSFWorkbook(ExcelFileToRead);
        XSSFSheet sheet = wb.getSheetAt(0);
        XSSFRow row;
        XSSFCell cell;
        Iterator<Row> rows = sheet.rowIterator();
        int numRows = sheet.getPhysicalNumberOfRows();
        int numColumns = sheet.getRow(0).getPhysicalNumberOfCells();
        double[][] values = new double[numRows][numColumns];
        int rowIndex = 0;
        while (rows.hasNext()) {
            int colIndex = 0;
            row = (XSSFRow) rows.next();
            Iterator<Cell> cells = row.cellIterator();
            while (cells.hasNext()) {
                cell = (XSSFCell) cells.next();
                values[rowIndex][colIndex] = cell.getNumericCellValue();
                colIndex ++;
            }
            colIndex = 0;
            rowIndex = rowIndex + 1;
        }
        return values;
    }
}
```

Listing C.2: Java class to take excel file containing topic-term weights and convert to Java MAtrix
C.1.3 CosineSimilarity.Java

CosineSimilarity.java class implements equation in C.1.3.

\[
similarity(A, B) = \frac{(A \cdot B)}{|A| \cdot |B|}
\]  

(C.1)

where:

\[
A \cdot B = \sum Ai \cdot Bi \\
|A| = \sqrt{\sum Ai^2} \\
|B| = \sqrt{\sum Bi^2}
\]

for \( i = [0..n-1] \),

where \( n \) = number of terms in topic-term matrix.

\( n \) in this project equals to 9,911

\( \cos \) is in range of \([0 1]\) and the higher this value the more similar the A and B.

##### Listing C.3: Java class to compute Cosine similarity: takes topic-term matrix and outputs the topic-topic similarity matrix
C.1.4 PrintMatrix.Java

package TopicSimilarity;

import java.io.PrintWriter;
import Jama.Matrix;

public class PrintMatrix {

    /*
     * pretty print matrix:
     * example:
     * === Cosine Similarity From Excel file ===
     *     T0    T1    T2
     *     T0  1.0000 0.6030 0.0767
     *     T1  0.6030 1.0000 0.8325
     *     T2  0.0767 0.8325 1.0000
     */
    void prettyPrintMatrix(String legend, Matrix matrix,
        String[] topicNames, PrintWriter writer) throws Exception {
        writer.printf("\n=== %s ===\n", legend);
        writer.printf("%8s", " ");
        for (int i = 0; i < topicNames.length; i++) {
            writer.printf("%12s", topicNames[i]);
        }
        writer.println();
        for (int i = 0; i < topicNames.length; i++) {
            writer.printf("%8s", topicNames[i]);
            for (int j = 0; j < topicNames.length; j++) {
                writer.printf("%12.6f", matrix.get(i, j));
            }
        }
        writer.println();
        writer.flush();
    }

    public String[] getTopicNames(Matrix matrix){
        int numTopics = matrix.getColumnDimension();
        String[] topicNames = new String[numTopics];
        for (int i=0 ; i< topicNames.length; i++){
            topicNames[i] = "T"+i;
        }
        return topicNames;
    }
}

Listing C.4: Java class to print a Matrix to excel file
Appendix D

Interface prototypes and implementation

D.0.1 Making results of Topic Modelling ready for visualisation

D.0.1.1 What files from topic modelling results I used? which information.

D.0.1.2 Clustering Algorithm

Python code that I used for clustering algorithm was donated by Mr. Pierre Le Bras. This code is then modified to accommodate different models. Unfortunately I can not publish his code due to his current research projects. Algorithm uses average distance function and checks following two conditions:

1. The size of each cluster is less than or equal to number of models,

2. There is only one topic from each model in each cluster.

D.0.2 Format of dataset used for visualisation

Output from clustering algorithm is a JSON dataset which serves as input data to D3.js. Complete JSON dataset is available in the code of the project. This dataset has two parts: 1. information about clusters and 2. information about topics.

