A Hybrid On-line Topic Groups Mining Platform

Cheng-Lin Yang
Yun-Heh Chen-Burger

THE UNIVERSITY
of EDINBURGH

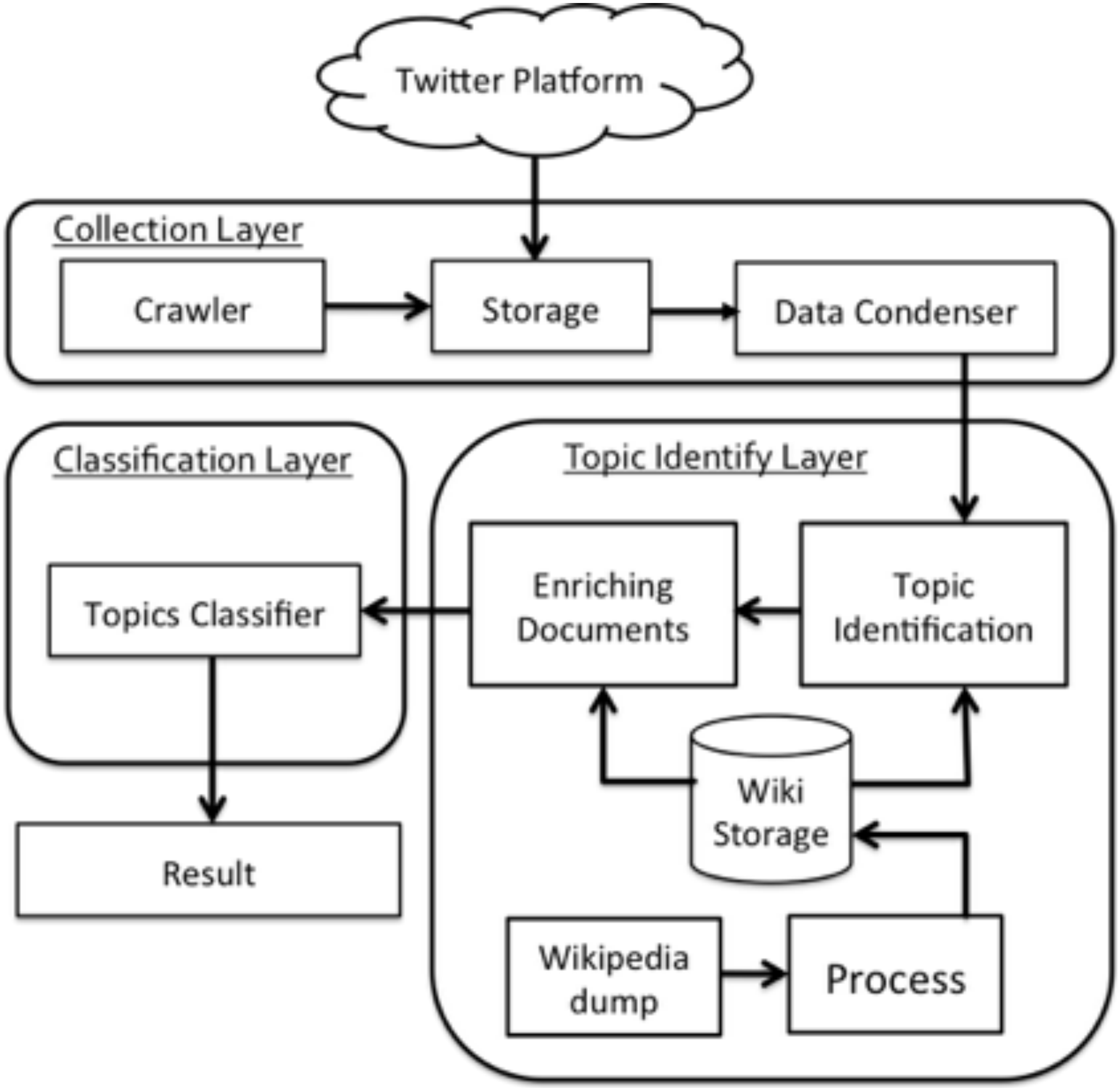
Target

- Given a large set of tweets, identify all possible topics of each tweet and cluster tweets with similar topics into communities.

