Find me a Dream Job: Using NLP to build a recommender system for Job Listings

Author: Julian Matthis Kurz
Student Number: H00259335

Supervisor: Dr Ekaterina Komendantskaya
Second Reader: Dr Diana S Bental

A thesis submitted in fulfilment of the requirements for the degree of MSc. Artificial Intelligence with Speech and Multimodal Interaction in the School of Mathematical and Computer Sciences

August 2018
Declaration of Authorship

I, Julian Kurz, declare that this thesis titled, 'Find me a Dream Job: Using NLP to build a recommender system for Job Listings' 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:
Julian Kurz

Date:
12th of April 2018
“(Let the AI) Choose a job you love and you will never have to work a day in your life.”

- based on Confucius circa 551-479 BC
Abstract

There is little empirical evidence for companies concerning the benefits of using data to inform their technologies in Germany, especially for small to medium sized companies. In this report, we outline an experiment design and project plan to use the business data of "Jobsuche Regional", a German company owning several job hunting websites, to build a recommender system. We focus on the usage of natural language processing to build the recommender system.

A corpus was prepared from the job listings posted on several regional job hunting websites. This corpus was used to train a word2vec model in order to make recommendations based on a similar semantic context between words. This was also done based on a recent corpus of the German Wikipedia, to compare how these two corpora compare in creating a recommender system. Both systems where shown to be more capable than the current system used by the company. The system built directly from the job data had a smaller vocabulary and was less consistent. However, compared to the system built on Wikipedia, it was more specialised on the language of job listings. This enabled it, for instance, to take qualifications required for a position into account more strongly.

While a prototype was prepared to be deployed at "Jobsuche Regional", this was unfortunately not possible yet. Instead, five German natives annotated an extract of the corpus on "Jobsuche Regional" and the recommender systems, as well as the original company algorithm, were tested on this data as a classifier. While the company algorithm proved to be mostly random, both classifiers based on natural language processing were considerably more capable, with the Wikipedia model reaching an accuracy of 30 percent and the model trained directly from "Jobsuche Regional" with 43 percent, compared to an expected random baseline of 16 percent.

This experiment provides empirical evidence that data can be used to positively impact the business model of a small German company. Further, it showed that word embeddings are a valid approach when trying to build a recommender system.
Acknowledgements

I would like to thank Ekaterina Komendantskaya for her advice, motivation and simply always making time and Vanessa Frenz who has been a constant help and support.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declaration of Authorship</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iv</td>
</tr>
<tr>
<td>Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>xii</td>
</tr>
</tbody>
</table>

### 1 Introduction
1.1 Introduction .................................................. 1  
1.1.1 Motivation .................................................. 1  
1.1.2 Aims and Objectives ......................................... 2  
1.2 Customer Journey ............................................... 3  
1.3 Hypothesis .................................................... 6  
1.4 Challenges .................................................... 6  

### 2 Literature Review
2.1 Recommender Systems ........................................... 8  
2.1.1 Motivation .................................................. 8  
2.1.2 Tasks ....................................................... 9  
2.1.3 Recommendation Techniques ................................. 10  
  Content-based recommendations ................................ 10  
  Collaborative filtering ........................................ 11  
  Hybrid Systems .................................................. 11  
2.1.4 Multidimensionality of Recommendations .................... 11  
  Critical comparison of the existing methods ................... 11  
2.2 Natural Language Processing .................................. 12  
2.2.1 Word Embeddings ............................................ 12  
  2.2.1.1 Bag of Words ........................................... 12
<table>
<thead>
<tr>
<th>Contents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D  File Glossary</td>
<td>96</td>
</tr>
<tr>
<td>Bibliography</td>
<td>99</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 Jobsuche Regional Start ........................................ 3
1.2 Jobsuche Regional Selection Page ................................. 4
1.3 Jobsuche Regional Main Page ..................................... 5
1.4 Jobsuche Regional Detailed False Positives ....................... 6

2.1 Simple word vectors .............................................. 15
2.2 Modelling concepts by adding word vectors. ....................... 16
2.3 Doc2Vec .............................................................. 17

3.1 Building the recommender system ................................ 23
3.2 Proposed System 1 ................................................. 24
3.3 Proposed System 2 ................................................. 25
3.4 Proposed System 3 ................................................. 26

6.1 The Software Development Impact Statement ...................... 37
6.2 Gantt for all milestones ........................................... 42
6.3 Gantt for Milestone 1 ............................................. 42
6.4 Gantt for Milestone 2 ............................................. 43
6.5 Gantt for Milestone 3 ............................................. 43
6.6 Gantt for Milestone 4 ............................................. 44

7.1 Raw User Data ...................................................... 46
7.2 Google Analytics Overview ........................................ 46
7.3 Google Analytics Single item ...................................... 47
7.4 Job Data Noisy ...................................................... 47
7.5 Job Data Corrupted .................................................. 48
7.6 Job Object .......................................................... 49
7.7 Input JSON for Prototype Corpus ................................ 53
7.8 Input JSON for Prototype Similarity ............................... 54

8.1 PCA Visualization "Erzieher" Job Data ......................... 63
8.2 PCA Visualization "Erzieher" Wikipedia .......................... 64
8.3 PCA Visualization "Entwickler" Job Data ........................ 64
8.4 PCA Visualization "Entwickler" Wikipedia ....................... 65
8.5 Jobs And Locations ................................................ 66
8.6 Synonym Jobs Versus Similar Jobs ............................... 67
8.7 Normalized Confusion Matrix Keyword Count ................... 70
8.8 Unnormalized Confusion Matrix Keyword Count ................ 71
8.9 Accuracy Report Keyword Count ................................ 71
List of Figures

8.10 Normalized Confusion Matrix Wiki Model . . . . . . . . . . . . . . . . . . . . 72
8.11 Unnormalized Confusion Matrix Wiki Model . . . . . . . . . . . . . . . . 73
8.12 Accuracy Report Wikipedia . . . . . . . . . . . . . . . . . . . . . . . . . . 73
8.13 Normalized Confusion Matrix Jobs Model . . . . . . . . . . . . . . . . . . 74
8.14 Unnormalized Confusion Matrix Jobs Model . . . . . . . . . . . . . . . . 75
8.15 Accuracy Report Job Data . . . . . . . . . . . . . . . . . . . . . . . . . . 75
List of Tables

6.1 Stakeholders ................................................................. 38
6.2 Tasks for Creation of Research Report ................................. 39
6.3 Tasks for Development and Deployment of Prototype ................. 39
6.4 Tasks for Improvement of Prototype ................................... 40
6.5 Creation of Dissertation .................................................. 40
6.6 Risk Analysis ............................................................... 41
Abbreviations

NLP       Natural Language Processing
NLU       Natural Language Understanding
MoSCoW    Must have, Should have, Could have, and Won’t have
SoDIS     Software Development Impact Statement
Dedicated to those that do not know what they want, yet know they want something different...
Chapter 1

Introduction

1.1 Introduction

1.1.1 Motivation

Steria [2016] conducted a study regarding the state of Data Science in Germany. 220 executives from companies with more than 500 employees were surveyed. Results clearly show that Data Science is increasingly becoming a concern for the industry, with 98% of German businesses investing or considering to invest in this field. However, 70% were merely investing in hardware and infrastructure and while a strong desire to change this was clearly expressed, 47% of businesses admitted that they were still in the planning phase or had not yet taken any measures whatsoever. Only 11% had done more than experiments in isolated departments. Several reasons for this were recorded. Most businesses listed institutional problems, such as a shortage of skilled employees. Another major concern, regarded the state of technology, more precisely its effectiveness, the suitability of data and an uncertainty concerning the extent of work involved. This lack of direction and empirical evidence of success is a major obstacle to be overcome in German economy.

Additionally, most research of this kind in Germany only considers large scale enterprises and is done by consultancies with direct economic interest in proving success of their services. It seems logical that this is problematic especially for companies with smaller budgets and even less proof of effectiveness available to them.
1.1.2 Aims and Objectives

This thesis aims to provide a field study for the use of Data Science to further the business goals of a typical small to medium sized German company. ”Jobsuche Regional” is a provider of 60 employment websites, each specific to a region in Germany, Austria or Switzerland. As of 26.01.2018 the company has 1.7 million site hits, 210,677 Facebook fans, 17,827 followers on Twitter and 10,973 subscribers for e-mail job listing updates.

”Jobsuche Regional”s business model is primarily based on employers paying to upload their job postings to the various job websites. Further, the company collaborates with several other providers of employment websites and profits from redirecting users to their websites.

As a result, ”Jobsuche Regional” is interested in a variety of insights that can be gained from their data:

- The company wants to increase the number of employers uploading job postings to its websites.
- The company wants to provide better service to their users, such as an optimized search process.
- The company wants to gather insight on the effectiveness of previous marketing strategies.
- The company wants to gain insight on how to direct users in order to maximize time spent on their website, while still ensuring they eventually leave for one of the partner websites. This is in their interest because ”Jobsuche Regional” hosts external content and profits from link clicks on it. However, whenever users access that content, they leave the ”Jobsuche Regional” website. As a result, the company ideally wants users to first access as much content as possible staying on their website, but then eventually leaving via a high value link.
- The company wants to gain a clearer understanding of the segmentation of their users with the primary focus being to spot real users from bot traffic.

This study focuses on developing a typical addition to their technology stack using data technologies. Recommender systems have proved largely successful for companies such as Netflix and Amazon.

We aim to use an recommender system to fulfill three of ”Jobsuche Regional”s objectives. By using the recommender system to provide a better service to the users, mainly
through improving the search functionality, we will direct users through the website in a desirable manner. As a result, we hope to increase the number of job listings posted by employers by improving the attractiveness of the website via better functionality.

1.2 Customer Journey

This section aims to provide a clear understanding of the product of "Jobsuche Regional", by presenting and discussing screenshots of the website, depicting a typical customer journey.

Figure 1.1 shows the landing page of "Jobsuche Regional". Considering structure, it is a very simple page. The box labelled "2" on figure 1.1 shows references. This design is deliberate, because the actual concept of "Jobsuche Regional" is having an individual website dedicated to a single town or county within Germany. These can be accessed via a selection Menu in the website header shown on "1" in figure 1.1.

Figure 1.2 shows the selection page of "Jobsuche Regional". Using "1" users can use a postal code to find a specific region, "2" is a selection menu based on the 16 different counties in Germany and "3" directly offers individual websites. However, most users do not actually reach their regional website through this selection menu. "Jobsuche Regional" is operated by a company with specialises in search engine optimization, which uses its expertise in this field to market its product. The unique selling point of
"Jobsuche Regional" is that they directly target Facebook users for each individual job and region and optimize for search terms such as "Jobs in Berlin". As a result, most users directly reach the website covering their region through either Google or Facebook.

No matter what channel users enter one of the company’s websites, there are some variations of the webpage shown in figure 1.3. In the top, users can enter search terms and select other relevant options. Users will then be shown the most fitting jobs listed on the website. However, on the screenshot the difficulties for "Jobsuche Regional" in implementing a new recommender system become clear. Even the search function is less than adequate - and the recommender system as well as the search function are supposed to be based on the same back-end technology.

For instance, the search term "Koch", which is the german word for "chef", of the eleven listings shown only two are actually jobs for chefs, indicated by a red box on figure 1.3. The other nine job listings are a heterogeneous mix of entirely different fields, ranging from team leading positions for logistics to a professor position at a business school.

This is the case because the company currently uses a simple, keyword based approach to make job recommendations. This approach counts the number of times the char-pattern is found within the HTML document. Therefore, if a user would search for the term "developer" the term "softwaredeveloper" would also score one hit, the imaginary word "Softwaredeveloperdeveloper" would score twice, even though it is one word.
This large number of false positives has multiple sources. The most common one is that many professions are commonly part of names within the German language. Thus, all job listings of a company which is named after its founder "Mr. Baker" will show when users search for positions as a baker. The same applies simply if the human resource contact person responsible for the listing has a name or e-mail address containing a matching word. Further, if the perks or responsibilities of a job include the search term, the job will be listed as well. In this example, "Koch", a management position is listed because a personal chef was mentioned in the job description as a job perk. Another example: A social worker position is listed because its list of responsibilities includes supervising people with disabilities while they prepare their own meals. The list of similar reasons is long and extremely varied, for example one of the jobs is listed because it is located in a village called "Unterkochen".

Figure 1.4 shows these false positives specifically. With "1" showing instances of the term being part of the job title, "2" showing instances where the term is part of the job description and "3" showing the term being part of a completely different word which actually refers to a specific region.
1.3 Hypothesis

The data of "Jobsuche Regional" can be used to build a recommender system based on natural language processing that can show more relevant job listings to the user than the current system.

1.4 Challenges

In the process of gaining insight on the business data in order to build a recommender system, there are several challenges which need to be addressed:

First, the existing business data needs to be extracted, transformed and loaded (ETL). "Jobsuche Regional" provides heterogeneous data sourced from several social media platforms, Google website analytics and server-logs. This data needs to be evaluated and integrated in a uniform manner, so that it can be used for implementation of a recommender system.

Second, after the data is in a usable state, exploratory data analysis has to be done to discover correlations and other attributes of the data which we can used to train the recommender system.

Third, the insights gained from the data then need to be evaluated considering the business goals of "Jobsuche Regional". This evaluation should identify a few key points...
of interest which should then be used to tune the recommender system to the specific
needs of the company.

Fourth, after the insights from the data are aligned with the company goals, a decision
can be made on suitable machine learning techniques to be used for the recommender
system, primarily focused on feature extraction and natural language understanding.

Fifth, following the implementation of the recommender system and other processes,
these need to be evaluated. It will be challenging to measure long term business impact
accurately within the scope of this thesis. However, some effects, such as the click-
through rate, can be tested quite immediately. It seems likely that the ultimate success
of the hypothesis has to be assessed by interviewing the stakeholders at ”Jobsuche
Regional”. This seems appropriate given the background of investigating the success
and feasibility of using data science in the German industry landscape.
Chapter 2

Literature Review

2.1 Recommender Systems

Recommender systems are the primary focus of this thesis. More precisely, this dissertation is concerned with designing a special recommender system, tailored to work job hunting websites. As a starting point, a general survey of recommender systems and their uses in the industry is done.

According to Ricci et al. [2011] recommender systems are information processing systems which use data to present items useful to users. These items are suggestions which are useful in a decision process, for example in Ricci et al.’s case, the decision which countries to travel to.

2.1.1 Motivation

One problem noticed in [Ricci et al., 2011], are likely discrepancies between the motivation of service providers and the interests of users. While users are typically interested in suggestions which are most useful to them, service providers may seek to design systems which further their business goals.

From the perspective of dissertation’s project, most of the typical examples given in Ricci et al. [2011] are not directly relevant since they regard the sale of items. However, with "Jobsuche Regional", just as with similar web services such as Facebook or Amazon, there is no longer a simple customer – service provider relation. The users are not customers in the classical sense, since they do not generate income directly (i.e. nothing is sold directly to them). Instead, users generate the income by accessing content provided by human resource representatives or affiliated job hunting websites. However,
Ricci et al. [2011] state that when designing recommender systems, this is mostly a semantic distinction, as selling products to a user is functionally very similar to optimizing articles read by a user.

Of the motivations discussed by [Ricci et al., 2011], the following are relevant for this thesis:

- Increase user satisfaction: [Ricci et al., 2011] argue that giving useful recommendations will result in more enjoyment of the system among users, thus increasing system usage and acceptance of recommendations. Both are desirable when trying to have users access more content.

- Increase user fidelity: Increasing loyalty to the website will lead to a further increase in system usage. Interestingly, this in turn also enables improvement to the recommender system since stronger user profiles can be used to enhance suggestions.

Outside of the aspects discussed by [Ricci et al., 2011], the following additional motivations are relevant within this thesis:

- Increase accessed content: Instead of selling items to users, ”Jobsuche Regional” simply wants them to access content. With better recommendations it seems likely that users will access more content.

- Guide order of user interactions: As discussed in chapter 1, the value users generate differs between types of content, with some content leading users to leave the original company website. As a result, recommendations should be designed in such a way that users do not only access items, but access them in a desirable order.

2.1.2 Tasks

[Herlocker et al., 2004] describe several tasks for recommender systems. The ones relevant to this thesis are:

- Find some good items: [Herlocker et al., 2004] distinguish between the necessity of showing all good or just some good items. For ”Jobsuche Regional”’s purposes, it seems sufficient to just show some good items.
• Annotation in context: This describes giving information according to a context, for example user behaviour. As discussed in section 2.1.1, this project is more interested in influencing user behaviour than in the quality of recommendations.

• Find credible recommender: [Ricci et al., 2011] explains that some users do not trust recommendations. In fact, several users in the past have contacted the service providers regarding the credibility of recommender systems. [Herlocker et al., 2004] argues that a good way to increase faith in the recommender system is to let the user experiment with it. This seems practical, since the recommender system built here will be based on users’ declarations of interests through search terms. However, when discussing context, it seems imperative that user behaviour or other factors outside users’ ability to control do not influence the recommendations to the point where suggestions appear to be arbitrary to the user.

2.1.3 Recommendation Techniques

The problem of how to make recommendations to users has been tackled by research from several angles. Originally, they started in other fields such as [Rich, 1979] using cognitive science, namely stereotypes, [Lilien et al., 1995] using marketing models or [Armstrong, 2001] using forecasting techniques. However, [Adomavicius and Tuzhilin, 2005] proposed using a utility function to describe the usefulness of an item to a user. They describe that this is typically based on ratings, for example how many stars they would rate a restaurant with. Such a feature is not implemented on ”Jobsuche Regional”. However, [Adomavicius and Tuzhilin, 2005] explicitly state that a profit function can be used as a utility function as well, which is available in this case. Further, it seems likely that a recommendation can be considered good when the user engages with the content that is offered.

[Balabanović and Shoham, 1997] describe two main classifications of recommendation techniques. This is recognized by several other authors, such as [Adomavicius and Tuzhilin, 2005, Ricci et al., 2011].

Content-based recommendations  Content-based recommendations are based on the principle of recommending items which are similar to items the user has liked in the past. For example, if a user has rated Indian restaurants positively in the past, more Indian restaurants will be recommended. This is certainly useful and will be worked with in this thesis. However, we do not have access to an extensive backlog for the user because users typically engage with a job hunting website only for a short time and intensively rather than over long periods of time.
Chapter 2. Literature Review

Collaborative filtering [Schafer et al., 2007] first proposed to use items as recommendations which are similar to what users have liked in the past. This similarity is calculated using the utility scores for items in the past of both users. [Ricci et al., 2011] describe this as the most popular and widely implemented technique in recommender systems. However, they recognize one fundamental flaw in this approach: It cannot be used to recommend new items.

[Ricci et al., 2011] describe several enhancements to collaborative filtering techniques, of which knowledge based techniques are the most interesting to this thesis. They describe these techniques as using domain-knowledge to guide recommendations on a case-basis.

Hybrid Systems Hybrid Systems are often chosen in order to combine the strengths of both approaches while trying to mitigate their weaknesses. [Adomavicius and Tuzhilin, 2005] describe several ways to structure these hybrid systems, such as combining several recommender systems or adding content based characteristics to a collaborative approach.

2.1.4 Multidimensionality of Recommendations

The proposed models so far focus on users and items. However, [Adomavicius and Tuzhilin, 2005] propose that several other criteria could be important for a recommendation. For instance, when recommending vacations, the time of year could be an important factor. Similar considerations were made when designing the model for this thesis. For example, the behaviour of users from a different region should not be included in a collaborative approach, since they are likely to be bot traffic and provide no monetary value.