```
"clusters": [
  {
    "agglomerative": 0,
    "nodes": [
      { "model": 5, "id": "id48"},
      { "model": 1, "id": "id1"},
      { "model": 4, "id": "id31"},
      { "model": 2, "id": "id18"},
      { "model": 3, "id": "id25"}
    ]
  }
],
```
Figure D.1: Clusters Object in JSON dataset has information about 12 clusters and the member nodes of each cluster.

```json
"topics": {
  "model": 4,
  "agglomerate": 11,
  "id": "id34",
  "name": {
    "TopWords": [
      {"label": "receptor", "weight": 0.02628},
      {"label": "cell", "weight": 0.01717},
      {"label": "channel", "weight": 0.01567}
    ]
  }
}
```

Figure D.2: Topics Object in JSON dataset has information about 50 topics for example which cluster they belong to and their top words.

D.0.3 Evolution of layouts

Figure D.3: Dendrogram produced using linkage table is the first prototype I developed. I used this layout to identify the best number of clusters to use in clustering algorithm.

Figure D.4: Dendrogram produced by using complete distance function to generate the linkage table. Also colour codes clusters.
Appendix D. Interface prototypes and implementation

**Figure D.5:** Hexagons layout is used to visualise one solution of topic modelling on Wellcome Trust dataset. Each hexagon is one topic and inside each topic top 5 words are displayed.

**Figure D.6:** Hexagons layout: in this layout each hexagon is one topic, each island is one cluster and each colour is one model.

**Figure D.7:** Example implementation of transition of labels and colours on tabular layout.

**Figure D.8:** Example of tooltip feature implemented on tabular layout.

**Figure D.9:** 10 Topics each showing weighted 5 top words.
Figure D.10: 3 solutions - 10 Topics with weighted top 5 words.
Row 1: running LDA Topic Modelling algorithm on Wellcome Trust grant funding dataset with random seed 1.
Row 2: result of running same algorithm on same dataset with same setting using random seed 500.

Figure D.11: Tabular web interface showing relationships between different models. Each row represent a model and each column is a cluster of topics. Topics in each column are the most similar topics across 5 models.
D.0.4 Code of tabular layout

D.0.4.1 index.html

<!--------------------------------------------------------------------
Module: MSC Project Main HTML file - Machine Learning Explanation for Boardroom
5 models -

Author: Azimeh Gharavi

What it does:
This interface shows 5 solutions of LDA topic modelling on Wellcome Trust dataset
For each solution (Model): 10 Topics are displayed with weighted top 5 words of topic in each tile.
This interface will be used by user to select the BEST and WORST solution (Model) among 5.

Dependencies
D3.js v3
tabular.js

Version history

Comments:
-------------------------------------------------------------------->

<!DOCTYPE html>
<meta charset="utf-8">
<body>
<script src="http://d3js.org/d3.v3.min.js"></script>
<link rel="stylesheet" type="text/css" href="./css/style.css" />
<script type="text/javascript" src="./js/tabular.js"></script>
<script type="text/javascript" src="./js/tooltip.js"></script>

<div ID="tabDiv"></div>
<script>
//============================creating SVG element for layout ======================
//objects of different renderers
var tab = tabular("#tabDiv");

// nasty globals to ease debug - should really use function args and private variables
var topicModelData = {};

//data sources
var urlTopicData5Models = ".//jsondata/agglomerated_nodes_5models.json";

//------------- READ JSON Data --------------------------

//Read Topic data
d3.json(urlTopicData5Models, function(error, tJsonData) {
    if (error) return console.log("Failed attempting to load JSON from:", urlTopicData5Models);
    topicModelData = tJsonData;
    processData();
});
Listing D.1: HTML code of the interface. This code imports data from a JSON file then renders tabular layout by calling tabular.js finally adds overwrite tooltip function.

D.0.4.2 tabular.js

// Module: tabular layout renderer CLASS
// Author: Azimeh Gharavi
// What it does:
// This JavaScript module implements a tabular layout (small multiples) in D3.js
// It adds an svg to the targetDOMElement where it renders the chart
// Dependencies
// D3.js v3
// Version history
// v001 04/07/2017 AG Created.