Problems we face

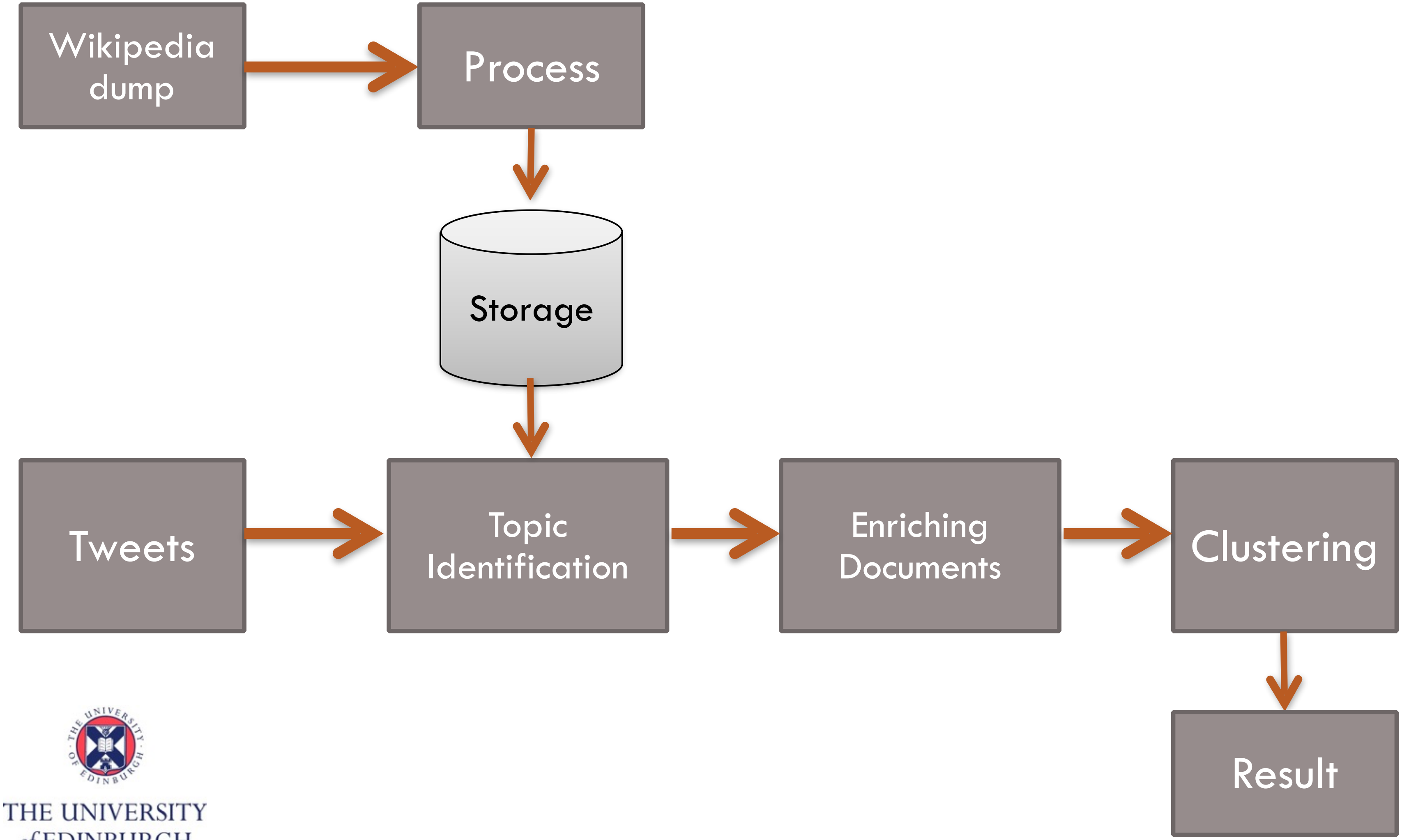
- Unstructured Data
  - Big data
  - Multiple users conversations
  - Uncontrolled topic threads
    - Up-to-date topic
  - Short content with little reference or information
  - Noise
    - emoticons: Orz / :) / :D
    - Internet slang: LOL / BRB
    - Meaningless strings: !@#%!!