Critical comparison of the existing methods The collaborative approach is a very good fit for this thesis, as there are lots of user interactions and it can be assumed that seeking employment is a rather homogeneous endeavour. Further, there completely new employment items are not likely a common occurrence.

Nevertheless, there is no reason not to try a hybrid system. The personal history of the user is additional data that can be utilized. Moreover, it seems likely that in this case, it will largely be a simpler version of the collaborative approach.

The most important point why the hybrid system is an attractive solution is the possibility of including the multidimensionality aspect, enabling saturation of the output of the recommender system with as much additional features as are needed.
Chapter 2. Literature Review

It seems natural that this is a key improvement. For instance, job listings which have been published years ago are most likely of very little interest to users, yet it is possible that they or similar users have looked at these items in the past.

2.2 Natural Language Processing

Natural language processing is the area of research concerned with using computers to understand natural language and manipulating it to create useful tools and techniques Chowdhury [2003]. Using natural language processing in recommender systems is a known topic in research Fleischman and Hovy [2003], Kamath and Kanakaraj [2015]. It is of particular interest to this thesis since, as opposed to many other commercial applications, users do not rate the job listings advertised on "Jobsuche Regional". Additionally, they are not in a homogeneous format. As a result, trying to use traditional approaches of building recommendations seems impossible. The only usable basis are the job listings themselves, however, since they are heterogeneous, this dissertation needs to find a way to process the language itself to build our recommendation system.

2.2.1 Word Embeddings

Using common machine learning techniques would create taxonomies for different words, meaning they would directly be classified as concepts. A common problem in natural language processing is that, as the order of these taxonomies starts to increase, the amount of training increases and the language modelling becomes increasingly ungainly.

Word Embeddings is the general term for modelling language and learning its features not through the words themselves, but instead representing them as numbers. This approach originated in the vector space model Salton et al. [1975], which represents language as a vector of numbers that has evolved over the years into different ways to achieve this.

2.2.1.1 Bag of Words

Bag of Words is the most simple approach to word embeddings and largely the same as the original vector space model. Here, a simple vector is built as a histogram of words within a piece of language Goldberg [2017]. This means that individual word counts are considered as features for building a dataset.
Since Bag of Words can just be seen as a count of all words present in a sentence, it could also be understood that, instead of representing the words in a certain way, actually the information regarding the structure and order of words is discarded to simplify the representation.

However, the information present in this representation still often proves ungainly. Since each individual symbol, word or even permutation of a word often requires its own coordinate in the vector space presentation, this can quickly escalate. For instance, imagine someone wanted to use a Bag of Words approach to model the works of Shakespeare: Shakespeare’s collective works use 31,534 different words, which means we would need a vector with at least this many dimensions, in fact it would be even more due to punctuation and other language artifacts. This is problematic for an additional reason: If there was only one sentence containing 8 words, this vector would almost exclusively contain 0 counts.

In order to tackle this, several techniques can be used. One is ignoring different permutations of language, such as case, alternative spellings, punctuation and frequent words with little information value such as ”a”. Another is decreasing noise, for example in the form of misspelled words and stemming (i.e. referring to the process of reducing all words to their word stem) can reduce the number of features in a data set.

2.2.1.2 N-Grams

Building on the limitations of the Bag of Words approach, N-Grams were formed. An N-Gram is a sequence of N number of words, called a gram Jurafsky and Martin [2009]. For instance, two words would be called a 2-gram, or more commonly a bi-gram. This has the advantage of reducing complexity significantly in comparison to Bag of words. It even starts to incorporate semantic meaning, for example ”Shut Off” would be a single coordinate with already clear meaning. Naturally, only combinations which actually occur are used, often even only common ones, instead of every possible combination, which would increase the dimensions into complete impracticality.

In fact, Bag of Grams are already much more powerful than Bag of Words and excel other methods in terms of accuracy Goldberg [2017]. They commonly use Markov’s Models, meaning that they are probabilistic language models Goldberg [2017]. This means that they essentially try to predict the likelihood of a certain word being next in sequence. This approach can be used, for instance, in predicting which search term a user is meaning to write while they are still typing. Additionally, n-grams can be used to represent users as a histogram of their search history, masking their identity.
2.2.2 Word2vec

Mikolov et al. [2013a] created a new word embedding technique for Google. They are neural networks, but not deep learning. The interesting thing about them regarding word embeddings is that the actual embedding is more of a side product.

The basic principle is that a simple neural network is trained with a single hidden layer to predict its surrounding words. This essentially means that the output layer represents all the available surrounding words. Each node in the output layer, combined with its weight, is used to represent a word as the probability of its surrounding words.

In comparison to n-grams, this approach has the distinct advantage that meaning is actually preserved. The neural net learns interesting relationships and connections between the different words in the vocabulary.

Training works by analysing a corpus with a certain window size. For instance, let’s look at the sentence “I want to receive an A for my report.”. If the neural network is supposed be trained to predict all words within a distance of 2 to each word, the training set for the word “I” in the format (input, output) would contain the following two tuples: (I, want), (I, to). For “want” it would be: (want, I), (want, to) and (want, receive).

Using this, we essentially map the meaning of a word by its context to a space. Moreover, we can actually do vector arithmetic with the vectors. This works because the neural network learned to predict the word by counting the frequency of words in its context, so the relative difference in the words ”King” and ”Queen” will be decided by the relative difference in words in their surroundings. The basic theory is that this difference is the semantic difference between these words, so if the only difference between a king and a queen would be the difference in the word ”Woman” in their neighbourhood, you would expect the only difference in their word embeddings to be the frequency of the word ”Woman”. As a result, if you were to subtract the word ”Woman” as a vector from the vector for ”King”, you would expect to point to the same location in space as the vector for ”Queen”.

This means that similar word embeddings actually point to similar locations in space, outlining areas of a broader meaning. Figure 2.1 show how simplified, the words for ”Prince”, ”King” and ”Queen” could point to a rough concept of a ”Royal Person” in space.
2.2.2.1 Doc2vec

Doc2vec is a continuation of the principles of word2vec aiming to tackle the problem of classifying whole documents Le and Mikolov [2014]. When considering word vectors as pointing to semantic concepts in vector space, keeping in mind that we can do vector arithmetic with them, the interesting part is that we should be able to extrapolate concepts we had not present in our training data. Meaning even if we had not had "Queen" represented in our corpus, we would expect to model the concept of a female king by adding female words to the vector of "King".

Figure 2.2 shows how we could try to model new concepts by adding different vectors of words to each other.

As it is possible to theoretically build new concepts out of concepts in training data, we do not actually need to know the labels of the considered data to know if they are the same. For instance, if we had the vector for "King", the vector "Woman" and the vector for "Queen" but we did not actually have the labels, we would still be able to recognise that the first two vectors combined are essentially the same thing as the third. We would not know what that is, but we would know it is the same.

When identifying a large number of documents, such as in the case of thousands of job listings on "Jobsuche Regional", it can be impractical to label all data for training a classifier. Considering the above principle, we should be able to get the concept for a
whole document by summing up every single word it contains. For instance, we could define a semantic concept for a job listing for a software engineer position as the area in vector space every vectorized document of a software engineering job listing points to. By training a classifier on this area, we should be able to accurately predict all software engineering job listings. This has important implications. First, if we have a sufficient amount of hand-labelled software engineering positions, we should be able to classify whether unknown job listings are software engineering positions.

Second, and much more importantly, since we are building a recommender system we do not actually need to know the semantic meaning of each category of job listing. We simply want to know if a job listing is similar to other job listings a user (or users similar to them) were interested in in the past. As a result, we do not actually need to label any job listings using this approach, except for measuring the accuracy during initial training.

Thus, we do not need to label an impractical amount of data to make sure that a relevant number of each feature is present. Also, we do not know anything about the semantic content a user is interested in, therefore we are able to protect their privacy. Additionally, if new types of concepts, for example new jobs, emerge in the future, our model can handle them and thus scales more robustly over the lifespan of the application.

Figure 2.3 illustrates how documents can be represented as added word vectors and used to describe the same concept.
Paragraph2vec  Paragraph2vec is a fairly new topic. To our awareness there are conflicting opinions in the research community about what it actually entails. While many seem to consider it to be the same thing as doc2vec, others seem to imply that it is the solution to the problem of documents with varying length.

Limitations  Of course, there are limitation to this approach. A "Queen" in reality is not exactly the same thing as a female king. Considering that a word2vec model might have been trained on historical data, the context of a queen might certainly share some overlap with that of a king but also be used in very different contexts.

This leads us to the next problem: Each word vector basically claims to have a universal unifying semantic representation of a word, which is exactly the same in every context. This of course not true, words change depending on their context and there certainly will be situations where a "Queen" does mean female King. However, sometimes it will just refer to a playing card. This is even more absurd as it is also semantically connected to the concept of a "King" in terms of cards, but certainly no longer relative by being female.

Another problem of this approach is that the semantic space is not necessarily continuous, meaning it can be hard to keep discrete from other spaces and as a result is prone to over-fit on test data.

**Figure 2.3:** All word vectors from two documents are added together, building two vectorized documents.
Considering that documents often are of varying length, the vector representations might point to vastly different places in space. Keeping space no longer discrete, even with a perfect classifier, if the length of the document is not considered.

As a result, doc2vec has to compete with other neural computing approaches to word embeddings [Xu et al., 2017].

### 2.2.3 Evaluation

This section aims to inform the evaluation process of the thesis. First, we evaluate the models we trained from the corpora. Then, we evaluate the classifiers. Lastly, we measure the impact of our recommender system.

### 2.2.4 Exploring the Model

Our models are based on word embeddings. These are large dimensional vectors. The standard method for evaluating word embeddings is cosine similarity, as this has been shown to map semantic meaning found in language Mikolov et al. [2013b]. Cosine similarity is an expression of the similarity between two vectors, where the cosine of the angle between them is calculated. Therefore, only their direction in space, but not their magnitude, is taken into account. This means we can evaluate our model by exploring how similar vectors for individual words are.

Moreover, the direction between two vectors, meaning the subtraction of them, was found to carry semantic meaning Mikolov et al. [2013a]. This is further useful in exploring our data set, as we can use this to test if our model has learned concepts. For example, removing "Man" from "Woman", leads to the concept of "Female". Or, even more interestingly, this can be used to evaluate if two terms are linked by the same semantic concept. For example, if subtracting "Teacher" from "School" leads to the same vector as subtracting "Mayor" from "Town Hall", that is an indicator for the model having learned and understood the concept of a workplace.

### 2.2.5 Classifiers

For evaluating our classification model we will use a confusion matrix. It matches the results of a classifier into the a matrix in a way that each row shows the prediction results, whereas each column symbolises the actual class Powers [2011]. This shows us whether, and in what way, the classifier "confuses" those two classes.
From this, certain measures of performance can be calculated to evaluate our model. We will focus on the following measures: Precision, Recall and F1-Score. Powers [2011] explain them as follows: Precision describes the ratio of true positives among our predicted positives. This is important in order to know how selective our classifier is. Recall describes the ratio of positives we find, meaning it shows how much of content interesting to the user we access. The F1-Score is the harmonic average over recall and precision, as such it is a unifying measure of performance of both these points disregarding true negatives. It is widely used for performance measures in machine learning and especially in natural language processing Tjong Kim Sang and De Meulder [2003].

2.2.6 Interrater-agreement

For labelling our data, we are using human participants. The agreement between them is measured using several well-known algorithms such as Cohens Kappa Cantor [1996], Fleiss Kappa Fleiss [1971], Krippendorfs Alpha Krippendorff [2013] and Scotts Pi Scott [1955]. They all inform how big the agreement between those labelling the data was, a higher value indicating validity of the test set. They are part of nltk library for python.

2.2.7 Visualization

The advantage of both vectors and confusion matrices is that they can be easily visualised to help in evaluation. While the confusion matrix needs only to be shown as a table, the situation is a little more complex for word embeddings. They are vectors which naturally lend itself to visualisation. Moreover, the meaning of the word embeddings is actually signified by their relative positions in space, which makes it perfect for visual evaluation. This is routinely done in the machine learning community and research, such as Smilkov et al. [2016].

However, word embeddings have a large amount of dimensions, 300 in our case. This is hard to visually grasp for a human. As a result, the vectors need to be reduced to map them in two-dimensional space. A common technique used for this is principal component analysis Smilkov et al. [2016]. Principal component analysis is a relatively straight-forward technique, where orthogonal transformation is used to reduce the dimensionality while aiming to preserve as much as possible of the variability between those vectors Pearson [1901].

This is especially useful for identifying the larger context and semantic meaning of word embeddings Smilkov et al. [2016], which is our main interest. However, we do not actually use principal component analysis to gather any insight directly. It shows problems with
identifying clusters or building a correct direction between vectors Smilkov et al. [2016]. However, this is not what we are doing, as we are evaluating calculations first on the full dimensional vectors using cosine similarity and vector subtraction and are only using principal component analysis to visualise those results for the better understanding amongst human readers. It is not used to derive any scientific meaning.

While all the models and calculations are done by the authors of this thesis, he would like to thank Mueller [2015] for his code for easily visualising word embeddings using principal component analysis and scikit learn [2018] for their code on visualising confusion matrices. While they and the data they were built from were all created by the author of this study, all visualisation of vectors using principal component analysis and confusion matrices where generated using these two pieces of code.

2.2.8 Measuring the Impact of the recommender system

For measuring the impact of the recommender system, we will simply be using Google Analytics. Google Analytics tracks how much money was earned and how often each URI was accessed. We will measure how the amount of jobs that were accessed and how much monetization was influenced after our prototype compared to before our prototype.
Chapter 3

Methodology

This chapter is dedicated to a multitude of aspects concerning this thesis. It outlines the experiment design from a theoretical point of view and discusses the technologies chosen for it. Further, it aims to provide the goals of the experiment implementation as artefacts of the software architecture design process.

This chapter is difficult to place within this document, since of course it is being informed by the literature review and requirements analysis. However, this was more of a fluent, bi-directional approach. Also, a lot of the technologies are specifically chosen to satisfy relevant professional, legal, ethical and social issues. Chronologically, this is placed early in the project planning phase 6. Stakeholders were identified first and used to gather requirements, the tasks and timeline were informed by the methodology.

3.1 Development

3.1.1 Building the recommender system

Figure 3.1 shows the experiment design. Documents in the case of our project are job-listings posted on ”Jobsuche Regional”. First, we need a word2vec model. We can either train a new one using a selection of available documents or from an existing corpus. However, we can also use a ready trained word2vec model, such as Google’s. Which method will prove superior will have to be discovered in early experimentation. A possible problem which could arise at this point lies in finding a good model or corpus for the German language, as English is much more popular, leaving us with no choice but to train our own model.
After that, we will need to vectorise all words within our documents and add them to build vectorized documents. These documents will not be semantically identifiable, but rather we will hopefully be able to train our model to recognize documents with similar meaning, as explained in 2. We will test this by hand-labelling several documents and training the doc2vec model on these documents.

In order to build a hybrid recommender system, we will need to build a content-based and a collaborative system. For both, we will need to represent the user. This will be done by using all terms the user has searched for in the past as a counter, although these will likely be too many features and will have to be drastically reduced in later experimentation.

Possible ways to reduce features will be selecting those with the highest correlation coefficient, merging semantically similar or identical terms or clustering similar professions closer to each other.

For the content-based system, we will train such a user model combined with a vectorized document and the binary class attribute stating whether the user has clicked on the document or not. In order to train this, we will experiment with traditional machine learning methods, trying to reach as high an accuracy as possible within a set timeframe.

For the collaborative-based system, we will consider similar users. In order for this to work, we will need a similarity function modelling how similar users are. Since all users will be counter-based, a heuristic used in search algorithms, such as the Manhattan function, could be used to model the ”distance in similarity between users. Afterwards, we will train predicting whether our given user has clicked on an document only by looking at similar users. We can later use this to predict whether users are likely to click on these documents.

Finally, we can build our hybrid system by building a new data-set, with the first two attributes being the confidence-scores produced by both our individual systems. As mentioned in 2, it is common practice to add additional features as attributes at this point. Possible candidates could, in this case, be the date the job listing has been posted or the term used by the user to avoid over-training on the users and similar users’ histories, making recommendations which may be good but hard to understand for the user.

Experimentation at any of these stages will be limited by time as scheduled in 6, trying to maximize accuracy in the meantime. However, if accuracy fails to beat the existing keyword-based system, we will need to cancel parts of the project in order to try to improve at this stage.
Figure 3.1: Visualization of steps involved in building the recommender system.
3.1.2 Proposed additional systems

In order to test our recommender system, we will produce a prototype to integrate on one or more of the different websites of "Jobsuche-Regional". In an interview with their chief developer, three APIs were formulated, which we can integrate into the website. We can later measure whether click-through rates, traffic and ratio of content with high monetization value accessed have changed during deployment of the prototype.

It is important to note that only these APIs will be provided by the author of this project and any necessary integration will be done by the company.

3.1.2.1 Conversion of job-listings to documents

3.2 shows the first API to be implemented. In order to work with vectorized documents, all job-listings must be vectorized naturally. In order to achieve a simple API will be developed, which will be called with a job-listing. Inside our model, the document will be vectorized and stored in a separate data-base specifically for vectorized documents.

This API will initially be called on all job-listings with a script and later automatically be called every time a human-resource user uploads a job-listing to "Jobsuche Regional".

![Figure 3.2: Creating vectorised documents whenever new job listing is created.](image-url)
3.1.2.2 Integration of the recommender system into the search function

In order to use the recommender system within the on-site search function, we will need another API. There are several things to keep in mind. People have the choice whether or not to log-in with a personal profile on the website, thus we may not know the user. Also, while the user types, we may not recognize the search term. We want to start recommending as soon as we either know the user or can make a confident guess towards the term the user wants to search for. We will not recommend anything if we do not know either.

3.3 shows the architecture for this second API. It will be called with the current user and search term, every time the search term changes. In our module we may try to guess search terms using n-grams or simply wait for a signal that the user is done searching. We will transform the user and look for similar users within our module.

If we find a good set of recommendations we will return this via the API.

![Proposed System II vs Existing System](image)

**Figure 3.3:** Giving all recommendations for a user, used in mailer system and when logging in.

3.1.2.3 Making recommendations just for the user

The last API we want to prototype, shown in 3.4, may look similar to the second proposed system. However, it is distinctly different because it will not be used in context of a search. This means we only have the user and show all good recommendations solely
Based on that. This API will simply be called with a user, which will be converted within our module. A list of recommendations will be returned.

This API will be used to show recommendations on the site, for instance on the home-screen when a user logs-in or when a user signs up for the email-based recommender.

![Proposed System II vs Existing System Diagram]

**Figure 3.4:** Giving all recommendations for either a user, a search term or both, depending on what is available when search function is used.

### 3.2 Technologies

In this chapter, the different technologies which will likely be used within the experiment design are discussed.

**Google Analytics**  Google Analytics is a free service offered by Google to track traffic on websites. It has been integrated on "Jobsuche Regional" websites for years and as mentioned in 1 has gathered a lot of data usable to train the recommender system on. Google Analytics has been shown to be a useful tool in guiding user behaviour Fang [2007].

**Python**  Schutt and O’Neil [2013] describe Python as a primary candidate for data science and machine learning applications, only rivalled by R. It offers lots of libraries and is well used within the research community.
Additionally, we chose python over other alternatives for our implementation because we are working with a web-application. Since python is also extensively used in web-technologies, this is a more natural choice for our prototype.

**NumPy**  NumPy is a popular library for python, specifically designed for the scientific community. It offers lots of mathematical representations such as matrices and high-level functions.

**Pandas**  Pandas is a library specifically designed for data manipulation and analysis. Most of our data conversion, cleaning and pre-processing will be handled using this library.