//
"use strict" //This catches acidental global declarations

function tabular(targetDOMelement) {
  //Here we use a function declaration to imitate a 'class' definition
  //This returns the object below with attached public methods
  var tabularObject = {};

  //=================== PUBLIC METHODS =========================
  tabularObject.loadAndRenderDataset = function (data) {
    //Load d3 format hierarchy
    dataset=data.clusters;
    topicsDataset = data.topics;
    render();
    return tabularObject; //enable chaining
  }

  tabularObject.overideTooltip = function (tooltipFunction) {
    //provides custom labelling for parent nodes
    tooltip = tooltipFunction;
    return tabularObject; //enable chaining
  }

  //=================== PRIVATE VARIABLES ====================================
  //Layout constants
  var svgWidth = 1000,
      svgHeight = 600,
      tileWidth = 70,
      tileHeight = 70,
      tileMargin = 25;

  var dataset = [],
      clusterDataset = [],
      topicsDataset = [];

  var borderRadiusX = 10,
      borderRadiusY = 10;
  var tabColour = "green",
      border=1,
      borderColor="black",
      labelcolor = 'red';

  var colorScale = d3.scale.category20();

  //Declare and append SVG element // Create the table
  var svg = d3.select(targetDOMelement)
    .append("svg")
    .attr("width", svgWidth)
    .attr("height", svgHeight);

  //=================== PRIVATE FUNCTIONS =========================
  function render () {
    GUP_tiles();
    GUP_labels();
  }

  var tileClick = function(d){

// console.log("tabular clicked!");
}
// TO BE CHANGED
var tileLabel = function(d,i){
    return d.id;
};
// TO BE CHANGED
var tooltip = function(d){
    //console.log("BBBBB");
}; //simple default

function GUP_tiles(){
    //GUP = General Update Pattern to render small multiples
    //GUP: BIND DATA
    var selection = d3
        .selectAll("rect")
        .data(dataset);

    selection
        .exit()
        .remove();
    //GUP Update 1
    selection
        .attr("border", tabColour)
        .style("stroke", bordercolor)
        .style("fill", "none")
        .style("stroke-width", border);

    //GUP: Enter selection
    //i.e. create necessary DOM elements if they don’t already exist
    selection.enter()
        .append("g")
    //GUP Update2 (this selection = update1 + enter)
    //i.e. update all selected visual elements on page
    selection
        .each(function(d2, i){
            d3.select(this).selectAll("rect")
                .data(d2.nodes)
                .enter()
                .append("rect")
                .on("click", tileClick)
                .transition()
                .delay(function(d) { return (( i )) / dataset.length * 5000; })
                .duration(4000)
                .attr("y", function(d2) {
                    if (d2.model == 1 || d2.model == 2){ return d2.model * tileWidth +tileMargin
                        ;}
                    else if (d2.model == 4 || d2.model == 5) { return (d2.model- 1) * tileWidth +
                        tileMargin}
                    else { return (d2.model+2) * tileWidth +tileMargin}
                })
                .attr("x", function(d2) { return (i+1)*tileHeight + tileMargin;})
                .attr("width", tileWidth)
                .attr("height", tileHeight )
        })
}
.attr("border", tabColour)
 .style("stroke", bordercolor)
 .style("fill", function(d, i){ if (d2.aggloCluster >= 0 ) { return colorScale(d2 .aggloCluster); } })
 .style("stroke-width", border)
 .attr({ry : borderRadiusX, rx : borderRadiusY });
})
 .append("text")
 .transition()
 .delay(function(d,i ) { return (i / dataset.length *2000 ) ; })
 .duration(4000)
 .attr("x", function(d,i){return ((i+1.1))*tileHeight + tileMargin;})
 .attr("y", 80)
 .text(function(d,i){ return "Topic " + d.aggloCluster; } );

//GUP exit selection
selection
 .exit()
 .remove();

function GUP_labels(){
//GUP = General Update Pattern to render labels

//GUP: Bind data
var selection = svg
.selectAll(".tile-label")
 .data(topicsDataset);

//GUP Update1 - do nothing for now
selection
 .exit()
 .remove();
//GUP: Enter selection
//Here we create the basic DOM text elements if they don't exist