Proposed Framework



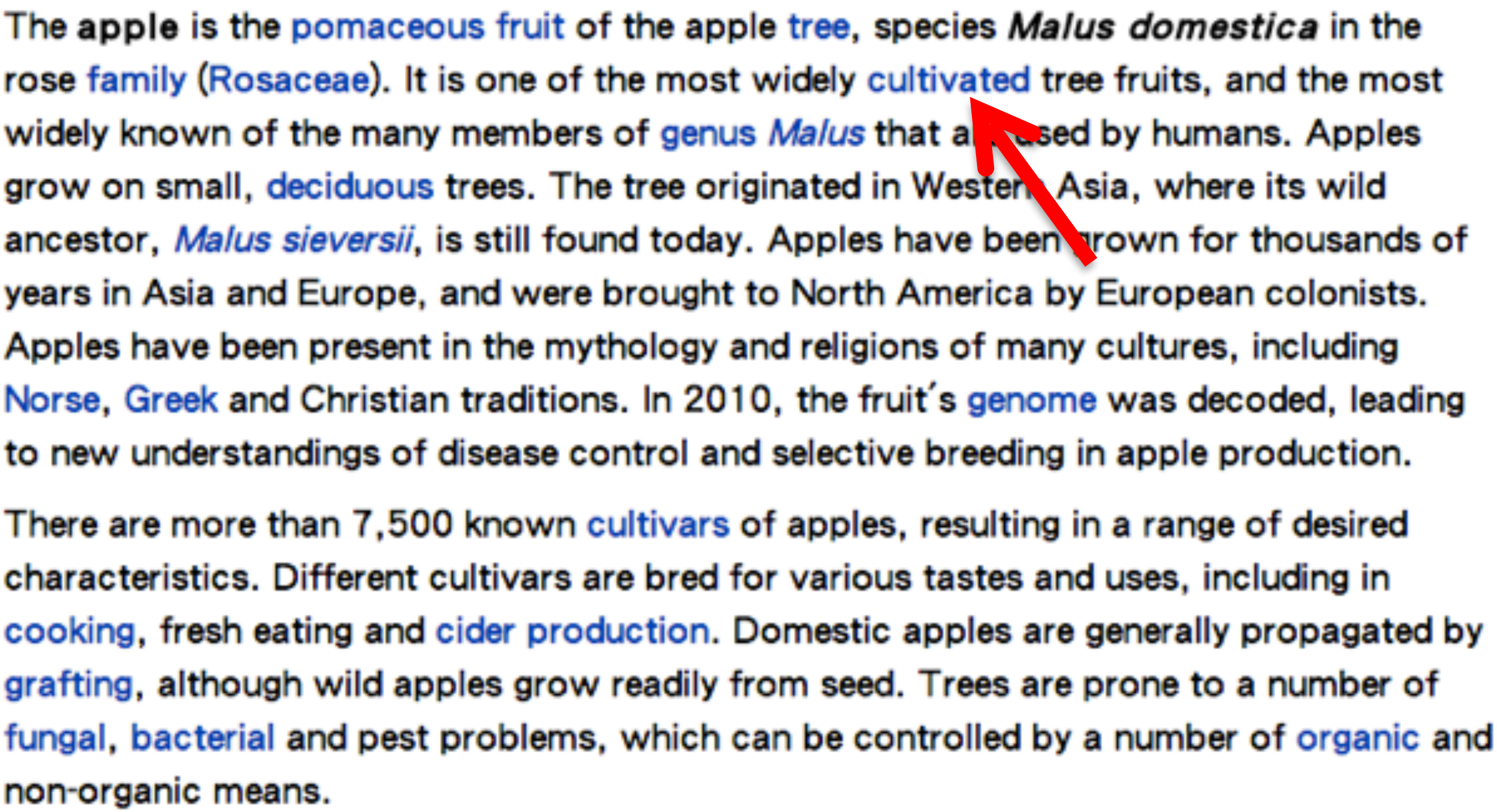
THE UNIVERSITY *of* EDINBURGH

Proposed Method Overview



Anchor Identification

- What is Anchor in Wikipedia



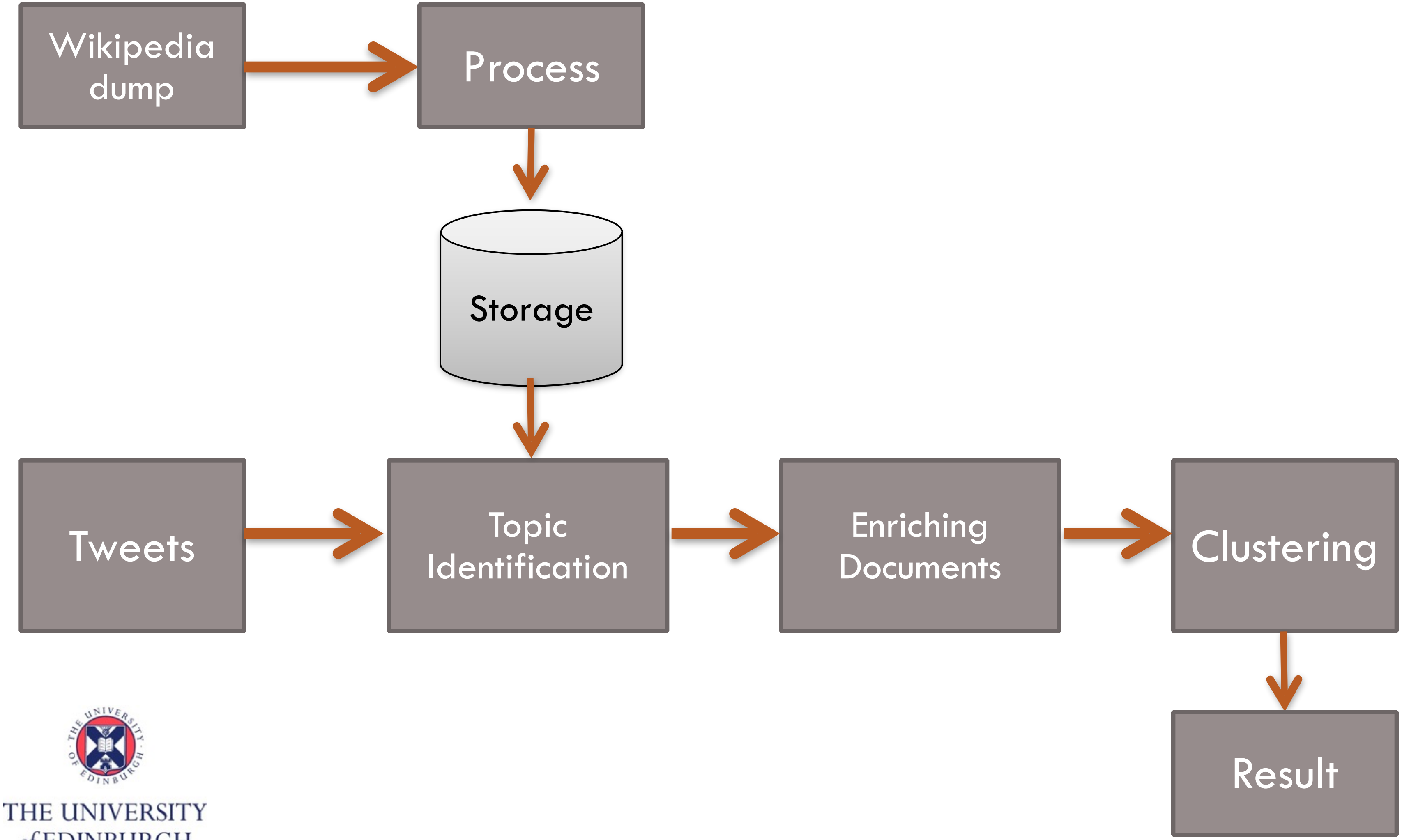
The **apple** is the pomaceous fruit of the apple tree, species ***Malus domestica*** in the rose family (Rosaceae). It is one of the most widely cultivated tree fruits, and the most widely known of the many members of genus *Malus* that are used by humans. Apples grow on small, deciduous trees. The tree originated in Western Asia, where its wild ancestor, *Malus sieversii*, is still found today. Apples have been grown for thousands of years in Asia and Europe, and were brought to North America by European colonists. Apples have been present in the mythology and religions of many cultures, including Norse, Greek and Christian traditions. In 2010, the fruit's genome was decoded, leading to new understandings of disease control and selective breeding in apple production.

There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. Different cultivars are bred for various tastes and uses, including in cooking, fresh eating and cider production. Domestic apples are generally propagated by grafting, although wild apples grow readily from seed. Trees are prone to a number of fungal, bacterial and pest problems, which can be controlled by a number of organic and non-organic means.