**Scikit vs Tensorflow**  For handling deep-learning and more elaborate machine-learning algorithms, there are two interesting choices.

Scikit is built directly on top of NumPy and is designed under the same principle of offering a wide range of functions. Tensorflow is developed by Google and much more low-level, excelling especially in creating new models by offering great control in using the GPU to ensure more efficient training.

Thus, if we need a lot of new implementations, we will use Tensorflow. Otherwise Scikit will be used if it proves sufficient for our needs.

**Gensim and NLTK**  While, as mentioned in 6, we already trained a small corpus from scratch as proof of concept for the feasibility of our approach, for the actual implementation we will want to handle word vectorization with a specialised library. Gensim is a Python library, optimized to work with NumPy specifically for word vectorization. A natural fit for this project.

If we need a more general approach to natural language understanding due to insufficient performance, we will move to NLTK, a more general natural language processing library for python.
IEEE [1990] defines the requirements analysis as the process of analysing user needs in order to formulate a definition of the requirements of the system. It further describes an requirement as one of three things:

1. Something a user needs to reach an objective.
2. Something needed to meet a formal agreement.
3. A documented representation of 1 or 2.

Requirements are important in order to gain a clear and measurable overview of what certain characteristics of a system are supposed to deliver. Notice the strong context with user desires and documentation. It is the first step in software development and as much of a communication tool as it is a planning tool. It helps different stakeholders to get the same idea of the project.

As such, identifying the stakeholders is usually the first step in formulating clear requirements. In this case, this has already been done as part of the Software Development Impact Statement in the project planning phase in table 6.1.

4.1 Requirements for the Recommender System

In order to analyse the requirements for the recommender system, we are using the MoSCoW method Clegg and Barker [1994]. This is a popular method used in many fields, including software development. The acronym stands for Must have, Should have, Could have, and Won’t have. These are quintessential categories which will be classifying the different requirements.
This clear separation is very useful for prioritising the right features and front-loading effort on the critical and appropriate places. However, as mentioned earlier, one of the main reason for a requirement analysis is actually communication between the stakeholders.

Communication and expectations have been identified as key risks for this project in table 6.6 because of its dual nature of academic and industry requirements. As such, the MoSCoW method seems especially appropriate, because it separates requirements very clearly in workflow-appropriate classes. It may intuitively seem most relevant what requirements the recommender system must and should have. However, this method has precisely been chosen because it enforces discussing what requirements it won’t have at the beginning of the project, preventing misleading expectations and conflict of interests from slowing down the process along the line. In much a similar manner it is very important as a tool in respecting the boundaries of all stakeholders and keeping ethical responsibility, which is extensively discussed in chapter 5.

As discussed in risk mitigation in table 6.6, ”Jobsuche Regional” is aware and has agreed that in the worst case scenario, the academic requirements of the dissertation will take priority within the timeframe. However, the work from this project can certainly be continued solely under industry conditions after the dissertation. As a result, within this project, ”Must have” should be seen as requirements that need to be fulfilled that neither society, the users or the business of ”Jobsuche Regional” is actively negatively effected by. ”Should have” shows the desired results for this project. ”Could have” shows what is allowed to happen and what could be done if time allows for it. ”Won’t have” are requirements which could be incorporated into future work.

All requirements have been formed as part of the research done in chapter 2, as dictated by the ethical issues chapter 5, required by the project plan chapter 6 or as part of interviews with the stakeholders.

**Must have:**

- Keep the anonymity and integrity of private user data as discussed in chapter 5.
- No major disruption of the productive systems in a way that will harm the business model of ”Jobsuche Regional”.
- Make recommendations for job listings relevant to users or recognize insufficient confidence defaulting to the system in use right now.
Chapter 4. *Requirements Analysis*

**Should have:**

- Use word2vec to build a model for natural language processing capable of identifying similar job listings. The overall accuracy is not as much of a priority as that a full page of recommendations, namely 10, can be made with high accuracy (80/)
- Use statistical models or neural computing to give recommendations of greater accuracy than the current system (not yet tested).
- A bug-free and modular integration in the productive system for testing-purposes as described in chapter 3.
- After being fully trained, produce recommendations in under 1000ms.
- Increase the click-through and conversion rate in Google Analytics.
- Increase the ratio of content with higher monetization value, accessed by users.
- Work for all job listings in all fields offered on the web applications.
- One test-period of prototype on productive system.

**Could have:**

- Real time updating of recommendations while typing (possibly using n-grams to predict the search term before the user has finished typing).
- A small testing suite, such as unit tests, to easily assure proper functionality of the models.
- Feedback for human resource users how likely their job listing is to attract users.
- Incorporate additional personal data via the Facebook API to improve recommendations, as long as the user remains unidentifiable.
- Additional features as discovered after first deployment on productive system.
- If the recommender system cannot be trained to handle all job-listings, either a valuable and practical subset could be chosen. If this fails as well, a proof of concept for one specific type of job listing.
- Use recommender system not only on site, but also for the automatic e-mails generated and sent to users.
Won’t have:

- A fully tested and robust long-term integration into the company technology stack.
- A web-crawler capable of searching the web for job-listings to incorporate on the web-application.
- Full integration to build recommendations from linked Linked-In or Xing profiles.
- Testing on E-Mail based job recommendations.
Chapter 5

Professional, legal, ethical, and social issues

This chapter discusses all of the professional, legal, ethical and social issues which might arise during the project. Each area will be discussed individually, although they can influence each other, they should largely be seen as separate and disjunct.

5.1 Professional Issues

Gotterbarn [2002] noticed that students continuously fall short in regards to professional issues in software development. This leads to misunderstandings in judging the full impact of their projects, both on their role within the larger system and their impact on society. They describe the biggest factor in leading to these shortcomings as considering professional issues not early enough in the development process, with many only be when they start testing their developments.

Further, the British Computer Society BCS [2015] has published a code of conduct. It is specifically also applied to students.

As a result, this study will take measures to assure a high standard of professional conduct. The professional impact will be considered early, before development even starts. This will be done using the "Software Development Impact Statement", as shown in figure 6.1. In accordance with this, it will also be broadened in regards to additional project planning to include societal and ethical issues in this section. Further, this study is committed to consistently adhere to the code of conduct of the BCS.
Lastly, special care will be taken to needs of the company this project is associated with. This includes trying to minimize impact on business processes to a minimum. Beyond that, special care will be taken in the documentation of the process and presentation of the code, as this may be the only reference available to the company in the future. However, it should be noted that this study is a prototype and proof of concept. Thus, focus will remain on technical feasibility, rather than the needs of implementation.

5.2 Legal Issues

The biggest consideration regarding legal issues involved in this project is the data protection of users of the web application. "Jobsuche Regional" has extensive user agreements Webfeinschliff [2017], and users are informed of every use of their data relevant to this project. The following is a rough summary, translated from the legal document provided in German:

- Users are informed that Google Analytics is used within this website, that it is used to analyse their behaviour and that this data is saved on servers of Google. In rare circumstances, their full ip-address is saved, however, never when they specially request anonymous ip addresses. This data is not used with other Google services and users are informed about how they can prevent their browser from storing cookies. They are further informed that this may lead to a decrease in quality of service.

- Users are informed that cookies are used to recognize them as the same user when they login multiple times. This cannot be used to discover their identity.

- Users are informed that their user data is used to create a hash recognizable by Facebook. This is done to enable re-targeting and marketing in accordance with Facebook APIs, while again ensuring that the identity of the user cannot be restored.

- Users are informed that the double-opt-in method is used for contacting them via e-mail. This means they have to specifically allow, via e-mail, that they want to recieve further e-mails and that they understand that it is possible to opt-out of this agreement at any point in time.

- Users are informed that the data associated with monetary transactions is only used for accounting purposes and never shared with third parties.
All software developed in this project will either be owned by "Jobsuche Regional" or be open source, such as python. Any software developed within this project will be owned by the author of this document and as such, full control is maintained.

5.3 Social Issues

The purpose of this project is to use technology to help efficiency in people finding employment. While this does seem intuitively positive, this does indeed not come without issues.

Employment is a fundamental cornerstone of our society as it is the primary source of income for most people. Therefore, competing for employment should be possible in a fair and accessible manner. Certain groups of society experience harder access to technology, either because of physical disability, poverty or limited exposure. If they cannot access the most efficient means of seeking employment, this creates social issues.

Wallace and Sheldon [2015] discuss ethical parameters in business research. They outline four categories, one of which is justice. In this category, they stress the importance of accessibility to participate in research. This is vital, because if research is inaccessible, it fails to include these people and consequently prevents society and science from helping them to a certain extend.

A similar argument can be proposed for the technology in this project or similar projects like it. If there is an unreasonable barrier for people to participate in having their data analysed, because they cannot or choose not to, they are barred from its benefits. If companies use data to analyse customer needs, the needs of these potential customers will remain unmet. Moreover, if the technology in this project becomes successful to the point where companies will use the user behaviour to make employment more attractive to potential candidates, again these people will not be taken into consideration.

? has described the filter bubble - a concept describing that users tend to be isolated when software attempts to filter content for them. Originally, ? had focused on services such as Facebook or Google presenting news, noticing that users mostly saw news related to their expressed political stances, meaning they only interacted with these types of news, reinforcing the filtering algorithm leading to the range of political perspective being shown to them decreasing steadily over time. The same concept has been observed for recommender systems ?. This would be problematic, if for example our hybrid systems would develop sexist tendencies leading to certain professions only being recommended to a specific gender. However, ? found that this effect is relatively small for recommender systems, especially if users regularly interact with the content. To alleviate this problem,
content actually accessed by the user will be weighted more highly and new content will randomly be represented to users in this project.

Building on this problem, we are using natural language understanding. The German language is inherently sexist towards male generics. In German grammar, all nouns have a clearly determined gender. For instance, the word "Schreiner" means carpenter and the word "Schreinerin" means female carpenter. Commonly, however, when not talking about a specific entity, the masculine version is used generically. The author explains that these male generics lead to achievement being associated with males and argues that over the last 30 years, it has been proven that women are put at an distinct disadvantage in Switzerland, Austria and Germany by this trait of the German language. Further, ? found a gender-biased link towards information retrieval from a text and long-term memory by the inherent sexism in the German language. Interestingly, word embeddings in our experiment picked up this trend as well. When assembling the centroid around a word, male generic synonyms for this word were ranked as more similar than the generic female version of the identical word. However, this is an extremely difficult problem to address. We are not aware of any research that has tackled the problem of inherent sexism in a corpus built from German language. Further, looking at the search terms on "Jobsuche Regional" by users, less than one percent contains female generics. As a result, it seems clear that female users are also using male generics when using the website. This means that any measures to combat this problem might be misguided and instead could put female users at a disadvantage. For all of the reasons above, we have decided that trying to include this problem would be out of the scope of this dissertation. However, we strongly endorse any efforts towards research assuring gender neutrality within natural language processing.

Lastly, the gained efficiency in matching candidates with job positions could reduce employment in the human resource sector or drive companies out of business who do not have access to these technologies.

Tavani [2003] classes the perspective on cyber ethical issues in several categories and associates disciplines within them. They argue that computer science does have a professional responsibility in this and that sociology and behavioural sciences are better suited to explore and comment on the impact of cyber technologies on governmental institutions and socio-demographic groups.

This project will try to alleviate social issues by being accessible and putting a fair weight on the preferences of different users. However, it also understands its limitations on the possible impact it might have on these larger issues.
5.4 Ethical Issues

There are two main ethical issues in this project. The first is handling user data. It is of imperative importance that users have a chance to realise and influence how their data is being used. As discussed in section 5.2, users are informed of how their data is being used, how they can prevent this from happening and what impact this has on the services provided to them.

Further measurements are taken to handle the user data with care. As outlined, all information regarding personal identities, such as redirects from Facebook, are already hashed. However, of course, it still might be possible to identify a user by their location, the type of employment they seek or other information available to us. Thus, in order to preserve the use of personal data, the data is transformed to make this impossible.

Users will only be represented by their behaviour on the site, specifically by what they searched for and which job listings they accessed. Further, job listings will lose all semantic meaning after they are vectorized. The only information which will remain is how unknown users are best matched to patterns of text with unknown content.

Further, all the software and data will only be stored on sufficiently secure hardware and will not be accessible from outside networks.

The second ethical issue is the discrepancy in company and user motivations. The recommender system will not necessarily provide the content of most interest to the users, but the content which will be most profitable to “Jobsuche Regional”. However, a large overlap in these motives is expected and as Ricci et al. [2011] explain, this is common practice in recommender systems.

This project, regardless, will focus on not going into extreme directions with this. Both the author and the associated company are committed to giving good and interesting recommendations to help people find better employment with less effort. Only within recommendations believed to be useful and interesting to the user will monetization be used to select further.
Chapter 6

Project Plan

As mentioned earlier, when professional issues in section 5.1 were discussed, a proper project plan in the early stages of a project is of vital importance Gotterbarn [2002].

As a result, this project used the Software Development Impact Statement Gotterbarn and Rogerson [2005], which is a tried and proven method, specifically within thesis work of students. Based on environmental impact statement, it enhances the traditional software risk-analysis process by including qualitative metrics, such as identifying the stakeholders and formulating how the tasks involved in the project affect them. This is in accordance with the commitment of this study to work mindful of the impact of this
Table 6.1: Stakeholders

<table>
<thead>
<tr>
<th>Role</th>
<th>Needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>Recommendations for job listings relevant and interesting to them.</td>
</tr>
<tr>
<td>Company Owner</td>
<td>Increase profit of &quot;Jobsuche Regional&quot; by increasing the amount of content and ratio of content with high monetization value accessed by users.</td>
</tr>
<tr>
<td>Chief Developer</td>
<td>Clear and modular APIs he can use to integrate the prototype within the existing system without disturbing the production system in an unreasonable manner.</td>
</tr>
<tr>
<td>Designers</td>
<td>APIs need to provide all the relevant content to present recommendations in an visually appealing and informative way.</td>
</tr>
<tr>
<td>Project Supervisors</td>
<td>Awareness of the project in an ongoing and transparent manner. Clear execution of methodology with academic significance.</td>
</tr>
<tr>
<td>Author (myself)</td>
<td>Produce a fulfilling project in both an industrial and academic sense.</td>
</tr>
</tbody>
</table>

Stage 1  The project is a recommender system. In SoDIS Stakeholders are classified according to their roles. Table 6.1 shows all the stakeholders identified within this project and their roles.

Stage 2  Since this is quite an elaborate project, it seems useful to group the tasks required. Four clear milestones are identifiable within the project. The tasks are shown in an individual table for each milestone. Since communication with the company is required and the company is unable to plan too far ahead, especially due to a rather small in-house IT force, the dates for Milestone 2 and 3 require a certain flexibility in order to meet company approval.

- Milestone 1, shown in table 6.2: Creation of Research Report. Due 12th of April (firm).
- Milestone 2, shown in table 6.3: Development and Deployment of Prototype. Due 18th of June (preliminary, awaiting company approval.)
- Milestone 3, shown in table 6.4: Use data gathered from the prototype to improve it. Due 18th of July (not-firm, but will cut into Milestone 4.)
- Milestone 4, shown in table 6.5: Creation of Dissertation. Due 9th of August (firm).
<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature Review</td>
<td>Create an overview of existing research relevant to this project and take conclusions.</td>
</tr>
<tr>
<td>Create Proof of Concept</td>
<td>Create simple implementations of the larger themes required in this project in order to create an informed methodology.</td>
</tr>
<tr>
<td>Requirements Analysis</td>
<td>Formulate an analysis of the requirements of the project. Use this to better inform decisions on project plan and methodology.</td>
</tr>
<tr>
<td>Create Methodology</td>
<td>Identify and describe technologies to achieve the aim of this project. Plan and describe modules which need to be developed.</td>
</tr>
<tr>
<td>Explore Professional, Legal, Social and Ethical Issues</td>
<td>Research and formulate all of PLES issues in a meaningful manner.</td>
</tr>
<tr>
<td>Create Project Plan</td>
<td>Follow SoDIS and create a clear timetable for the rest of the project in form of a Gantt chart.</td>
</tr>
</tbody>
</table>

**Table 6.2:** Tasks for Creation of Research Report

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train Word2Vec</td>
<td>Explore different ways to create a meaningful corpus to build a word2vec relevant for this project.</td>
</tr>
<tr>
<td>Vectorize Documents</td>
<td>Vectorize all job listings in &quot;Jobsuche Regional&quot; for further use.</td>
</tr>
<tr>
<td>Test Accuracy of vectorized documents</td>
<td>Hand-label some job listings and test accuracy of vectorized documents in classifying them.</td>
</tr>
<tr>
<td>Create User Representation</td>
<td>Build data instance of a user by choosing attributes from the system.</td>
</tr>
<tr>
<td>Build Content-Based Recommender System</td>
<td>Use machine learning techniques to train predicting whether a given user representation would click on a given vectorized document.</td>
</tr>
<tr>
<td>Build Collaborative Recommender System</td>
<td>Use machine learning techniques to train predicting whether a similar user and a score to represent how similar he is to the user in question would click on a given vectorized document.</td>
</tr>
<tr>
<td>Build Hybrid Recommender System</td>
<td>Use output of both recommender systems and possibly identify additional attributes in training a system to give the final recommendation.</td>
</tr>
<tr>
<td>Testing and optimizing the prototype</td>
<td>Measure accuracy and experiment with parameters of previous tasks until a satisfactory accuracy is reached.</td>
</tr>
<tr>
<td>Develop the three APIs</td>
<td>Develop the three APIs in accordance with chapter 3.</td>
</tr>
<tr>
<td>Deployment on productive system</td>
<td>Since they already have one web application dedicated to testing, the APIs need to be integrated on it and tested.</td>
</tr>
</tbody>
</table>

**Table 6.3:** Tasks for Development and Deployment of Prototype
### Chapter 6. Project Plan

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gather data from prototype</td>
<td>The prototype is planned to be deployed for two weeks on the productive system to gather data.</td>
</tr>
<tr>
<td>Evaluation of Data</td>
<td>Evaluate the data gathered during employment on how click-rates and monetization value of users have changed.</td>
</tr>
<tr>
<td>Analyse how to improve prototype</td>
<td>Try to find ways in which performance of the prototype can be improved, comparing it to the system without the prototype.</td>
</tr>
<tr>
<td>Create and test patch to prototype</td>
<td>Modify the prototype, for examples classifiers used or features extracted and how they are modelled.</td>
</tr>
<tr>
<td>Apply patch to prototype on productive system</td>
<td>Apply the patch on the productive system and monitor performance.</td>
</tr>
</tbody>
</table>

**Table 6.4: Tasks for Improvement of Prototype**

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluate gathered data</td>
<td>Evaluate all the data gathered while building the system. First deployment of prototype, compare to systems without prototype. After deployment of patch, compare to first prototype and systems without prototype.</td>
</tr>
<tr>
<td>Present data in a meaningful way</td>
<td>Create informative and clear presentation of results</td>
</tr>
<tr>
<td>Create Conclusion</td>
<td>Discuss how the results compare to similar studies and what insight they offer in regards the objectives of this project.</td>
</tr>
<tr>
<td>Present Limitations</td>
<td>Discuss the limitations and shortcomings of this project and comment on them.</td>
</tr>
<tr>
<td>Discuss Future Work</td>
<td>Show what further approaches could be used to build on the work of this project.</td>
</tr>
</tbody>
</table>

**Table 6.5: Tasks for Creation of Dissertation**

Stage 3-4  Gotterbarn and Rogerson [2005] recognizes that stage 3 and 4 are fluent and can be addressed simultaneously and that even movement between them is possible. The general goal is to formulate risks in the following way:

\[
\text{Might} \ [\text{task}] \ [\text{effect}] \ [\text{stakeholder}]
\]