selection.enter()
 .append("text")
 .attr("class", "tile-label");

//GUP Update2 (this selection = update1 + enter)
//Here we update the labels and their positions
selection
 .call(d3.helper.tooltip(tooltip))
 .transition()
 .delay(function(d,i) { var order = d.aggloCluster;
 if ( order == 0 || order ==2 ) { return (order + 2) / dataset.length * 5000; }
 else if( order == 7 ) {return 3 / dataset.length * 5000; } 
 else if( order == 4 ) {return 5 / dataset.length * 5000; } 
 else if( order == 10 ) {return 6 / dataset.length * 5000; } 
 else if( order == 3 ) {return 7 / dataset.length * 5000; } 
 else if( order == 11 ) {return 8 / dataset.length * 5000; } 
 else if( order == 9 ) {return 9 / dataset.length * 5000; } 
 else if( order == 1 ) {return 10 / dataset.length * 5000; } 
 else if( order == 6 ) {return 11 / dataset.length * 5000; } 
})
else if (order == 5) {return 12 / dataset.length * 5000; }
else if (order == 8) {return 13 / dataset.length * 5000; }
}
.duration(5000)
.attr("y", function(d2) {
  if (d2.model == 1 || d2.model == 2) { return (d2.model + 0.5) * tileWidth + tileMargin; }
  else if (d2.model == 4 || d2.model == 5) { return (d2.model - 1 + 0.5) * tileWidth + tileMargin; }
  else { return (d2.model + 2 + 0.5) * tileWidth + tileMargin; }
})
.attr("x", function(d, i) {
  var order = d.aggloCluster;
  if (order == 0 || order == 2) {return (order + 1 + 0.1) * tileHeight + tileMargin; }
  else if (order == 7) {return 2.1 * tileHeight + tileMargin; }
  else if (order == 4) {return 4.1 * tileHeight + tileMargin; }
  else if (order == 10) {return 5.1 * tileHeight + tileMargin; }
  else if (order == 3) {return 6.1 * tileHeight + tileMargin; }
  else if (order == 11) {return 7.1 * tileHeight + tileMargin; }
  else if (order == 9) {return 8.1 * tileHeight + tileMargin; }
  else if (order == 1) {return 9.1 * tileHeight + tileMargin; }
  else if (order == 6) {return 10.1 * tileHeight + tileMargin; }
  else if (order == 5) {return 11.1 * tileHeight + tileMargin; }
  else if (order == 8) {return 12.1 * tileHeight + tile Margin; }
})
.text(tileLabel);

//GUP exit selection
selection.exit().remove();

//================================================ IMPORTANT do not delete ===============
return tabularObject; // return the main object to the caller to create an instance of the 'class'
}
Listing D.3: `tooltip.js` function which when called will create a floating DOM element and display some information inside the DOM.
Appendix E

Evaluation

E.1 Advert for recruiting participants

Following advert were published on social media among friends.

---

Dear friends in Edinburgh (preferably not computer scientists)

I need your help :) 

For my Master's dissertation project I need to perform a user experiment in the form of focus group and I appreciate if you can spare 1 hour for me.

What is my project about?
People's trust and confidence on machine learning algorithms in decision making process.

However measuring trust is not as easy as it sounds so I reduced the scope and the purpose of this experiment will be:

"How people see the difference between solutions of machine learning algorithms and how they choose the "BEST" solution."

What is a focus group?
It is a kind of user experiment where a few of you (~ 5 - 8 people) will gather in a room, then I will present my project to all of you, demonstrate my web interface then ask you some questions. Then you will discuss your answers with other participants. For example you will get to choose the best option among available options, then explain why you have chosen this answer, why you think the other answers are not the BEST answer.

What I will do?
I will ask you to fill a few questionnaires (~ 1 - 3 parts) and record the whole sessions so I can transcript discussions to use your useful and interesting points and insights in my report.

How long will it take?
The experiment will last for around 1 hour (not more than 1 hour, 15 minutes).

When?
Wednesday 25 July - from 2 pm - 3:15
OR
Saturday 29 July - from 11 am - 12:16

Where?
Heriot-Watt University. (Details of room will be confirmed 3 days before the experiment)

Notes:
- You do not need to know about machine learning algorithms.