These risks are associated broad tertiary risk-level: critical, significant, minor and might be given a probability. In any case, a solution is proposed. Table 6.6 shows the risk analysis for this project.
<table>
<thead>
<tr>
<th>Task-Effect</th>
<th>Stake</th>
<th>Sev</th>
<th>Prob</th>
<th>Solution</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Might milestone 2 be delayed due to the companies fault, causing project delay for the</td>
<td>author</td>
<td>significant</td>
<td>Medium</td>
<td>Move back timeline, be prepared to cancel Milestone 3 and only evaluate prototype, use extra time for more in-depth development and testing of prototype.</td>
<td></td>
</tr>
<tr>
<td>Might milestone 2 be cancelled due to the company not having the resources or termination of the collaboration for other business related reasons, causing forced cancellation of milestone 2 and 3 for the</td>
<td>author</td>
<td>critical</td>
<td>low</td>
<td>Move focus of the study on the academic and machine learning side. Put additional time towards creating a more solid and accurate doc2vec method, possibly enhanced by research on additional methods. Move hypothesis towards using doc2vec to classify job listing in a more general manner.</td>
<td></td>
</tr>
<tr>
<td>Might Create User Representation lead to too many attributes, causing to a loss of accuracy for the</td>
<td>author</td>
<td>minor</td>
<td>high</td>
<td>Reduce number of attributes, merge similar attributes and try to investigate related attributes and move them closer to each other. Worst case: show proof of concept for smaller number of different job listings.</td>
<td></td>
</tr>
<tr>
<td>Might milestone 2 create a prototype of such a low accuracy that it will harm the business of the</td>
<td>company</td>
<td>significant</td>
<td>medium</td>
<td>take extra time on milestone 2, possibly cancelling milestone 3. If complete failure cancel milestone 2 and 3 and investigate and discuss the academic significance of being unable to build a suitable doc2vec model for job listings.</td>
<td></td>
</tr>
<tr>
<td>Might milestone 2 or 3 create a model so focused on the past behaviour of the user and similar users that it does not or does barely take the search term into account, causing low-trust and credibility for the</td>
<td>users</td>
<td>significant</td>
<td>medium</td>
<td>force algorithm to take search term into account at the cost of accuracy. Potentially re-confirm semantic suitability with labelled job-listing data.</td>
<td></td>
</tr>
<tr>
<td>Might milestone 2 or 3 create a model so focused on the past behaviour of the user and similar users that it does not or does barely take the search term into account, causing low-trust and credibility for the</td>
<td>users</td>
<td>Significant</td>
<td>medium</td>
<td>Force algorithm to take search term into account, even possibly at the cost of accuracy. Potentially re-confirm semantic suitability with labelled job-listing data.</td>
<td></td>
</tr>
<tr>
<td>Might milestone 2 or 3 take up so much time that academic meets are not meet, or vice verca causing insufficient time for the</td>
<td>author</td>
<td>significant</td>
<td>high</td>
<td>open communication about expectations. The company is informed and has agreed that in the worst case, the academic requirements must have priority. The work on the prototype will be continued and</td>
<td></td>
</tr>
</tbody>
</table>
6.1 Timeline

This section illustrates the timeline for this project using Gantt charts. Figure 6.2 shows one Gantt chart for the entire project, having all tasks merged into their respective milestones. Figures 6.3, 6.4, 6.5 and 6.6 show an individual Gantt chart for each milestone. This was done since a Gantt chart with all individual tasks proved to be of little informative use.

**Figure 6.2:** This Gantt chart illustrates all the 4 major milestones for the project.

**Figure 6.3:** This Gantt chart illustrates the tasks for Milestone 1. (Completed)
Figure 6.4: This Gantt chart illustrates the tasks for Milestone 2.

Figure 6.5: This Gantt chart illustrates the tasks for Milestone 3.
Figure 6.6: This Gantt chart illustrates the tasks for Milestone 4.
Chapter 7

Implementation

In this chapter, all parts of the dissertation that required implementation are described in chronological manner. The process started with evaluating and cleaning the data provided. Then, two corpora were built, one using the actual data taken from "Jobsuche Regional" and the other from the complete German version of Wikipedia. Afterwards, a first prototype to be deployed on "Jobsuche Regional" was developed and sent to the company. Unfortunately, after several weeks of difficult communication it was brought to my attention that a medical emergency with their chief developer made it impossible for the company to continue with the project within the designated time frame of the thesis. As a result, the planned for risk in table 6.6 of the project being delayed due to the company’s fault was invoked. Milestone 3 was cancelled and instead, the produced work so far was evaluated more in-depth. This was especially achieved by setting up a small-scale experiment, where real humans pretended to be job seekers on the available data. Results were recorded into a test set. The original company algorithm was tested as a classifier against the new algorithm used for the prototype, using models built from both corpora. A glossary over the files provided in the package accompanying this dissertation is available in Appendix D.

7.1 Analysis, Processing and Cleaning of Data

7.1.1 Types of Data

This section discusses the state of the data received. Unfortunately, only three weeks worth of data was available, which is much less than originally discussed with "Jobsuche Regional".
7.1.1.1 User Data

The user data was collected on a per day per individual job hunting website basis. All in all, it was 12.8 GB worth of data, distributed over 3313 files. Each file contained several user actions, one line with three pieces of information per action. One such line is shown in Figure 7.1. The first item is a cut down ip address, able to uniquely identify the user, without being able to identify them individually. The second item is the date and time of access. The last item is the HTTP-Request for the user action.

![Figure 7.1: Snippet showing raw user data.](image1.png)

All in all this data was in pristine condition. It contained 1455630 unique users, with 9903300 actions.

7.1.1.2 Google Analytics

Google Analytics registers all the website traffic. Figure 7.2 shows the overview page for Google Analytics. This, however, is only a small part of what is recorded and available. Extremely thorough analysis over money made per individual website, all user traffic on a collective as well as individual basis, geographic, time data and much more is available.

![Figure 7.2: Small snippet from Google Analytics Overview.](image2.png)

However, what we were really interested in was connecting Google Analytics to our user data. As shown in figure 7.1, we could see the HTTP-request for each individual user action. This request included the URI, which could be linked to Google Analytics, where an individual report was available for each URI. Figure 7.3 shows such a report. This way, we could get critical information such as how money each URI was worth per visit, how much time an average user spent on this site or how many users left after accessing this site.
7.1.1.3 Job Data

The job data itself, unfortunately, was corrupted while it was being prepared by "Jobsuche Regional". I informed the company immediately, however, "Jobsuche Regional" does not regularly store this data and had prepared it particularly for the purposes of this thesis. Therefore, it was not possible for the company to recollect several weeks worth of data in time. "Jobsuche Regional”s in-house developer was not sure were the error occurred, thus it can only be theorised, but the problem was likely a persistently saved encoding conversion error and a mixup between HTML and Full-Text files, as well as the server being unable to handle special characters used in the german language (e.g."Ä", "Ü", "Ö"). In any case, figure 7.4 illustrates how noisy the data was. There was a lot of additional information beyond the information needed.

In addition the broken format, likely due to the suspected encoding conversion error, parts of the data were corrupted. Figure 7.5 shows what most likely was supposed to be the term "Stadt: München”. However, spending several hours looking through the data with the informed knowledge of being a German native speaker suggests that the data was consistently, not randomly, corrupted.

All in all 26.1 GB of collected job data was available as text-files.
7.1.2 Discussion of Data

While it was unfortunate that the job data was in such a poor state, the completeness and size of the data was very satisfactory. Especially for this kind of thesis, it was very helpful to have access to such a rich amount of real-world data. All of German Wikipedia is about 5 GB in size, so having such a large corpus was very promising.

Also, the interconnectedness of the data was very good. Using the user data, we could easily build a profile for each user of all HTTP-requests they had made. These could be linked to Google Analytics and are additionally linked in the job data, so that all three files can be clearly linked.

Unfortunately, some of the data requested by us was not recorded, specifically the terms actions taken on site by the user and the list of items presented to the users. This was unfortunate for two reasons. First, we had no data available of what the users typed in the search box or what list of jobs was presented to them. This issue was worked around after we noticed that completed searches where included in the ”?s” variable of the URI in user data. Thus, at least complete search terms were available. Second, the list of items which was presented to users after a search was not accessible via work-arounds. This meant it would be impossible to build the large parts, such as the collaborative part of the recommender system. Therefore, we agreed to instead move the collection of this data back in time until deployment of the first prototype and collect it together with all data which was to be collected during deployment of the prototype.

7.1.3 Processing and Cleaning of Data

Google Analytics as well as the user data was of good quality. The important and arduous part was the cleaning and processing of the job data. Since the cause of the noise and corruption could not be traced for sure, the data was transformed using informed knowledge by the author in an iterative fashion. Transforming the code, re-reading it and transforming it several times in the hopes of making it continuously more usable.
The following code shows an small, exemplary fraction of necessary steps in the cleaning process, such as converting corrupted signs to their intended meaning, removing unnecessary white spaces and HTML artefacts:

```python
test = line.partition("<!--")
text = test[2].partition("-->") [0]
text = text.replace("Å", "ue")
text = text.replace("Å", "ae")
text = text.replace("Å", "oe")
text = text.replace("Å", "Ue")
text = text.replace("Å", "ß")
text = text.replace(">>", "")
text = text.replace("^ a", "")
text = text.replace("^ a", "")
text = text.replace("~A", "Ae")
text = text.replace(" ", " ")
text = text.replace(" ", " ")
```

It should be noted that this example does not show the complete extend of the problem, since a lot of the signs which had to be filtered were unavailable in PDF-Format. For a more informed impression please reference the exemplary file jobs.py accompanying this report.

However, after a long and careful cleaning process, the job data was transformed into a ”Job Object”, where for each job listing an attempt was made to filter the available information into semantically relevant attributes. Figure 7.6 shows one such job object. While certainly not perfect, and much corruption was still present, it was in a much better state than before. It was ultimately agreed that since new data would be gathered anyway after deployment of the first prototype, this was a workable solution and would fix itself if the project was to run continuously. Large parts of the data remained corrupted since it was impossible to reliably notice all corruptions in 26GB of text. However, this was not as important an issue as it may seem, because word embeddings do not know semantic meaning and if corruption is constant, it can learn from corrupted data to a certain extend as well. Nevertheless, it certainly is a problem which has to be fixed long-term. For the purpose of the following development, the focus was set on using the identified ”URL” attribute to link the job data to an URI, the ”Position” attribute was the title of the job listing and was isolated in order to experiment with
assigning different weights to the text and the headline and the ”Volltext”, which means ”full text”, was used for most of the language processing. All jobs which did not have all three of these attributes present were discarded.

Additionally, it was noticed that some of the job listings were actually in English, not German, which was information ”Jobsuche Regional” had not been aware of before. A quick test calculated that this was only the case in around 0.02 percent of all job listings. It was decided to ignore these listings for the purpose of this project, however they are an interesting concern for future work.

7.2 NLP Model

7.2.1 Creation of Job Data Corpus and Model

The NLP models were trained using gensim. In order to train a model, gensim needs a corpus, however, these are fairly simple: All that is needed is a file with one document per line, which in our case is a job listing.

For the job listings this was simply achieved by writing the full text to a corpus.txt. This was then processed by using a generator to work around memory issues. Standard gensim pre-processing library functions were applied. At the end, a model was trained from the corpus with a window-size of 10 and 150 vector dimensions:

```python
def read_input(input_file):
    with open(input_file) as input:
        for i, line in enumerate(input):
            yield gensim.utils.simple_preprocess(line)

        if (i % 10000 == 0):
            logging.info("read {0} listings".format(i))
```
yield gensim.utils.simple_preprocess(line)

abspath = os.path.dirname(os.path.abspath(__file__))
documents = list(read_input("corpus.txt"))
logging.info("Done reading data file")

# build vocabulary and train model
model = gensim.models.Word2Vec(
    documents,
    size=150,
    window=10,
    min_count=2,
    workers=10)
model.train(documents, total_examples=len(documents))
model.save("word2vec.model")

By default, gensim uses 5 epochs to iterate over the model, which makes it very accurate, however the process takes a long time - the longest time a model was trained in this thesis was 52 hours. In order to execute the code more quickly, iter=1 was added to the parameters of training the model.

### 7.2.2 Creation of German Wikipedia Corpus and Model

In the interest of comparing and choosing the most accurate corpus for our task, additionally to the job data we decided to train a model using all of German Wikipedia. Fortunately, Wikipedia makes up-to-date dumps available regularly and the gensim library is by default able to build a corpus from Wikipedia. Thus, building and training word embeddings for German Wikipedia, while extremely lengthy, was very straightforward using the following code:

wiki = wiki.load('wiki.corpus')
tfidf = tfidf.load('wiki.tfidf.model')

class MySentences(object):
    def __iter__(self):
        for text in wiki.get_texts():
            yield [word for word in text]
sentences = MySentences()
params = {'size': 300, 'window': 10, 'min_count': 40, 'workers': max(1, multiprocessing.cpu_count() - 1), 'sample': 1e-3,}
word2vec = Word2Vec(sentences, **params)
word2vec.save('wiki.word2vec.model')

7.3 Development of first Prototype

In accordance with the available data, we decided that the first prototype to be deployed on the company systems would be in accordance with Figure 3.2 and the first half, namely the search term part of Figure 3.3. With these in place, we could gather the needed data for the rest of the planned prototypes. We further agreed that this time the company would handle the responsibility of passing clean and processed data to the prototype. This was agreed upon for the main reason that this way, the company would not have to adjust the code if their data format would ever change. Please refer to the folder "jobsucheregional" for the ready prototype that was passed onto the company. The readme.txt is a protocol explaining the proper usage of the prototype. Unfortunately, however, it is in German.

7.3.1 Creation and Continuous update of model

As Figure 3.2 shows, we needed a a simple API which you could pass documents and then update and permeate the model. In a joined decision with the company, we decided that the said piece of software would be designed to be run once a day, during night, since training the model could be quite a time consuming process, which would be infeasible to do for each individual new job listing. Instead, the system got to train on all new listings for the day at once. New listings could still be recommended based on the trained model from the previous day, which should only make a marginal difference.

The following is in reference to buildword2vecmodel.py. Instead of the database proposed in figure 3.2, we decided to simply store the documents in a corpus file, as required by the gensim library.

The prototype gets passed a JSON-File:

logging.info("Updating corpus")
with open(sys.argv[1], 'r') as j:
    update_json = json.load(j)
Figure 7.7 shows the input format of the JSON. "Ueberschrift" means headline and "Text" means text. The JSON is supposed to have one object for each new listing of the day, "Ueberschrift" containing the header of the listing and "Text" the full text of the job listing.

```json
[
    {
        "Ueberschrift": "", "Text": ""
    },
    {
        "Ueberschrift": "", "Text": ""
    },
    ...
]
```

Figure 7.7: This figure shows the JSON input format for the continuous update of the prototype corpus.

The corpus file, which is continuously present, is then loaded and the content of the JSON is added line by line. Notice the parameter "a" for append, which means the corpus will grow larger each day:

```python
text_file = open("corpus.txt", "a")
for o in update_json:
    text_file.write(o["Ueberschrift"]+"\n")
    text_file.write(o["Text"]+"\n")
```

Finally, the input is read the same way as when creating our test-corpus by using a generator to preserve memory and using the standard gensim preprocessing. Finally, the model is trained and saved to the folder of the prototype.

### 7.3.2 Comparison of similarity

For retrieving recommendations, we decided on building a similarity score for documents passed to the system in relation to a search term. To achieve this, we decided that a selection of job listings would be passed to the prototype via a JSON. This was done for several reasons. First, while it was valuable that we could combine the data from
all of the job sites for machine learning purposes, the reverse was not true for document retrieval. Therefore, having the documents up for selection be present in JSON would grant control as to what content would be relevant for the search in question and give greater control to "Jobsuche Regionals" team of developers without a need to interact with our code. Also, this enabled to include current job listings into the search which where not yet integrated into the corpus.

Figure 7.8 shows the structure of the agreed upon JSON-format. Each job listing which should be considered for the recommender system would be single JSON-Object. The "Id" enables them to uniquely identify each listing. "Ueberschrift" and "Text" are again the header as well as the full text of the listing.

```json
[
    {
        "Id": "",
        "Ueberschrift": "",
        "Text": ""
    },
    {
        "Id": "",
        "Ueberschrift": "",
        "Text": ""
    },
    ...
]
```

**FIGURE 7.8**: This figure shows the JSON input format for getting a recommendation from the corpus.

In order to enable the company to be independent from our work in the future, extensive additional parameters were provided to enable use of the prototype according to "Jobsuche Regional"'s needs. In addition to the JSON, the search term, a header weight and a text weight would be passed to the prototype. Those weights enabled the company to decide in what ratio the header and the text of the job listing should be considered for similarity.

The code would save an Output.json to the folder, again to enable them to flexibly use the prototype according to their needs, which would simply contain the estimated similarity for each Id. Thud, the format was: 

```json
{ "Id1": Score1, "Id2": Score2, ... }
```

The following is in reference to similarityy.py. As stated the prototype loads the model and the JSON passed as an argument:
text_file = open("corpus.txt", "a")
for o in update_json:
    text_file.write(o["Ueberschrift"]+'\n')
    text_file.write(o["Text"]+'\n')

Afterwards, it goes over each entry in the JSON-File and builds a JSON for their similarities:

for o in input_json:
    get_text_similarity(o["Id"],o["Text"],o["Ueberschrift"], sys.argv[2], sys.argv[3], sys.argv[4])

And finally writes the results to an Output.json:

text_file = open("JSONOutput.json", "w")
text_file.write(json.dumps(similarity_dictionary))
text_file.close()

The code for the "get_text_similarity" function gets passed all the parameters, builds a similarity for each word in the text as well as the header in comparison to the search term and finally averages them according to the defined weights:

```python
def get_text_similarity(textid,input_text,input_header, keyword, weightText, weightHeader):
    words = input_text.split()
    wordsHeader = input_header.split()

    similarity = 0
    similarityHeader = 0
    offset = 0
    offsetHeader = 0

    for word in words:
        try:
            similarity += model.wv.similarity(w1= keyword, w2=word)
        except:
            offset += 1

    for word in wordsHeader:
```
try:
    similarityHeader += model.wv.similarity(w1= keyword, w2=word)

except:
    offsetHeader += 1

similarity = similarity / (len(words) - offset)

similarityHeader = similarityHeader / (len(wordsHeader) - offsetHeader)

finalSimilarity = (similarity*int(weightText) + similarityHeader*int(weightHeader)) / (int(weightText)+int(weightHeader))

similarity_dictionary[textid] = finalSimilarity
return(finalSimilarity)

7.4 Development of second Prototype

Right after finishing the first prototype, development on the second one began on a general basis of code structures, which we felt would be needed. These were mainly a way to represent a user as well as a way to compare complete documents to each other, both for collaborative filtering. Both of these were completed. The next step would have been to add Google Analytics Data, include the data gathered from the first prototype and build a deployable version of the second prototype. However, at this point we received information about the medical emergency within "Jobsuche Regional", invoked our prepared risks and continued with a different experiment. Nevertheless, the work done on the second prototype is included here for the sake of completeness. Please refer to the folder "jobsucheregional2" for the second prototype. Also notice that it is not a finished prototype.

7.4.1 Building a user model

In order to build a user model, we felt it was necessary to build a vector collection of all actions done by the user. This was achieved by iterating over all the user actions as shown in Figure 7.1. A dictionary was built where one key was created for each ip-address. Under this key, another dictionary was created for all the actions for this ip-address. This resulted in a complete dictionary for every user:
with open("/home/julian/Documents/workthesis/jobs_access/"+filename) as input:

    for line in input:

        useraction = line.split()

        if useraction[0] not in users:
            users[useraction[0]] = []

            users[useraction[0]].append(useraction)

        users.append(line.split())

    input.close()

This information could have later been easily used to use any form of feature extraction to build a vector to represent our users. For instance, a sum of all the terms they searched for could have been built when identifying the variable for search requests, identified by the variable ”?s”. Also, all URIs could have been used to link their behaviour to Google Analytics or view the job listings they accessed.

Similarity could have been assessed by comparing the distance between the vectors of two users, for instance using Cosine Similarity, Manhattan Distance or simply Euclidean Distance.

### 7.4.2 Comparing Documents

After having access to the documents which were presented to users and the information which they chose, we only needed a method of retrieving the most similar documents to a given document, instead of just a search phrase. This could have been achieved by using Latent Semantic Analysis Deerwester et al. [1990]. Here, again the higher dimensionality of a vector is reduced to a fixed number using orthogonal transformation to make complete documents easily comparable.

In reference to comparedocuments.py, we aimed to achieve this by first transforming our corpus into an LSI space and then transforming the corpus into this space and indexing it:

```python
lsi = models.LsiModel(corpus, id2word=dictionary, num_topics=2)

index = similarities.MatrixSimilarity(lsi[corpus])
```
Then, we can define any document, transform it to an word embedding and transform this to LSI space:

```python
doc = "This can be any old document."
vec_bow = dictionary.doc2bow(doc.lower().split())
vec_lsi = lsi[vec_bow]  # convert the document to LSI space
```

Finally, we just need to use Cosine Similarity to find the most similar document in LSI-Space and its Index:

```python
sims = sorted(enumerate(sims), key=lambda item: -item[1])
```

Unfortunately, at this point we had to discontinue the development of the second prototype.

### 7.5 Experiment

With the news that our prototype would not be deployed in time, we needed to get meaningful data for evaluation. We decided to quickly set up an experiment in line with recent research. There is considerable precedent in NLP research, specifically on corpora derived from online content, to extract a part of the corpus, annotate it and try to build a classifier using the annotated data as a test set Dinakar et al. [2011], Nobata et al. [2016], Schmidt and Wiegand [2017], Waseem et al. [2017].

#### 7.5.1 Experiment Design

We aimed to design our experiment to be in very close accordance with the designs of similar studies. First, we needed to get a sample of our corpus. In contrast to other studies, we were not interested in a specific subset of the corpus, but instead the whole of it. Thus, we could sample randomly. However, we were much more interested in the relative relevance of those items, not just in rigidly clustering them. So, for each line we randomly took three job listings from our corpus and matched one of the most popular search terms to them.

We named the listings in each line as "A", "B" and "C". The annotators were asked to rank them in accordance with the presented search term. For instance, if "B" was the most appropriate listing for the given search term, then "C" and then "A" they
would be asked to classify the line as "BCA". This way, we could evaluate our classifier both on terms if it could get the relative relevance correct and if it could find the most appropriate listing.

The annotators were chosen under the criteria of being German native speakers who had used a similar platform to seek employment before.

Unfortunately, we had neither the time nor the resources to get as big of a data set annotated as the cited studies. Especially since a job listing can be considerably longer than the Tweets or YouTube comments they used.

Since this would be a relatively long experiment as it was, we decided that subjecting the annotators to more than 30 instances would be unreasonable. Therefore, each participant annotated 90 job listings.

Appendix A shows the consent form presented to the participants, all their forms are available in the folder "participationforms".

### 7.5.2 Creation Experiment

In order to create the experiment, we created a .CSV file, so that the annotators could conveniently carry out the experiment using accessible software such as Excel or Libre Office calc. This was the code used to create the .CSV:

```python
array = random.sample(range(1, 43984), 120)
samples = []

with open("Output.txt", "r") as text:
    counter=1
    for line in text:
        if counter in array:
            samples.append(line.lower())

    print(counter)
    counter+=1

random.shuffle(samples)
```
with open('experiment.csv', 'w') as csvfile:
    filewriter = csv.writer(csvfile, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)
    filewriter.writerow(['Word', 'A', 'B', 'C', 'Similarity'])

    counter2 = 0

    for i in range(40):
        filewriter.writerow(['', samples[counter2], samples[counter2+1], samples[counter2+2], ''])
        counter2 += 3

Appendix B shows the experiment as it was presented to the participants.

### 7.5.3 Processing of Results

Appendix C shows the results obtained from the participants. In order to process them, in majorityvote.py, the classifiers for the different models were built.

The base-line was the algorithm the company is currently using as a classifier, which is fairly simple. It counts the number of occurrences of a char sequence, independent from the actual words:

```python
def key_word_similarity(input_text, keyword):
    words = input_text.split()

    similarity = words.count(keyword)

    return similarity
```