- I won't record any personal and sensitive information.
- You can withdraw from experiment at any time before end of experiment.
- Experiment will be in English.

Please send me a private message if you can make any of these sessions.

---

121
E.2 Facilitator’s discussion guide

Welcome
Welcome and thank you for coming to my focus group. I appreciate your time.

Introduction
This focus group is all about your views on Machine Learning algorithms and what are your criteria for selecting the BEST solution among a number of valid solutions.

Presentation

1. Introducing project

2. Aims (Project and Focus Group)

3. Topic Modelling

4. Dataset

5. Show an example Topics

Presentation slides are included in Section E.3

Demonstration

1. Show 1Model – (Seed = 1)
   
   a. Explain Topic,
   
   b. Terms (Top words), and weights of the words.

2. Using same algorithm on the same dataset, same number of topics and words per topic. By only changing the RANDOM SEED, algorithm will result in slightly different results.
   
   a. RANDOM SEED relates to slide number 12 of presentation in Section E.3.
   
   b. Show 2Models – ( Seed = 1 & Seed = 1001 )
   
   c. Point out the differences between two models for example Topic 1 pretty much stays the same but Topic 4 changes.
Appendix E. Evaluation

3. Show comparison of another two models with more obvious differences. (Seed = 1 & Seed = 500)
   
   a. Some topics stay stable for example Topic 1
   
   b. Some topics change slightly for example Topic 4, Topic 6
   
   c. 2 Topics disappeared and 2 new topics formed

4. Show 5 models
   
   a. This interface shows 5 solutions among many available models.
   
   b. Please look carefully and try to make a note of which model you think is the WORST and which model is the BEST among these 5 models.

   You have 5 minutes to decide.

Anonymity

Despite being recorded, I assure you that the discussion will be anonymous. The recording will be kept safely in a password protected laptop until they are transcribed word for word, then they will be destroyed.
Legal requirement
This analysis uses openly available grant funding information made publicly available by the Wellcome Trust. Wellcome have not been consulted in or have reviewed the subsequent analysis and interpretation of the data undertaken.

Ground rules

1. Only one person speaks at a time.

2. There are no right or wrong answers

3. I’m just as interested in negative comments as positive comments, sometimes negative comments are the most helpful.

4. When you do have something to say, please do so. There are many of you in the group and it is important that I obtain the views of each of you

5. You do not have to agree with the views of other people in the group

6. Any questions?

7. OK, let’s begin.

Group discussions
This part will last around 20 to 30 minutes. I will start the conversation with question 1 and 2 and let people to discuss among themselves.

Question 1 : Which model you chose as BEST model if you only had access to this interface?

- Explanation: Here BEST model is the one that participant is most likely to choose by just examining the interface;

- Probes: this question don’t need any probes.

Question 2: How did you choose this model?

- Explanation: I will explain that they might feel their reasons are very basic and obvious, but they are important to me.

- Probes:
  1. "Would you explain further?"
  2. "Would you give an example?"
  3. "I don’t understand."
  4. "Does anyone else have some thoughts on that?"
Question 3: 3. *What are the advantages of this interface when selecting the BEST model?*

- **Explanation:** I will display interface on screen for them and ask them to tell each other at least 1 advantage of the interface.

- **Probes:**
  1. "Can you explain further?";
  2. "Does anyone else have some thoughts on that?";
  3. "X (X here is the name of the person who is not involved in discussion), do you agree with this?"

Question 4: *What are the disadvantages of this interface?*

- **Explanation:** For some people finding what they don’t like is easier than what they like. Same as question 3, I will display interface on screen for them and ask them to tell each other at least 1 advantage of the interface.

- **Probes:**
  1. "Can you say more about that?";
  2. "Does anyone else have some thoughts on that?"

Question 5: 5. *What suggestions do you have to improve this interface?*

- **Explanation:** What extra features participants like to have.

- **Probes:**
  1. "Can you explain how this feature will help you to select the BEST model?"
  2. "Does anyone else have some thoughts on that?"