Additionally, both the model built from Wikipedia as well as the one from the job data of "Jobsuche Regional" were turned into a classifier, which simply averaged the 5 highest similarities within the presented document:

```python
def get_text_similarity(input_text, keyword):
    words = input_text.split()

    # uncomment when using wikipedia corpus
    # keyword = keyword.replace("ue","\u")
    # keyword = keyword.replace("st\u", "angestellter")
similarity = 0
offset = 0
wordsims = []

for word in words:
    try:
        wordsims.append(model.wv.similarity(w1= keyword, w2=word))
    except:
        offset += 1

try:
    similarity = sum(nlargest(5, wordsims)) / 5
except:
    similarity = 0
return(similarity)
Chapter 8

Evaluation

8.1 Natural Language Models

As discussed previously, Cosine Similarity can be used to compare vectors, to do semantic analysis by comparing the similarity between the products of vector arithmetic and to visualize findings using principal component analysis.

8.1.1 Exploration of Vectors

First, there are the general surroundings of word embeddings trained on the job data of "Jobsuche Regional" by finding the 10 most similar words. Figure 8.1 shows a visualization of the closest words for the term "Erzieher", which means "Child Care Worker". While the results are certainly not perfect, they show the tendencies hoped for: Similar jobs which have the same requirements and would be interesting to trained child care workers, such as such as "social pedagogue" or "social worker", are shown as the closet terms. Notice how these words share no topological similarities with the original word, meaning that their similarity is solely derived from semantic meaning. Further, qualifications required for this position, such as "staatlich" which means "officially recognized" are also shown. While this qualification might apply to other positions as well, this is certainly an interesting finding.

Figure 8.2 shows the same visualization for the Wikipedia corpus. It is somewhat similar, with certain words such as "pedagogue" shown as well. However, a lot of variations of the term for teacher show up. This is interesting, because while not expected, it makes sense within German culture. A teacher and a child care professional are expected to have a similar semantic meaning, however this does not show up at all for the job data. This is likely, because teachers are civil servants in Germany, meaning their jobs are
not advertised but appointed by the German government. As such, they would not be part of the corpus of a job hunting website. In general, it can be said that Wikipedia has a lot more synonyms among the closest neighbors, which is probably largely due to a larger vocabulary and indicative of job listings being less heterogeneous in structure than the rest of the German language.

However, the job data is somewhat inconsistent. For instance, figure 8.3 shows the embedding for the word "Entwickler", the German word for "Developer". Interestingly, it almost exclusively lists technologies or specializations which might be interesting for developers. But it does not list similar jobs, such as "Computer Scientist". Although, albeit not being in the top 10, it still shows very strong similarity.

In contrast, figure 8.4 shows the same embedding trained on Wikipedia. Here, the results are much more consistent and show synonyms, such as "Programmer", and related terms, such as "Software". Interestingly this also shows important companies in the business, such as Microsoft or LucasArts. However, it does not show common qualifications.

In general, both systems perform very well, and when comparing similarities the system always performed as the author, a German native, would expect - usually even by large margins. For instance, engineers are more similar to computer scientists than to teachers; surgeons more similar to doctors than to butchers and a mechanic is connected to being
Figure 8.2: PCA Visualization “Erzieher” Wikipedia.

Figure 8.3: PCA Visualization “Entwickler” Job Data.
certified. In fact, the only times the system did not provide firm answers, was when the author was not quite sure either - for instance, which is more similar to an engineer, a mechanic or an electrician?

8.1.2 Gaining Insight from Vectors

As discussed, beyond similarity, we can actually use vector arithmetic and compare the result to try and gain semantic meaning about our corpora. Figure 8.4 shows several professions subtracted from workplaces. The figure shows very similar, almost parallel vectors for the vast majority of those relationships. Chefs work in a restaurant, politicians at a town hall, teachers in a school et cetera. The model clearly learned the concept of a workplace from the data. That being said, it is not perfect and sometimes fails. At the bottom right a doctor and a hospital can be seen. They should most likely have the same relationship as the examples earlier, however, for unknown reasons they do not, within our corpus.
Figure 8.5: Figure showing that the model learned the concept of a profession and the according workplace.

Figure 8.6 is another interesting indicator that our model has actually learned meaning and gained semantic insight. It recognizes "Informatiker" (Computer Scientist) and "Entwickler" (Developer) as well as "Baecker" (Baker) and "Konditor" (Pastry Chef) as essentially doing the same job. However, it also recognizes "Architekt" (Architect) and "Ingenieur" (Engineer) as well as "Mechaniker" (Mechanic) and "Mechatroniker" (Electrical Technician) as being in related fields of works. This is especially interesting when building a recommender system for a job hunting platform.

8.1.3 Comparision of models

The Wikipedia and the Job Data model are fairly similar. However, they are distinct in a few aspects. Job Data is clearly more geared towards the job hunting language, whereas Wikipedia provides a more general depiction of the German language. This has advantages, for instance there are considerably fewer words out-of-vocabulary. However, it also adds a lot of noise in language that will never be witnessed in the context of a job hunting website.
8.1.4 Short-Comings

The biggest shortcoming of the Job Data corpus is the poor state of the data. A lot of the corpus is still corrupted and the corrupted words were trained and incorporated, meaning that it might think that certain words are out-of-vocabulary and will not recognize them, even though they are just corrupted. Also, the corpus is still fairly small.

The Wikipedia corpus is not semantically familiar enough with the language of job listings and misses connections, such as qualifications, which are linked to a profession.

8.2 Beyond Ethics

As discussed in chapter 5, we were cautious of the possibility that a corpus build from the German language might carry inherent sexism. Interestingly and unfortunately, this could in fact be observed clearly. Nearly all desirable traits as defined by Alicke [1985] were classified as being more similar to a "man" than to a "woman". Beyond that, more specific to our recommender system, certain tasks such as cleaning where learned to be more similar to a "woman" than to a "man".
Chapter 8. Evaluation

<table>
<thead>
<tr>
<th>Similarity &quot;Putzen&quot; (Cleaning)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Frau&quot; (Woman)</td>
<td>0.18311532</td>
</tr>
<tr>
<td>&quot;Mann&quot; (Man)</td>
<td>0.10662878</td>
</tr>
</tbody>
</table>

The corpus is even quite machismo in that, as it, for instance, classifies a stove as being more similar to a woman than to a man.

<table>
<thead>
<tr>
<th>Similarity &quot;Herd&quot; (Stove)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Frau&quot; (Woman)</td>
<td>0.15377194</td>
</tr>
<tr>
<td>&quot;Mann&quot; (Man)</td>
<td>0.105765164</td>
</tr>
</tbody>
</table>

This can lead to practical problems. For instance, this means that certain professions are more associated with a certain gender, which means the system could very well categorize users into gender-stereotypical clusters in a collaborative recommender system. This would be even worse with certain considerations for future work, such as building recommendations from reading a person’s CV.

Moreover, we can see the difference in only the generic versions of the same word. The male-generic word for "Developer" is four times more associated with the word "Promotion" than the female-generic version of the identical word.

<table>
<thead>
<tr>
<th>Similarity &quot;Beförderung&quot; (Promotion)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Entwickler&quot; (Male-Generic Developer)</td>
<td>0.2058322</td>
</tr>
<tr>
<td>&quot;Entwicklerin&quot; (Female-Generic Developer)</td>
<td>0.056173928</td>
</tr>
</tbody>
</table>

What is more, it seems that the corpus is not only sexist, but also intolerant in other ways, for instance, it associated black skin significantly more with prisons than white skin.

<table>
<thead>
<tr>
<th>Similarity &quot;Gefängnis&quot; (Prison)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Schwarze&quot; (Person with black skin color)</td>
<td>0.061881553</td>
</tr>
<tr>
<td>&quot;Weißer&quot; (Person with white skin color)</td>
<td>-0.0024472426</td>
</tr>
</tbody>
</table>

This is clearly a problem with no redeeming factors. From this evaluation it seems clear that future research into keeping NLP models neutral towards aspects such as gender or race and other minorities is very much needed.
8.3 Experiment

8.3.1 Interrater-Agreement

Interrater-Agreement was measured as discussed in section 2.2.6.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Agreement Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohens Kappa</td>
<td>0.798723933637878</td>
</tr>
<tr>
<td>Fleiss Kappa</td>
<td>0.7981046559538525</td>
</tr>
<tr>
<td>Krippendorfs Alpha</td>
<td>0.7992465904091509</td>
</tr>
<tr>
<td>Scotts Pi</td>
<td>0.7978992520897492</td>
</tr>
</tbody>
</table>

Overall, agreement was very strong, which is indicative of our test set being coherent and non-random.

8.3.2 Experiment Results

8.3.2.1 Original Company Algorithm

Figure 8.7 and figure 8.8 show the confusion matrices for the company algorithm. Figure 8.9 shows the accuracy report for the company algorithm. Clearly, the results are very poor. For 29 out of 30 test-cases, the algorithm simply made a guess because the keyword in question was not present at all. This is especially strange since the annotators claimed that most keywords were quite relevant to the presented listings. It only recognized the most relevant listing in 30 percent of cases, which is in line with this classifier mostly guessing for the task at hand.

8.3.2.2 Wikipedia Classifier

Figure 8.10 and figure 8.11 show the confusion matrices for the classifier built from the Wikipedia model. Figure 8.12 shows the accuracy report for the classifier built from the Wikipedia model. These results are much better. The accuracy is already 0.3 compared to the randomly expected 0.16. Also precision, recall and f1-score are relatively uniform, indicating that there are no certain specific problems within the model. This classifier already recognized the most relevant listing in 53.3 percent of cases.
Figure 8.7: Normalized confusion matrix for the classifier build from the keyword counting algorithm.

8.3.2.3 Job Data Classifier

Figure 8.13 and figure 8.14 show the confusion matrices for the classifier built from the Jobs Data model. Figure 8.15 shows the accuracy report for the classifier built from the Jobs Data model. This classifier performed even better and is achieving a range where it is clearly useful. Again, the performance is mostly uniform across the different performance measures. However, compared to the Wikipedia model, it was only marginally better in classifying the most relevant job listing, with only one more case, reaching 56.6 percent.
Figure 8.8: Unnormalized confusion matrix for the classifier build from the keyword counting algorithm.

Figure 8.9: Accuracy report for the classifier build from the keyword counting algorithm.
Figure 8.10: Normalized confusion matrix for the classifier build from the wikipedia model.
Figure 8.11: Unnormalized confusion matrix for the classifier build from the Wikipedia model.

<table>
<thead>
<tr>
<th>True label</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Accuracy: 0.3

<table>
<thead>
<tr>
<th>Report</th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.09</td>
<td>0.88</td>
<td>0.60</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>0.69</td>
<td>0.43</td>
<td>0.50</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>0.59</td>
<td>0.25</td>
<td>0.33</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>0.33</td>
<td>0.25</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>0.50</td>
<td>0.44</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.98</td>
<td>0.00</td>
<td>2</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.37</td>
<td>0.38</td>
<td>0.31</td>
<td>30</td>
</tr>
</tbody>
</table>

Figure 8.12: Accuracy report for the classifier build from the Wikipedia model.
Figure 8.13: Normalized confusion matrix for the classifier build from the jobs model.
Figure 8.14: Unnormalized confusion matrix for the classifier build from the jobs model.

Figure 8.15: Accuracy report for the classifier build from the Job Data model.
Chapter 9

Conclusion and Discussion

9.1 Conclusion

In this dissertation, two word embeddings were trained on corpora derived from the most recent German Wikipedia dump and on job data made available by "Jobsuche Regional".

Compared to the requirements specified in chapter 4 and the project plan formulated in chapter 6, not all aims of the thesis were achieved. While the project was very ambitious, the biggest hindrances proved to be a medical emergency at "Jobsuche Regional" and the shortcomings in the data provided by the company rendering any work on user actions as well as Google Analytics impossible. Reviewing the requirements formulated in chapter 4, outside these external restrictions all aims set initially were achieved: According to the project plan, in chapter 6’s milestones 1, 2 and 4 were achieved completely. Milestone 3 was only achieved partially due to the restrictions mentioned, instead, a complete experiment showing the use of the recommender system as a classifier was prepared and executed.

To our knowledge, this is the first work which used a German language model to build a recommender system for job listings. The models trained on both corpora were explored and shown to be useful in finding semantic similarities as well as learning concepts and gaining insights about the used corpora. Additionally, the classifiers built from these recommender systems were shown to perform significantly better than both the existing system and random distribution: It showed an accuracy of 43 percent in categorizing the exact order of a complex language problem in 6 possible classes. This is certainly an indicator for the usefulness of this prototype and of using an word embedding approach to build a recommender system for job listings. Therefore, it is an indicator for our hypothesis 1.3 to be true.

76
9.2 Limitations

Two main limitations stand out in the work of this thesis. Firstly, the corrupted corpus of "Jobsuche Regional". Ideally, the corpus needs to be completely rebuilt from scratch, especially since the degree of corruptness is not known. Any word could be randomly chosen out of vocabulary leading to an extremely inconsistent behavior of the prototype. Moreover, the corpus size is still quite small and it would be interesting to see how the performance of the model would change after running on "Jobsuche Regional" for a longer amount of time, continuously updating its corpus.

Secondly, while the results of the experiment where impressive, they could be quite misleading while significance tests are not routinely done for classification tasks. As none of Dinakar et al. [2011], Nobata et al. [2016], Schmidt and Wiegand [2017], Waseem et al. [2017] did so, it still stands to reason that a test set of 30 is too small to inspire much confidence in those results. Especially compared to labeled data sizes of 25,000, as in the case with Waseem et al. [2017]. However, it should be noted that all of the above used supervised learning techniques, compared to our unsupervised learning technique. It would still be very desirable to increase this test set and see if the same level of performance could be maintained.

9.3 Future Work

The obvious next step for future work is deploying the prototype on "Jobsuche Regional" and continuing from there as previously planned. A lot of work can still be done to improve the recommender system. The first step here would be to finish the partially completed second prototype for the collaborative system to include similar users. Afterwards, a third prototype could be built, trying to improve the recommender system further by adding business metrics from Google Analytics. For instance, average time spent at an URI could be an indicator of how interesting the job listing is to users. Finally, taking into account the monetization data available on Google Analytics, we could expand the recommender system to not only consider which jobs are most appropriate, but also which ones offer the most monetary value to the company. Although, at this point, the trade-off between those two goals, especially long term, should be carefully considered.

Furthermore, it was discovered that there are English job listings as well on "Jobsuche Regional". There is fascinating work done on building word embeddings from multilingual corpora Coulmance et al. [2016]. It would be very interesting to see whether the
same could be achieved for the corpus of "Jobsuche Regional". Searching for "Developer" and finding job listings including "Computer Scientists" would be very helpful to users.

As discussed in beyond ethics, among other aspects, the inherent trends towards sexism and racism in our model are worrisome. We are not aware of any research being conducted to solve these issues. However, the author strongly believes that this research would be worthwhile and should be endorsed. In the meantime, it seems imperative to be aware of this danger when using word embeddings to build recommender systems.
Appendix A

Participation Form Experiment
Study Title
Gaining insight from real world business data of a German company.

Invitation Paragraph
You are being invited to take part in a study concerning matching appropriateness of job-listing compared to a keyword. Before you decide, it is important for you to understand why this research is being done and what it will involve. Please take time to read the following information carefully. Please ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

Purpose of study
A lot of people use web-sites to seek employment these days. Better matching job-listing to the terms job-seekers search for is an important aspect in helping people find employment.

Why have you been chosen?
You have been chosen to participate in this study to help identify how real humans would match job-listing to a keyword.

Do you have to take part?
It is up to you to decide whether or not to take part. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part, you are still free to withdraw at any time, without giving a reason. A decision to withdraw at any time, or a decision not to take part, will not have any consequences.

What will happen to me if I take part?
You will receive a spread-sheet containing keywords, job listings and be asked to match them according to appropriateness.

Anonymity and Confidentiality
All personal information obtained from you during the study will be kept strictly confidential. Your contact details will be kept safe and will only be accessed by the researcher, Mr. Julian Kurz. Any information that is reported or published will be anonymised prior to dissemination, and thus will not contain information that would reveal your identity. You can ask for your data to be deleted at any time by emailing or contacting one of the people listed below.

What will happen to the findings of the research study?
The findings will be used to evaluate a mathematical model aimed to make recommendations to job seekers.

Right to Withdraw
You have the right to withdraw from the study at any time if you wish to do so.

Complaint Procedure
If you have any complaints about any issues regarding the study or any of the proceedings of the study, you can address informal complaints to Mr Julian Kurz.

Contact for Further Information

**Researcher Mr Julian Kurz**  
School of Engineering and Physical Science  
Heriot Watt University, Edinburgh, Scotland, UK EH14 4AS  
Email: jmk12@hw.ac.uk

Thank you for your interest in this study
CONSENT FORM
Title of Research: Gaining insight from real world business data of a german company.

Researchers: Julian Kurz

Please Initial Boxes

I confirm that I have read and understand the information sheet for the above study and I have had the opportunity to consider the information, to ask questions and have these answered satisfactorily.

I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, and without consequences.

I understand that any personal data (including direct quotes) used in the dissemination of findings will remain anonymous.

I agree to take part in this study.

__________________________  _____________________
Name of Participant                             Date                         Signature
Appendix B

Experiment as presented to participants
15.03.2018 jetzt bewerben! Herzlich willkommen bei pluss bildung & sozialwesen, wir bieten Ihnen in einrichtungen der fruhekinderbildung sowie bei traeger der sozialen arbeit den passend zugeschnittenen arbeitsplatz. werden sie ein teil von uns fur unseren regionalen partner suchen wir sie als erzieher/in in teilzeit oder in vollzeit. individuell waehrbar einsatzzonen und flexibel gestaltete arbeitszeiten sowie wohnortnahe und vor allem sichere arbeitsplatze garantiert planbare freizeit und sichere urlaubsplanung idealer berufseinstieg zu erprobung verschiedener kindertageseinrichtungen, schulen oder einrichtungen der einkaufsdienstleistung mit unterschiedlichen k

join the force! zur erweiterung unseres teams suchen wir ab sofort einen motivierten head of operations (m/w) - muenchen. seine aufgaben verantwortung fur den bereich operations teamaufbau fuer die bereiche backoffice, operations und it steuerung der it aktivitaeten, u.a. steuerung unseres it systems in abstimmtung mit unserem head of finance, ausbau bi, vorbereitung und ausfuhrung des crm robots inkl. integration unseres ausflugdienstes, aufbau und bewerbung eines neuen webshops fuer b2b und b2c koordination, ausbau und kontrolle der professionalisiert unsere logisticspartners in deutschland koordination und kontrolle der zentralisierung von 4 weiteren lagertarifstellen in europa einfuhrung und nachkontrolle eines qualitaetsmanagements im backoffice anhand von k

www.peersoftware.com peer software ist ein dynamisches und schnell wachsendes softwareunternehmen, das sich mit den herausforderungen des it services in verteilten umgebungen beschaeftigt. als amerikanisches unternehmen legt peer software seinen buero in muenchen, um und fnden einen starken fokus auf europa. unsere produktionsstandorte sind in afrika und in china. unsere teammitglieder arbeiten in berlin, muenchen, ulm und stuttgart. ein learn & earn-kultur als international taetiges unternehmen bieten wir vielfaltige moeglichkeiten, sich bei uns beruflich zu entfalten. zu verstaerken unseren teams in ismaining bei muenchen und ulm suchen wir zum naechstmöglichen termine einen engagierten post sales support

vertrieb

und hier finden sie sich genau wieder? Ihre aufgabe ware es im wesentlichen praxistreu, termingerecht und stets freundlich die umsetzung der Kundenwunschen und interner organisationsverordnungen zu erledigen. sie stellen durch ihre aktive unterstützung die fuer uns sehr wichtigen ablaufe fuer eine bestmögliche kundenzufriedenheit. (m/w). ich erwartet von Ihnen einen mitarbeiter, der sich in der vorstandsfunktion bewerbt. mein team bei uns ist very friendly and flexible, and we are looking for a machtvolle salesperson, who can help us to expand our business in europe.

krankepfleger

es reizt sie in einem der fuenf logistikunternehmen ihre kommunikationstalente miteinzubringen und auftrage an land zu ziehen? dann haben wir die passende stelle fuer sie als arbeitgeber (m/w) in einem der logistikunternehmen die sie interessieren. unsere unfallarzneimittel GmbH sucht einen erfahreneren dozent fuer die medizinische fuhrung eines unfallarzneimittelunternehmens mit samtlicher berechtigungen in der arbeitnehmerfuhrung. kein gesundheits- und sicherheitszertifikat erforderlich.

informatiker

alle reden ueber innovationen aber selten mit hands-on-experience. aber die technologie verstehen wir einstellung in den api-bereich und arbeiten mit aktuellen neuentwicklungen.