**Conclusion**

1. Thank you for participating. This has been a very successful discussion;

2. Your opinions will be a valuable asset to my project;

3. I would like to remind you that any comments featuring in this report will be anonymous.
E.3 Presentation

This section includes presentation I used in focus group to introduce my project and topic modelling.

Outline of the experiment

<table>
<thead>
<tr>
<th>Activity</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consent form</td>
<td>5 Minutes</td>
</tr>
<tr>
<td>Presentation of the project</td>
<td>10 Minutes</td>
</tr>
<tr>
<td>Demonstration of Interface</td>
<td>10 Minutes</td>
</tr>
<tr>
<td>Questions</td>
<td>5 Minutes</td>
</tr>
<tr>
<td>First questionnaire</td>
<td>10 Minutes</td>
</tr>
<tr>
<td>Group Discussion</td>
<td>20 Minutes</td>
</tr>
<tr>
<td>Last questionnaire</td>
<td>10 Minutes</td>
</tr>
</tbody>
</table>

Project Motivations

1. Big Data → Machine Learning algorithms are used to analyse
2. Often users are given ONE solution
3. ML algorithms will result in different solutions (ALL VALID) by changing constraints or by using different model.
4. Users are often unaware of this fact

Research Questions:

1. How people react when exposed to NON-DETERMINISTIC nature of ML algorithms?
2. Can a group of people agree on a BEST solution based on visualisation of available options?

My Study

1. Corpus – Welcome Trust grant funding dataset
2. ML algorithm - Topic Modelling

Corpus

- Welcome Trust awarded grants – 1 Oct 2000 to 30 Sep 2015.
- Total grants (After cleaning dataset): 9,391
- Each document: Project Title + Project Abstract

<table>
<thead>
<tr>
<th>Project Title</th>
<th>Project Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>China Kohorel BioBank (CBB)</td>
<td>A longitudinal prospective study of 532,000 adults, recruited during 2000-08 from 10 disease regions of China, with extensive data collected at baseline and subsequent average 2.4 year follow-up. Participants had undergone comprehensive physical measurements, and stored biological samples. By 3.1.2016, 2,849 deaths and 1,541 incident disease events had been recorded among participants, through linkages with death and disease registries and...</td>
</tr>
</tbody>
</table>
Appendix E. Evaluation

Pre-processing data

1. Missing data
2. Stop word removal
3. Lemmatizing

Topic Modelling

Topic Modelling assumption

![Diagram showing the topic modelling assumption]

From: David M. Blei, Probabilistic Topic Models, communication of the ACM, April 2012

Topic Modelling algorithm

Iterative algorithm
1. Initialize parameters
2. Initialize topic assignments randomly
3. Iterate
   - For each word in each document:
     - Resample topic for word, given all other words and their current topic assignments
4. Get results
5. Evaluate model

1. Initialize parameters

![Diagram showing the initialization of parameters]

From: Christine Doig, Introduction to Topic Modelling in Python, PyTexas 2015

2. Initialize topic assignment

For each document do:
- Randomly assign each word to a Topic

![Diagram showing topic assignment]

From: Christine Doig, Introduction to Topic Modelling in Python, PyTexas 2015

So we will have:

![Diagram showing topic assignments and distributions]

From: Christine Doig, Introduction to Topic Modelling in Python, PyTexas 2015

3. Iterate

For each word in each document resample the Topics.
Based on current topic assignment and all other words

![Diagram showing the iteration process]

From: Christine Doig, Introduction to Topic Modelling in Python, PyTexas 2015
Model setting:

Algorithm: LDA
Corpus size: 9,901 documents
Number of Topics: 10
Number of words in each topic: 10
Number of iterations: 500

Discovered Topics - Example

5 Topics – Displayed with Top 5 words

brain, mental, health
study, policy, research
protein, cell, development

Legal requirement:

“This analysis uses openly available grant funding information made publicly available by the Wellcome Trust.
https://wellcome.ac.uk/funding/managing-grant/grants.awarded
Wellcome have not been consulted in or have reviewed the subsequent analysis and interpretation of the data undertaken.”
Bibliography