lehrer (m/w) fuer nachhilfe & privatunterricht online oder zu haufe - bundesweit superprof sucht aktuell in ganz deutschland privatlehrer / nachhilfelehrer (m/w) fuer einzelunterricht &
<table>
<thead>
<tr>
<th>Word</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
</table>
| Rahmen der personalvermittlung oder arbeitnehmeruberlassung mit interessanten und renommierten unternehmen aus ihrer region zusammen und sorgen so dafuer, dass sie einen job finden, der wirklich zu ihnen passt. werden sie ein teil von uns? fuer unseren regionalen partner suchen wir sie als front office manager (m/w) wir bieten ihnen: einen unbefristeten arbeitsplatz mit allen gesetzlichen und tariflichen sozialleistungen (inkl. urlaubs- und weihnachtsgeld) fachbezoogene, langfristige einsatze in der kaufmaennis behaeltst und zur prozessverbesserung beitraegt. mit zukunftweisendem change management madfust du auch selbst einen groafen schritt in dieser karriere. in einem team, das heute definiert, wie unsere welt morgen aussieht und nicht nur darauf redet: beschreibung starte als (junior) consultant (m/w) bei m3 management consulting in munich. bei dieser stelle handelt es sich um eine direktermittlung, der gesamte bewerbungsprozess wird von uns durchgefuehrt, wende dich deswegen bei fragen zum prosess bitte an deiner karrierepartner sina busch bei academic work: aufgaben du startest in zusammenarbeit mit einem senior consultant und wirst in folgenden bereichen zunachst unterstutzend tatig sein: in einem team arbeite gruppenkurse zu hause, online oder in der nahe des wohnorts. der unterricht kann sich an schuelerinnen &amp; schüler samtlicher altersklassen und nivellen richtet. deine vorteile: freie wahl der fachfer &amp; unterrichtsort: einzelunterricht, gruppenkurs, online-nachhilfe flexible unterrichtsplanung: du bestimmst zeit, ort und die kundin dieses unternichts attraktive &amp; klare vergütung: wir verlangen keine kommission haben wir dein interesse geweckt? | mechander ihre zukunftige arbeitstellte: wir suchen fuer unseren kunden, ein weltweit fuellrendender arbeiter fuer hochwertig veredelte ummantlungsfurniere und furnierannten, eine produktionshilfen (m/w): ihre aufgaben: sie arbeiten an furniermaschinen sie mussten fehlerhafte stuecke aus furnierstreifen rauschen die sie mussten das furniermaterial in die maschine einlegen: dort werden die furnier automatisch verbunden und auf groafen rollen aufgerollt sie mussten die teile auf die qualitat uberpriifen und sicherstellen, dass es keine fehler an der verbindungsschicht gibt: die tatigkeiten werden ubereinigend im stehen ausgefuehrt, kann sich aber auch mal hinten unter profi: sie fuessen ubere fingerfertigkeit und sind in der lage kleine und groe furnierarbeiten auszuftihren sie fuessen ubere handwerkliches geschick stehende tatigkeiten sind fuer sie ein kein unser angebot: attraktives einrichten und ruesten der maschinen (mit bauten und baulauren) qualitaetspriifungen und prozesskontrollen an flachbaugruppen, z.b. serienlaufkontrolle, druck-, test- und stueckueberpriifung durchfuehren von wartungsarbeiten gemafi wartungsplan optimierung der arbeitsablauf, z.b. optimales verhaelt zwischen ruest- und stueckueberpriifung zu maschineneinrichtungsgrad kontinuierliche weiterbildung die andritz gruppe ist einer der weltweit fuellrenden lieferanten von anlagen und serviceleistungen fuer wasseramp erwirtschaftlichen werkstoffe, fuer die zellstoff- und papierindustrie, die metal undindustrie sowie fuer die andere sideral special-industrien (fest-flussig-trennung, futtermittel und biomasse), der hauptsitz der gruppe, die weltweit rd. 25.700 mitarbeiter/innen beschäftigt, liegt in graz, â€œ sterreich. andritz verfuegt uber mehr als 250 produktionsstaetten sowie service- und vertriebsfirma in europa auf der ganzen welt, fuer den standort dueren sucht andritz kraftfahrer fuer folgende position eine/n mitarbeiter/in zum baldoengleischen eintritt: schichtfuehrer (m/w) und kraftfahrer gmbh. unsere kaufmaennern fur den standort dueren sucht andritz kraftfahrer fur die folgende position eine/n mitarbeiter/in zur baldoengleischen eintritt: schichtfuehrer (m/w) und kraftfahrer gmbh. unsere kaufmaennern fur den standort dueren sucht andritz kraftfahrer fuer die folgende position eine/n mitarbeiter/in zur baldoengleischen eintritt: schichtfuehrer (m/w) und kraftfahrer gmbh. unsere kaufmaennern fur den standort dueren sucht andritz kraftfahrer fuer die folgende position e | alterpfleger ihre aufgaben management verschiedener europaeische projekte gleichzeitig briefing des hauseigenen produktionsteams, sowie externer agenturpartner berichterstattung der projekterfolge an die europaeischen klienten anweisung der lokalen digitalen vermarktung abwicklung der rechnungsstellung bis hin zur zahlung mit dem finanzteam ihre aufgaben: zum tatigkeitsbereich gehoren alle aufgaben der rehabilitativen pflege wie z.b. unterstutzung bei der grundpflege einfuhrung der neuen aufgenommen patienten reinigung und desinfektion der stationseinrichtung auffuellen von wasche und pflegestoffen ihr profil: gerne erste berufserfahrung, idealweise in der neurologie oder onkologie, interessante arbeit bei der stadtwerk arbeiten erfolgreich mit onkologisch erkrankten patienten/innen bereitschaft zum 3: schicht system und zu wochendiensten; wir bieten Ihnen: ein interessen an und interessante verantwortungsvolle aufgabe in einem motivierten und engagierten team hohe qualitaetsstandards strukturierte einarbeitung ein umfangreiches angebot an optimierungs- und entwicklungs moglichkeiten, sowie umfangreiche angebote der internen und externen fort- und weiterbildung eine leistungsgerechte verguen | mechander 28.02.2018 jetzt bewerber! maschinen- und anlagenfuehrer (m/w) fuer heidelberg heiligenstadt gesucht! herzlich willkommen im geschaeftsbereich industrie der pluss unternehmensgruppe, dem specialisten im personalmanagement fuer die industri. mit der ubergebungs- und vermittlung von gesellen, fachleuten und kaufmaeinnern arbeiten sorgen wir in renommierten 25.04.2018 jetzt bewerber! herzlich willkommen in pluss care people, den specialisten im personalmanagement fuer medicin &amp; pflege. wir uberreichen Ihnen: unsere wuenchen entsprechend in der gesundheitswesen: werden sie ein teil von uns? wir suchen die als gesundheits- und krankpfleger (m/w) ubrigens auch fuessigkeit, betriebswirtschaftliche, unerfahrene oder erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeiter/in. die erfahrene arbeit
kundenbetrieben fuer mehr flexibilitaet und echte entlastung sowie passagenweise besetzung von offenen vakanten. viele namenhafte kundenunternehmen vertrauen seit jahren auf unsere qualitaetsdienstleistung. und der erfolg gibt uns recht als mittelstaendisches unternehmen sind wir nun seit uber 30 jahren mit unseren personalleistungen fuer menschen und unternehmen an uber 30 standorten am markt etabliert. wir sind uberauszeugt, dass die arbeitsewelt erfuellt drei.

steuer

30.04.2018 jetzt bewerber! herzlich willkommen beim pluss-team, spezialist im personalmanagement fuer das handwerk. wirubenlassen sie ihren wuenchen entsprechend in kleinstbetriebe genauso wie an industrialzulieferer, werden sie ein team von uns fuer unseren regionalen partner suchen wir sie als kaeselangebauer (m/w) wir bieten ihnen einen unbesichristenen arbeitsplatz mit allen gesetzlichen und tariflichen sozialleistungen (inkl. urlaubs- und weihnachtsgeld) abgezahlte auf grund des igz-tarifvergabes leisten ich profil umfangreiche praktische erfahrung in bezug auf verpflichtungen zu arbeitsverhandlungen, dauer und betriebliche erfahrung mit arbeitsverhandlungen von vorteil fur unsere angebot, attraktives arbeitsumfeld mit gutem perspektiven tarifliche entlohnung nach igz-igd-tarif plus branchenzuschlage betrieblich gefuehrte altersvorsorge fortlaufende gesundheitsfuersorge zusatzleistungen zulage, infolge besonderer eingliederung, die ihr volle aufmerksamkeit verlangen gratifikationen fuer betriebszugehoerigkeit von bis zu 2500EUR, exklusive preisnachlaesse bei reisen, moden, wohnen, freizeit und handyvertraege von zeit von zeitzu zeitzu freizeitplanung mitarbeiterangebote (z. b. adidas, zaland, sony, sky u)

mechatroniker

mit diesem wunsch sind sie bei asklepix gut aufgehoben. mit rund 150 gesundheitsleistungen in 14 bundeslaendern zaehlen wir zu den grossahen privaten klinikbetreibern in deutschland. der ken unsere unternehmensphilosophie es reicht uns nicht, wenn unsere patienten gesund werden wir wollen, dass sie gesund bleiben. wir verstecken uns als begleiter, der menschen ein leben lang zur seite steht. wir suchen ab sofort fuer unsere kuche (befrister) einen: mittarbeiter in der spuelkuche (w/m) fuer den arbeitgeber asklepix klinik sankt augustin am standort sankt augustin in vollzeit / teilzeit bewerben sie sich zur online-bewerbung ihren aufgaben: entfrachten und verpflegen von strukturbauenteilen (großraum- und kleinbauteile) nach zeichnung, putzplanen und kaufkarte bohren nach schadablen zeichnungsbleiben ihr profil umfangreiche praktische erfahrung in bezug auf verpflichtung zu arbeitsverhandlungen, dauer und betriebliche erfahrung mit arbeitsverhandlungen von vorteil fur unsere angebot, attraktives arbeitsumfeld mit gutem perspektiven tarifliche entlohnung nach igz-igd-tarif plus branchenzuschlage betrieblich gefuehrte altersvorsorge fortlaufende gesundheitsfuersorge zusatzleistungen zulage, infolge besonderer eingliederung, die ihr volle aufmerksamkeit verlangen gratifikationen fuer betriebszugehoerigkeit von bis zu 2500EUR, exklusive preisnachlaesse bei reisen, moden, wohnen, freizeit und handyvertraege von zeit von zeitzu zeitzu freizeitplanung mitarbeiterangebote (z. b. adidas, zaland, sony, sky u)

pfleger

mit diesem wunsch sind sie bei asklepix gut aufgehoben. mit rund 150 gesundheitsleistungen in 14 bundeslaendern zaehlen wir zu den grossahen privaten klinikbetreibern in deutschland. der ken unserer unternehmensphilosophie es reicht uns nicht, wenn unsere patienten gesund werden wir wollen, dass sie gesund bleiben. wir verstecken uns als begleiter, der menschen ein leben lang zur seite steht. wir suchen zum nachstmoglichen zeitpunkt einen that erarbeiten (w/m) fuer allgemein- und viszeralchirurgie fuer den arbeitgeber asklepix klinik nord am standort hamburg in vollzeit / teilzeit bewerben sie sich zur online-bewerbung endlich weit er frischen wind in schleswig holstein wertschaetzung, flexibilitaet, mitbestimmung? fremdwoerter fuer sie? dann bewerben sie sich bei uns als altapfleger (m/w) sie haben zu vieles zu leren, wir moechten sie wertschaetzen: wir machen besser moglich durch eine ubertarifliche bezahlung, einen unbesichristenen arbeitsvertrag, urlaubs- und weihnachtsgeld und mehr eine voll- oder teilzeitbeschaffung oder auf 450EUR stV (m/w) auf grund aller wahrnehmung wird fundiert werden uber aktuelle zuergemeinschaften, unternehmensunfaehigkeiten in den unterschiedlichsten landschaften. --- ich hoffe, dass sie sich bewerben und uns ihre laufende zusammenarbeit bei helfen!

im ballungsraum koeln/bonn/duesseldorf und setzt in vielerlei hinsicht maßstäbe; innovative, transparente, effiziente prozesse und hochmotivierte und extrem ambitionierte mitarbeiter/-innen. das unternehmen verfügt uber eine hervorragende reputation und einen sehr gut aufgestellten bestand und betreibt ein nachhaltiges portfoliomanagement mit dem ziel, den Wert des bestandes kontinuierlich auszubauen. zur verstarkung des unternehmensbereichs a&a2organisationa&a2 suchen wir fuer die zentrale in koeln zum nachrufzeitigen zeitpunkt den office manager (m/w) organisation/prozesse/ einheit (in voll- und teilzeit mind. 32 Stunden) in diesen verantwortlichen funktionen sind sie fuer die organisation, koordination und optimierung der administrativ materialumschlagsmaschinen, die leistungsfahigen maschinen haben ihren festen platz in unterschiedlichsten anwendungsbereichen vieler industriezweige und kommen weltweit zum einsatz. aufgaben thema des praktikums: sicherstellen der datenqualitat von 3d-cad-modellen. diese imponierende Behandlung von komplexen Themen der 3d-technik und ihre Erarbeitung in einem High-End-Umfeld werden zum einen die leistungsfaehigkeit und kreativitat der neuen Mitarbeiter/-innen in Frage stellen. zum anderen wird die leistungsfaehigkeit und begeisterung der gesamten Arbeitsgruppe gefordert. das gesamte Umfeld bietet eine einzigartige berufliche und berufliche Ausbildung. Aufgaben:

- Entwicklung und Verwaltung von Datenbanken und Datenmanagement-Systemen
- Erstellung von Dokumentationen und Berichten
- Planung und Steuerung von Projekten
- Optimierung von Prozessen und Arbeitsabläufen
- Koordination mit verschiedenen Abteilungen
- Teilnahme an Kundenauftritten und Meetings

Veranstaltungsort: a&a2organisationa&a2 koeln

kauflaemischen als auch im gewerblich-technischen sowie medizinischen Bereich. unsere vielfältigen Beschaffungsmodelle umfassen unter anderem die arbeitsnehmeruberlassung, die personalvermittlung und das interinmanagement. was wir suchen, sind folgende Qualifikationen: Wir suchen Mitarbeiter/-innen mit einem hohen Grad an Engagement und Aufgabenbereitschaft.

bueroaufmann

wir suchen siehen als kommissionierer m/e im zentrallager in leedebach kennen sie unseren kunden, einen fuhrenden energiemanager unterstützen und ihr wissen einbringen. ein angenommenes und zukunftsweisendes arbeitsumfeld mit weitreichenden perspektiven erwarten sie, bei diesem stellenangebot besteht eine hohe wahrscheinlichkeit fuer eine spätere festanstellung bei unserem auftraggeber. ihr neues aufgabengebiet: sichtkontrole der waren auf volltaedigkeit und funktionstuechtigkeit weiterleitung der rucklaufer an die entsprechenden stellen warenubernehmgskontrolle der rucklaufer waren aus den ausfalltagen uberprufung der begleitpapiere erfassung der rucklaufer im edv-system ihr profil: einschlagige erfahrungen im lagerbereich eine ausreichende Ausbildung im lagerbereich vorteilhaft gut

betriebskeller (m/e) fur unseren bosch car service johaneberg fuer neu- und erfahrenen. Ihre Aufgaben umfassen die organisatorische und technische Unterstützung des Teams, die Koordination von Arbeitsschutzmaßnahmen, die Durchfuhrung von Arbeitsaufgaben und die Abwicklung von Aufgaben im Bereich der Arbeitsdeckung.

mechatroniker

02.05.2018 jetzt bewerben! herzlich willkommen bei plus care people, den spezialisten im personalmanagement fuer medizin & pflege. wir uberlassen sie ihren wuenachen entsprechend in einrichtungen der alterpflege und in institutionen der krankenpflege. sie werden ein teil von uns fuer unseren regionalen partner suchen: wir bieten Ihnen ein individuell gestaltetes und flexibel gestaltetes arbeitszeitgebot sichere und freie arbeitszeitplanung kostenlose fortbildungen sowie weiterbildungs- und aufstiegsmöglichkeiten integration in ein tolles team mit weiteren in der betreuung vor ort und einsatzbegleitung unentbehrlichen arbeitsplatz

methatroniker 30.04.2018 jetzt bewerben! herzlich willkommen im plus-team, spezialist im personalmanagement fuer die industrie. wir uberlassen sie ihren wuenachen entsprechend in renommierte kundencarter. werden sie ein teil von uns fuer unseren regionalen partner suchen und sich in themen, wie maschinenbauzeichner (m/w) - industriezeugen oder maschinenbauer (m/w) - industriezeugen

30.04.2018 jetzt bewerben! herzlich willkommen beim plus-team, spezialist im personalmanagement fuer die industrie. wir uberlassen sie ihren wuenachen entsprechend in renommierte kundencarter. werden sie ein teil von uns fuer unseren regionalen partner suchen und sich in themen, wie maschinenbauzeichner (m/w) - industriezeugen oder maschinenbauer (m/w) - industriezeugen

lagerist

wir suchen fuer einen kunden im donauland helfen. streben nach einem derartigen arbeitsplatz, um so den Kundenbetrieb zu betreuen. Wir bieten Ihnen eine einzigartige Arbeitssituation mit den Vorteilen einer hoch motivierten Arbeitsgemeinschaft und einer ansprechenden Arbeitsumgebung.

wir als personaldiensleister die jobpulse-gruppe sucht nicht nur den
Technoskop (mikroskop) dokumentation der pruefung auf papier Ihr profil: sehr gutes sehvermoegen erfahrung in der qualitaetskontrolle von verteilt zuverlaessig, flexibel, serviceorientiert unsere leistungen fuer sie: ein sicherer arbeitsplatz mit unbefristetem arbeitsvertrag ubernahmeoption durch unseren kundenbetrieb intensive einarbeitu bei unserem kunden AÄÄA+ 150 starthereise / AÄÄA+ 500; empfehlau

Menschen in den mittelpunkt, sondern bietet daruber hinaus individuelle karrierechancen und eine hohe beratungsdienstleistung, werden sie teil unseres teams - starten sie im raum wolframshausen als produktionsmitarbeiter (m/w) smd/bauelemente/-aufgaben dotierung von smd-leiterplatten losten von kleinsteilen montage von smd-baugruppen feinarbeit

Informatiker wir suchen... fuer unsere kunden in f Firmen und umgebung fachkraftschwerpunkt/pfleger fuer die intensivpflege (m/w) werden sie teil unseres erfolgreichen akts... team unser anspruch bietet: einen unbeschriften arbeitsvertrag stundenlohn ab 22,00(AÄÄA, - urlaubs- und weihnachtsgeld f firmen-pkw zur vollen beruflichen und privaten nutzung (inkl. tankkarte) regelmassige fortbildungen ubernahme beim kunden moglich service mit herz: bei uns stehen sie im mittelpunkt unserer taeglichen arbeit ihre kompetenz: eine abgeschlossene ausbildung zum gesundheits- und Krankenpfleger (m/w/f) fachweiterbildung intensivpflege / anaesthesie oder mehrjahrige fachbezogene berufserfahrung zuverlaessigkeit, flexibilitaet und teamfaehigkeit sind fuer sie selbsts

Administration von windows- und linux-systemen den umgang mit aktueller hardware und netzwerktechnik modernster sicherheitstechniken rechenzentrumstechnik planung, installation und wartung von systemen und netzwerken unterstuetzung und schulung von kollege

Logistik

Wir als personaldienstleister verstehen uns auch als personaldienstleister die jobpulse gruppe sucht keine Mitarbeiter, sondern teammitglieder, die mit uns gemeinsam weiter wachsen moechten. diesen anspruch vertreten wir mittlerweile in 12 laendern der welt. die jobpulse gruppe stellt nicht nur den menschen in den mittelpunkt, sondern bietet daruber hinaus individuelle karrierechancen und eine hohe beratungsdienstleistung, werden sie teil unseres teams - starten sie im raum feldstadt als lagermitarbeiter/in mit ubernahmeoption ihre aufgaben kommissionieren von kundenauftraegen und fillauftraegen regelstelle und lagerauftragen

Ihre zukunftige arbeitstelle: unser namhafter kunde in eilen ist ein traditionsreiches mittelstaendisches Unternehmen, das seit mehr als 130 Jahren kunden weltweit mit produkten aus edelstahl und aluminium beliefert. in der maschinenfabrik wird erfolgreich auf dem gebiet der ventiltechnik gearbeitet, die im langjaehrigen Einsatz bewahrten Produkte werden bevorzugt in der chemischen, pharmazeutischen, getränke- und lebensmittelindustrie eingesetzt. ihre aufgaben: mitarbeit in der gasfahrt (pumpe) und verpacken von gro/B und kleingut. Ihre profil idealerweise haben Sie eine Ausbildung zum gasfahrtmechaniker (m/w) oder aehnliche ausbildung absolviert sie arbeiten gern im team, wie auch ebenso eigenstandig eine zuverlaessige und genaue arbeitweise ist Ihnen wichtig unser anspruch: attraktives Arbeitsumfeld mit guten perspektiven tarifliche entlohnung nach igd/dgb tar

Ingenieur


27.04.2018 jetzt bewerben! herzlich willkommen beim plus-team, spezialist im personalmanagement fuer die industrie; wir uberreichen ihnen ihren wuenschen entsprechend in renommierte kundenbetriebe. werden sie ein teil von uns/ fuer unseren renommierten kunden suchen wir sie als mechanikerin - kaethechnik (m/w) wir bieten Ihnen: einen unbeschriften arbeitsplatz mit allen gesetzlichen und tariflichen sozialleistungen (inkl. urlaubs- und weihnachtsgeld) fachbezogene, langfristige einsatze in der industrie beteiligung nach igd-tarifvertrag; ggf. uebertarifliche bezahlung sowie branchenspezialfirmae nach tv-bz hochwertige personliche schutzhausruestung und arbeitskleidung (z.b. engelbert strauss)

30.04.2018 jetzt bewerben! herzlich willkommen bei plus care people, den spezialisten im personalmanagement fuer medizin kampf; pflege, wir uberreichen ihnen ihren wuenschen entsprechend in einzelnen der altenpflege und in institutionen der Krankenpfleger. werden sie ein teil von uns/ fuer unseren regionalen partner suchen wir sie als kinder- anaesthesia technische/r assistent/in (ata) wir bieten Ihnen: individuell waehlbare und flexibel gestaltete arbeiterzeiten garantieren planbare freizeit und sichere urlaubsplanung kostenfreie fortbildungen sowie weiterbildungs- und aufstiegschancen integration in ein tolles team mit steter personlicher betreuung von vor ort und einseitbegeleitung unbeschriften arbeitsplatz mit fairer ver
<table>
<thead>
<tr>
<th>Word</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>informatiker</td>
<td>11.03.2018 jetzt bewerben!</td>
<td>herzlich willkommen im geschäftsbereich care people der pluss-</td>
<td>sie erfassen die anforderungen der kundenspezifikationen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unternehmensgruppe, dem spezialisten im personalmanagement</td>
<td>und die referentenleiter an die schaffkreise sie vermerken</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fuer die medicin &amp; pflege, mit der</td>
<td>und prufen die hardware gemaß fuer kundenanforderungen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uferberufserfassung und vermittlung von</td>
<td>und dokumentieren die entsprechenden testergebnisse</td>
</tr>
<tr>
<td></td>
<td></td>
<td>examiniertem personal</td>
<td>durch (gem. pdv) sie erfassen mogliche fehlerquellen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>und fachhelfern sorgen wir in einrichtungen der stationaeren</td>
<td>und besprechen diese mit den entwicklungsingenieuren</td>
</tr>
<tr>
<td></td>
<td></td>
<td>und ambulanten altenepple, in institutionen der</td>
<td>und referenten zur moglichen ueberuberarbeitung</td>
</tr>
<tr>
<td></td>
<td></td>
<td>krankenpflege sowie in behinderteneinrichtungen fuer mehr</td>
<td>im rahmen unseres kontinuierlichen wachstums suchen wir zum</td>
</tr>
<tr>
<td></td>
<td></td>
<td>flexibilitaet und echte entlastung sowie passgenaue</td>
<td>nachstmoglichen zeitpunkt sie als dipl.-ing. maschinenbau (m/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>besetzung von offenen vakancen. viele kundenumternehmen</td>
<td>w) entwicklungskonstrukteure ihre aufgaben umfassen:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vertrauen seit Jahren auf unsere qualitaetsdienstleistungen.</td>
<td>entwicklungskonstruktionen (maschinen, baueinheiten,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>und der erfolg gibt uns recht. als eines der marktfuehrenden</td>
<td>baugruppen), mitwirkung bei der entwicklung innovativer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unternehmen sind wir mittlerweile seit uber 30 Jahren mit</td>
<td>technologien, wir erwarten: dipl.-ing (th/rth) maschinenbau, ggf.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unseren dienstleistungen fuer menschen und unternehmen</td>
<td>maschinenbautechniker mit berufserfahrun, erfahrung im</td>
</tr>
<tr>
<td></td>
<td></td>
<td>an uber 30 standorten in der geschaeftswirtschaft</td>
<td>werkzeugmaschinenbau/ kenntnisse in der fertigungstechnik,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>etabliert. wir sind ubereinzogen, dass die ar</td>
<td>- cad-erfahrung • gute englischkenntnisse in word und schrift,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>boerokauffrau</td>
<td>• systematisches denken und handeln, strukturierte arbeitsweise.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ihre zukunftige arbeitstelle: fuer unseren kunden aus der</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>luft- und raumfahrtbranche am standort umf sind sie als</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>fachkraft (m/w) in bereich der optische revision taa.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ihre aufgaben: fehlererfassung und einleitung von nacharbeiten</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>uuberprufen auf vollstaendigkeit durch durchfluohnung von</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>waneineingangspruefung durch profil berufsausbildung als</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>zahntechniker/in, goldschmied (m/w), fehmecmechaniker/in</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>oder optiker/in bzw. vergleichbare qualifikation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>berufserfahrung in den bereichen feininstrument und erfahrung</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mit mikroskop sowie elektrofertigung sag-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>grundkenntnisse wertvorschenwert unser angebot:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>luftfahrtechnisches knowhow als langjahriger strategischer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>partner der airbus group spannende jobs bei interessanten</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>unternehmen wie airbus operations, premium aerotec sowie</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>der luftfahrtechnischen zulieferindustrie an</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>internationale reisetaetigkeit</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>als personaldienslouster verstehen uns auch als</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>personaldienslouster die jobpulse group sucht keine</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mitarbeiter, sondern teammitglieder, die mit uns gemeinsam</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>weiter wachsen mochten. diesen anspruch vertreten wir</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mittlerweile in 12 laendern der welt. die jobpulse group</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>stellt nicht nur den menschen in den mittelpunkt, sondern</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>bietet dazuer hinaus individuelle karrierechancen und eine</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hohe beratungsdienstleistung. werden sie teil unseres teams</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- starten sie im raum erzogen als einkaufer/in</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(nichtproduktivemittel) ihre aufgaben durchfliehen des</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>strategischen beschaffungsmanagements im bereich nicht-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>produktive gueter und facility ma</td>
<td></td>
</tr>
<tr>
<td>pfleger</td>
<td>Ihre zukunftige arbeitstelle: fuer unser erfolgreiches</td>
<td></td>
<td>zur unterstuetzung unseres teams suchen wir fuer</td>
</tr>
<tr>
<td></td>
<td>kundenunternehmen in ettingen suchen wir befristet bis</td>
<td></td>
<td>krankenhaeuser in pforzheim und stuttgart ab sofort</td>
</tr>
<tr>
<td></td>
<td>31.03.2019 eine/-kaufmann projektleiter/in mit folgendem</td>
<td></td>
<td>motivierte, erfahrene gesundheits- und krankenpfleger (m/w)</td>
</tr>
<tr>
<td></td>
<td>aufgabenbereichspunkt: ihre aufgaben: konzepterstellung</td>
<td></td>
<td>fuer die pneumologisch onkologische palliativstation die</td>
</tr>
<tr>
<td></td>
<td>versicherungen und umsetzung - aufbau risikomanagement</td>
<td></td>
<td>interesse, wissen und engagement mitbringen. *** arbeitszeit-</td>
</tr>
<tr>
<td></td>
<td>sicherstellung der liquiditaet und bereitstellung von</td>
<td></td>
<td>montage bis freitag im fruehdienst und</td>
</tr>
<tr>
<td></td>
<td>investitionsriemen - fremitmittelbeschaffung, koordination</td>
<td></td>
<td>spaziergang*** ihre aufgaben: ganzheitliche medizinisch-</td>
</tr>
<tr>
<td></td>
<td>von fordermittelantrag en unterstuetzung beim einkauf</td>
<td></td>
<td>pflegersische versorgung der patienten individuelle und</td>
</tr>
<tr>
<td></td>
<td>energie und bei der vergabe von neubauwerken - koordination</td>
<td></td>
<td>sorgfaltige betreuung der patienten enge zusammenarbeit</td>
</tr>
<tr>
<td></td>
<td>einkauf und claimmanagement - budget und</td>
<td></td>
<td>mit den kollegen aus dem pflege-team und den</td>
</tr>
<tr>
<td></td>
<td>investitionscontrolling - kalkulation / produktpreise ihr profil -</td>
<td></td>
<td>behandelnden aerzten fuellung der pflegedokumentation</td>
</tr>
<tr>
<td></td>
<td>betriebswirtschaftliche berufsausbildung mit mehrjahriger</td>
<td></td>
<td>Ihre qualifikation: abgeschlossene ausbildung</td>
</tr>
<tr>
<td></td>
<td>berufserfahrung in leitender funktion - erfahrung im aufbau</td>
<td></td>
<td>gesundheits- und krankenpfleger berufserfahrung,</td>
</tr>
<tr>
<td></td>
<td>neuer geschaeftsfelder losungsorientiertes arbeiten hohes</td>
<td></td>
<td>vorzugsweise in der onkologie eigeninitiative</td>
</tr>
<tr>
<td></td>
<td>verantwortungsbewusstsein und kommunikationsgeschick</td>
<td></td>
<td>patientenorientierte pflege teamfaehigkeit hohe soziale</td>
</tr>
<tr>
<td></td>
<td>unser angebot: attraktives arbeitsumfeld mit guten</td>
<td></td>
<td>kompetenz</td>
</tr>
<tr>
<td></td>
<td>perspektiven tarifliche entlohnung nach igd/dgb</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>buerokauffrau</td>
<td>einrichtungsleitung (m/w), gute bezahlung, kleiner treiger in</td>
<td>09.03.2018 jetzt bewerben! herzlich willkommen im</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>diese aufgaben bringen sie auf hochfrequent fehlerdiagnose und</td>
</tr>
</tbody>
</table>


bad honnau sie haben lust, eine einrichtung hinsichtlich image, marketing mit einem jungen, dynamischen und motivierten personaleam so rührig zu 'pimpen'! sie haben erste berufserfahrung als einrichtungsleitung gesamtem oder absolvieren gerade ein trainee-programm? ja? dann suchen wir genau so wie einrichtungsleitung (m/w)! natürlich ist die vermittlung durch uns in diese festanstellung ist fuer sie kostenfrei unser mandant ist ein kleiner privater traeger mit kurzen entscheidungswegern die einrichtung hat 80 betten, ach ja, naturlich erwarten sie auch noch folgende routineaufgaben: gesamtvorberatung fuer die einrichtung sicherstellung und gewahrlieferung der hohen pflegequalität personalaufnahme und entwicklung bewohnerakquisition und sicherung der kapazitaetsauslastung erfassung und einhaltung geschaeftsleitung office der pluss-unternehmensgruppe, dem spezialisten im personalmanagement fuer kaufmännische aufgabenerbereiche, mit der uberbezahlung und vermittlung von kaufleuten sorgen wir in der öffentlichen verwaltung, bei dienstleistungsunternehmen sowie in handwerks- und industriebetrieben fuer mehr flexibilität und echte entlastung sowie passgenauer besetzung von offenen vakancen. viele kundenunternehmen vertrauen seit Jahren auf unsere qualitätsperteinheiten, und der erfolg gibt uns recht als mittelstaedisches unternehmen sind wir nun seit uber 30 jahren mit unseren dienstleistungen fuer menschen und unternehmen an uber 30 standorten am markt etabliert, wir sind uberauszeugt, dass die arbeitswelt erfuellend sein kann und schaffen daher echte partnerschaften.

behebung an allen gaengigen kfz-marken qualitaetsicherung bei wartungs- / instandhaltungsarbeiten sowie verschiedenen reparaturen fachliche betreuung bzw. unterstützung der gewerblichen mitarbeiter und auszubildenden aktive mitarbeit in der werktatt mit diesem profil machen sie das rennen erfolgreich abgeschlossene kfz-meisterprüfung oder kfz-techniker-meisterausbildung idealerweise mehrjahrige berufserfahrung in vergleichbarer position hohe kommunikationsfaehigkeit mit ausgeprägter kunden- und servicesorientierung strukturierte und zufriedene arbeitseben gut evkennisse

schröner immobilien - makler / innen fuer die folgenden standorte: koeln, stuttgart, dusseldorf; sie arbeiten gerne selbstständig fuer sich und ihre erfolge bei freier zeitteilung und werden durch uns unterstützt - auch bei terminerung! - wir bieten ihnen lebenslang partnerschaft als immobilienmakler in unserem bundesweit aktiven unternehmen - ihre tätigkeit umfasst die vermittlung und den verkauf von immobilien als freier mitarbeiter bzw. mitarbeiterin - sie erhalten kostenlosen support an ausbildung, fortbildung, seminarbesuchen etc. - sie koennen sofort bzw. nach vereinbarung starten - sie haben keine finanzielle beteiligung bzw. vorauskosten an den kosten unseres unternehmens, wie internetkosten, werbung, call center etc. unser unternehmen ist kein franchise unternehmen! - sie erhalten staendig einkaufstermine durch unterstüttzung unseres call-centers - unser unternehmen ist premium kunde bei uber 1 000 internetportalen - provision bei erfolgreichem abschluss eines miet- oder kaufver...

zur verstaerkung unseres teams in der dienstleistungsorientierte heilbronn suchen wir ab sofort nachwuchs fleuhrungskraft (m/w) im bereich finanz- und rechnungswesen ihre zukunftigen aufgaben: als unterstützung fuer den teamleiter Â€Â€rechungspruefung ware Â€Â€sind sie neben eigenen operativen aufgaben in der rechnungspruefung fuer die koordination und organisation des reibungslosen aufflanges innerhalb des bereiches zuständig. dazu kommen auch aufgaben aus dem finanz- und rechnungswesen wie die debitoenbuchhaltung oder die verrechnung von handwerkerleistungen. sie berichten in dieser funktion dem leiter finanz- und rechnungswesen.

ihre aufgaben: endmontage hochwertiger laboreinrichtungen im haus montage der laboreinrichtungen vor ort bei externen kunden, verpacken und die mittile bei der verladung der ware ihr profil: abschluss als schröner/innen, zimmerer/innen oder eine vergleichbare ausbildung, mehrjahrige berufserfahrung, deutsch in wort und schrift unser angebot: attraktives arbeitsumfeld mit guten perspektiven tarifliche entlohnung nach igi/öbt tarif zzgl. branchenzuschlägen personelle einsatzbegleitung und qualifizierte beratung unser mitarbeiter-benefi-programm Â€Â€orizon pluspunkte Â€Â€ bis zu 30 jahre zehnertausd ihr Kontakt: orizon gmbh herr janewerem marlen marten straÂ€e 2 77656 offenau telefon: +49 781 65808 0 fax: +49 781 65808-88 e-mail: bewerbung.oberhein@orizon.de, Ihre bewerbung wir freuen uns auf Ihre aussagekraftige bewerbung

informatiker Ihre aufgaben: analysieren und abstimmen der projektanforderungen mit allen beteiligten entwicklungsleitern untersuchen des problemfeldes (problemstrukturierung) entwerfen und bewerten funktionaler losungssysteme und alternativen (bewertungsmatrix) erstellen von systemarchitektuern fuer adas (advanced driver assistance systems) sowie system- und konzeptgetalhauung uber den entwicklungs- und produktlebenszyklus hinweg ableiten von systemanforderungen an fahrzeug und softwarekomponenten koordinieren mit den vor- und nachgeschalteten entwicklungsleitern und unterstuetzen bei verifikationsmaßnahmen erstellen von automatisierungtools-erstellung eines konzepts und die entwickelung einer oberflaeche fuer ein elektronisches laborsystem

mit der pharmaindustry hattest du bisher keine bewerfungserfahrung? kein problem! du uberauszeugt durch dein kommunikations- und dein verstandes fuer zweckmässig undensive beziehungen? als einkaufer (m/w) auch in anderen branchen und industrien - konntest du bereits Erfahrungen sammeln und sucht nun nach einer neuen herausforderung? dann bieten wir dir einen einsteig in die internationale pharmabranche im kfz- und arzneimittelbereich in einem dynamischen und staedig wachsenden markt. deshalb suchen wir einen junior einkaufer (m/w) in der pharmabranche fuer das europaeische ausland, der das unternehmen ab sofort am standort trittau (bei heidelberg) unterstuetzt. bei dieser stelle handelt es sich um eine direktvermittlung, das bedeutet du bist direkt bei unserem kundenunternehmen angestellt. der gesamte bewerbungsprozess wird von

lagerist Ihre aufgaben entladen von lkw mit dem gabel stapler oder 30.04.2018 jetzt bewerben! herzlich willkommen beim pluss-
sie bringen ein gut fundiertes speditionelles know-how innerhalb
ameise warum kommissionieren ihr profit zuverlässigkeit stapelschein wunschenswert teamfaehigkeit und einfaches umfeld mit guten perspektiven tarifliche entloehung nach igf/dbg tarif zzgl. branchenzuweisungen personliche einsatzbegleitung und qualifizierte beratung untereiner der gewerbe markt unternehmens profil: zuverlässigkeit von staplerschein wuenschenswert teamfaehigkeit und einfaches umfeld mit guten perspektiven tarifliche entloehung nach igf/dbg tarif zzgl. branchenzuweisungen personliche

team, spezialist im personalmanagement fuer das handwerk, wir uberaufen sie Ihnen wuenschen entsprechend in ihrer betriebsnaturigen. wuenschen sie ein tiefe kunde aus uns werden sie uns als fluggerate einkaufschalter/m/w wir bieten Ihnen: einen unbefristeten arbeitsplatz mit allen gesetzlichen und tariflichen sozialleistungen (inkl. urlaubs- und weihnachtslohn) bezahlung auf basis des igf-tarifvertrages. ggf. uebertarifliche bezahlung hochwertige arbeitsteilung und schutzauflagen personliche betreuung einsatzbegleitung sichere uralteplanung Mitarbeiterangebote (z. B. adidas, zalando, sony,)

der charterverkehr mit tumors bewerben sie sich jetzt als disponent/in team- und komplettdurchlauf auf standort diese aufgaben begleiten Ihnen alltag: Disposition der teil- und komplettdurchlauf auf nationaler ebene unter beruecksichtigung der wirtschaftlichen auslastung des firmenweiten fuhrparks einlauf zusätzlicher kapazitäten auf dem spotmarkt inkl. der verhandlung von frachtpreisen stetig halten sie den kontakt zu kunden, subunternehmen und fahrern abgabe von tagespreisen bei kundenanfragen abrechnung Ihre qualifikationen: abgeschlossene ausbildung bis kaufmann fuer spedition und logistikdienstleistung oder vergleichbar einschaeglige erfahrung in der disposition von vorlauf durchsetzungsgestecke und verh

mechaniker

26.04.2018 jetzt bewerben! herzlich willkommen im geschäftsbereich care people der plus- unternehmensgruppe, dem spezialisten im personalmanagement fuer die medizin und pflege. wir suchen auch als gesundheits- und krankenpflegepersonal in dieser aufgabenbereiche: unterstützt die fachkraft (altepflege, gesundheits- und krankenpflege) bei der pflege und betreuung kranker; pflegebedarfsgerecht und behilflich bei menschen alle altersgruppen durchfuhrung der grundpflege und pflegephasen, nach aktuellen pflegewissenschaftlichen erkenntnissen, sozialen berufsunfallkontrolle durchfuhrung von deklarierbar behauptungspflege bei ärztlicher verordnung je nach qualifikation und delegation, unter berücksichtigung der spezielle

ihre hauptaufgaben in dieser funktion sind sie fuenf das zerlegen, ausgeben und dressieren von mind- und kaltfleisch und einen der Produktionsleitung in ihrem team werden sie als kunstleder- oder fleischfachmann/innen wir Ihnen: ein lebhaftes umfeld wohl zusätzliche zeichnen Sie sich durch eine zuverlässige und pflichtbewusste arbeitsweise aus. unser angebot bieten Ihnen eine herausforderung in einem zukunftssicheren und innovativen unternehmen. teamgeist, eigenverantwortung und eine leistungsfördernde entlohnung sind fuenf uns eine selbstverständlichkeit

zurueckbiegen einen umfangreichen angebot an qualitätssicherungsarbeiten in/aus dem operativen tagesgeschäft zum einen mit unserem service delivery administratorkin/innen (m/w) oder assistent (m/w) fuenf folgende interessante tätigkeiten. Ihre Aufgaben: unterstützung des service delivery manager im operativen tagesgeschäft einzige point of contact-management-schnittstelle zwischen dem kunden und unserem service delivery einlauf, dazu zählt auch die messung der kundenzufriedenheit (score card), die einleitungen qualitätsverbessernder maßnahmen sowie ggf. das management von eskalationen und managements der vertragsfleckige gefolgten service level sowie der operativen eingesetzten teams und partner vertragssachbearbeitung (anpassung, verlängerungen, end of services, usw.) erstellung und präsentation kundenspezifischer reports aufbau einer belastbaren kundenbeziehung durch kontinuierlich

mechaniker

suchen Sie abwicklung statt alttagroutine? teamsprit statt starrer hierarchien? dann sind Sie bei uns richtig! wir bieten Ihnen spannende Aufgaben bei einem fuchsenden mobildienstleister: ein professionelles, kollegiales arbeitsklima in einem hochmotivierten team. mehr noch: Sie durehen stark anfangen und sich starker weiterentwickeln bei diesem stellenangebot besteht die hohe wahrscheinlichkeit fuenf eine spätere festanstellung bei unserem auftraggeber: Ihr neues aufgabengebiet: kommissionieren mit handscannern nach lieferens bereitstellung von materialien bearbeitung des waren- und ausgangs transport und verpackungstätigkeiten unterstützende tätigkeiten eines ordnungsgemaßen tagesablaufs Ihr profil: interesse und freude an der arbeit im lager sind Sie körperlich fit, was Ihnen bieten: wohnortnahe einsatzige personliche betreuung soziale und rechtliche absicherung einsatz in/aus dem operativen tagesgeschäft zum einen mit unserem service delivery administratorkin/innen (m/w) oder assistent (m/w) fuenf folgende interessante tätigkeiten. IhreAufgaben: unterstützung des service delivery manager im operativen tagesgeschäft einzige point of contact-management-schnittstelle zwischen dem kunden und unserem service delivery einlauf, dazu zählt auch die messung der kundenzufriedenheit (score card), die einleitungen qualitätsverbessernder maßnahmen sowie ggf. das management von eskalationen und managements der vertragsfleckige gefolgten service level sowie der operativen eingesetzten teams und partner vertragssachbearbeitung (anpassung, verlängerungen, end of services, usw.) erstellung und präsentation kundenspezifischer reports aufbau einer belastbaren kundenbeziehung durch kontinuierlich

wenn Sie in Ihrem Bereich eine hohe verantwortung und koffer fuenf eine spezielle arbeit in einem kaiserlichen, gepflegten auftreten und freundlichkeit am arbeitsplatz sich die nennt skilful shutterstock und diese sicherheit handwerk fuenf ich kein problem, auf diese individuellen

ziemlich treffen sie in Ihrem Bereich eine hohe verantwortung und koffer fuenf eine spezielle arbeit in einem kaiserlichen, gepflegten auftreten und freundlichkeit am arbeitsplatz sich die nennt skilful shutterstock und diese sicherheit handwerk fuenf ich kein problem, auf diese individuellen

bueroaufmann

dafür bringt du mit: spezielle vorkenntnisse brauchst du als servicemitarbeiter nicht, denn mit deinem freundlichen, gepflegten auftreten und freundlichkeit am arbeitsplatz begeistern Sie unsere kunden dank ihrer flexibilität ist auch der schichtdienst fuenf ich kein problem, auf diese individuellen

ziemlich treffen sie in Ihrem Bereich eine hohe verantwortung und koffer fuenf eine spezielle arbeit in einem kaiserlichen, gepflegten auftreten und freundlichkeit am arbeitsplatz sich die nennt skilful shutterstock und diese sicherheit handwerk fuenf ich kein problem, auf diese individuellen

das institut der entreena impresario ag beraet existenz sehr engagiert. Sie müssen folgende test-scribes entwickeln von text-scrips zur auswertung des qualitätsstandards durchführen und auswerten von algorithmen- und software-tests designen und implementieren der benötigten software-algorithmus eigenverantwortliches organisieren und koordinieren von projektprospektspezifischen definieren der anforderungen an die testumgebung planen der notwendigen ressource teilnehmen an statusgesprächen in interdisziplinären projektteams einfacher abstimmen mit den entwicklungsabteilungen der geschäftsbereiche bzgl. testumgebung, tools und software abgleichen zwischen kunden- und internen spezifikationen
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
</table>
| bedürfnisse gehen wir dabei natürlich gerne ein. aufgabenbereich professionell zu begegnen, das zeichnet sie aus idealerweise eine abgeschlossene kaufmännische berufsausbildung z.B. kaufläche/mann im einzelhandel und/ oder mehrere jahre berufserfahrung im filialumfeld, insbesondere im kassenbereich berufspraxis in den bereichen behältnisabwicklung, kassenabschluss, warenrückgabe und umtausch sowie kenntnisse zu allgemeinen filialprozessen freude an dienstleistung, service und der naht zum kunden, gepaart mit einsatzbereitschaft, zuverlässigkeit, lernbereitschaft und teamgeist das erwartet sie bedienung einer modernen checkout-computerkasse und abwicklung von bargeldlosen zahlungsvorgängen mein eigener chef... und weiß wofür ich arbeite: für meinen eigenen erfolg und selbstbestimm... setzen auch sie in zukunft wie schon über 100 franchisepartner auf den marktführer zu unseren mandanten, seit 25 jahren der unangefochtene europäische marktführer seiner home-delivery-food-dienstleistung, suchen wir zur erschließung neuer franchisegebiete sie als führungskraft / chef / manager (m/w) home delivery als franchisepartner
Appendix C

Results experiment
<table>
<thead>
<tr>
<th>FIELD1</th>
<th>part1</th>
<th>part2</th>
<th>part3</th>
<th>part4</th>
<th>part5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>CBA</td>
<td>CBA</td>
<td>CBA</td>
<td>CBA</td>
<td>CBA</td>
</tr>
<tr>
<td>1</td>
<td>ABC</td>
<td>ABC</td>
<td>ABC</td>
<td>ABC</td>
<td>ABC</td>
</tr>
<tr>
<td>2</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>3</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
</tr>
<tr>
<td>4</td>
<td>BCA</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
</tr>
<tr>
<td>5</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>6</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
</tr>
<tr>
<td>7</td>
<td>CAB</td>
<td>CAB</td>
<td>CAB</td>
<td>CAB</td>
<td>CAB</td>
</tr>
<tr>
<td>8</td>
<td>CBA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>9</td>
<td>CAB</td>
<td>CAB</td>
<td>CBA</td>
<td>CBA</td>
<td>CBA</td>
</tr>
<tr>
<td>10</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
</tr>
<tr>
<td>11</td>
<td>CAB</td>
<td>CAB</td>
<td>CBA</td>
<td>CBA</td>
<td>CBA</td>
</tr>
<tr>
<td>12</td>
<td>ACB</td>
<td>ABC</td>
<td>ABC</td>
<td>ABC</td>
<td>ABC</td>
</tr>
<tr>
<td>13</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
</tr>
<tr>
<td>14</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
</tr>
<tr>
<td>15</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>16</td>
<td>CAB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
</tr>
<tr>
<td>17</td>
<td>CAB</td>
<td>CBA</td>
<td>CBA</td>
<td>CBA</td>
<td>CBA</td>
</tr>
<tr>
<td>18</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
</tr>
<tr>
<td>19</td>
<td>ABC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
</tr>
<tr>
<td>20</td>
<td>ABC</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>21</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
</tr>
<tr>
<td>22</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>23</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
</tr>
<tr>
<td>24</td>
<td>CAB</td>
<td>CAB</td>
<td>CAB</td>
<td>CAB</td>
<td>CAB</td>
</tr>
<tr>
<td>25</td>
<td>ABC</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>26</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
<td>ACB</td>
</tr>
<tr>
<td>27</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
</tr>
<tr>
<td>28</td>
<td>BAC</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
<td>BCA</td>
</tr>
<tr>
<td>29</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
<td>BAC</td>
</tr>
</tbody>
</table>
Appendix D

File Glossary

This Appendix acts as a glossary to navigate the files provided with this dissertation.
<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folder hwu-master-thesis</td>
<td>The LaTeX folder from which this document was generated.</td>
</tr>
<tr>
<td>importierte_jobs</td>
<td>Folder containing the raw job data.</td>
</tr>
<tr>
<td>jobs_access</td>
<td>Folder containing several thousand documents. Each document is the access log for users for one day for one particular job hunting website.</td>
</tr>
<tr>
<td>jobsucheregional</td>
<td>First prototype as presented to the company.</td>
</tr>
<tr>
<td>jobsucheregional2</td>
<td>Unfinished second prototype.</td>
</tr>
<tr>
<td>participationforms</td>
<td>Folder containing the template participation form, as well as all the forms signed by the participants.</td>
</tr>
<tr>
<td>tests</td>
<td>Containing tests for the first prototype, used during development.</td>
</tr>
<tr>
<td>buildword2vecmodel.py</td>
<td>Python script used to train the job data word embeddings.</td>
</tr>
<tr>
<td>createexperimentdata.py</td>
<td>python script used to create the .csv file for the experiment.</td>
</tr>
<tr>
<td>createwikitext.py</td>
<td>Create Wikipedia corpus and train word embeddings.</td>
</tr>
<tr>
<td>dewiki-20180720-pages-articles-multistream.xml.bz2</td>
<td>German Wikipedia dump</td>
</tr>
<tr>
<td>fleisskappa.py</td>
<td>Create Interrater-Agreement Statistics.</td>
</tr>
<tr>
<td>germanenglish.py</td>
<td>Estimate percentage of english job listings.</td>
</tr>
<tr>
<td>input.json</td>
<td>JSON File simulating possible input for prototype for testing purposes.</td>
</tr>
<tr>
<td>jobs.py</td>
<td>Exemplary cleaning and processing script.</td>
</tr>
<tr>
<td>JSONOutput.json</td>
<td>JSON output written during development and testing of prototype.</td>
</tr>
<tr>
<td>File</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>majorityvote.py</td>
<td>Python Script for measuring and visualizing the performance of the different classifiers on our experiment.</td>
</tr>
<tr>
<td>originalexperiment.csv</td>
<td>CSV of original experiment</td>
</tr>
<tr>
<td>Output.txt</td>
<td>Corpus Job Data</td>
</tr>
<tr>
<td>removeexcelheader.py</td>
<td>Used to convert excel data from participants back to csv.</td>
</tr>
<tr>
<td>resultsparticipants</td>
<td>Results experiment</td>
</tr>
<tr>
<td>similarity.py</td>
<td>Python Script used for experimenting on creating similarities using our word embeddings.</td>
</tr>
<tr>
<td>test.py</td>
<td>Python script which was used for quick experimentation to understand certain library functions etc.</td>
</tr>
<tr>
<td>updated_experimentresults.csv</td>
<td>Cleaned experimentresults.</td>
</tr>
<tr>
<td>users.py</td>
<td>Experiment on creating user representation.</td>
</tr>
<tr>
<td>visualizeCorpus.py</td>
<td>Python Script for visualizing experiments done on word embeddings.</td>
</tr>
<tr>
<td>models</td>
<td>Folder containing all the various models trained during experimentation.</td>
</tr>
</tbody>
</table>
Bibliography


BCS (2015). Bcs, the chartered institute for it trustee board regulations - schedule 3 code of conduct for bcs members.