Anchor Identification

- Why Anchor is useful?
  - We define that an Anchor is a topic in Wikipedia
  - It is defined by authors therefore is more trustworthy

Proposed Method Overview



THE UNIVERSITY *of* EDINBURGH

Topic lookup

- Divide the input tweet by n-gram where n=1 ~ 6
- Eg: Steve Jobs is CEO of Apple
  - Steve, Steve Jobs, Steve Jobs is, Steve Jobs is CEO, Steve Jobs is CEO of Apple
  - Jobs, Jobs is, Jobs is CEO, Jobs is CEO of, Jobs is CEO of apple
  - Is, is CEO, is CEO of, is CEO of Apple
  - CEO, CEO of, CEO of Apple
  - of, of Apple
  - Apple

Topic lookup

- Look up all divided term in the anchor dictionary
  - Keep all matched anchors as candidates:
    - Steve, Steve Jobs, CEO, Apple
  - Remove the anchor which is the substring of the candidate anchor
    - Steve Jobs, CEO, Apple
- Ambiguous anchor issue:
  - Apple = apple tree

    apple computers

    apple records

    …..

Disambiguation

- Voting for the most possible topic which is the most related to the given anchor
  - Using Google distance to calculate the relatedness between all ambiguous topics and given anchor
  - Calculate total score of each anchor
  - Remove topic with lower score by threshold

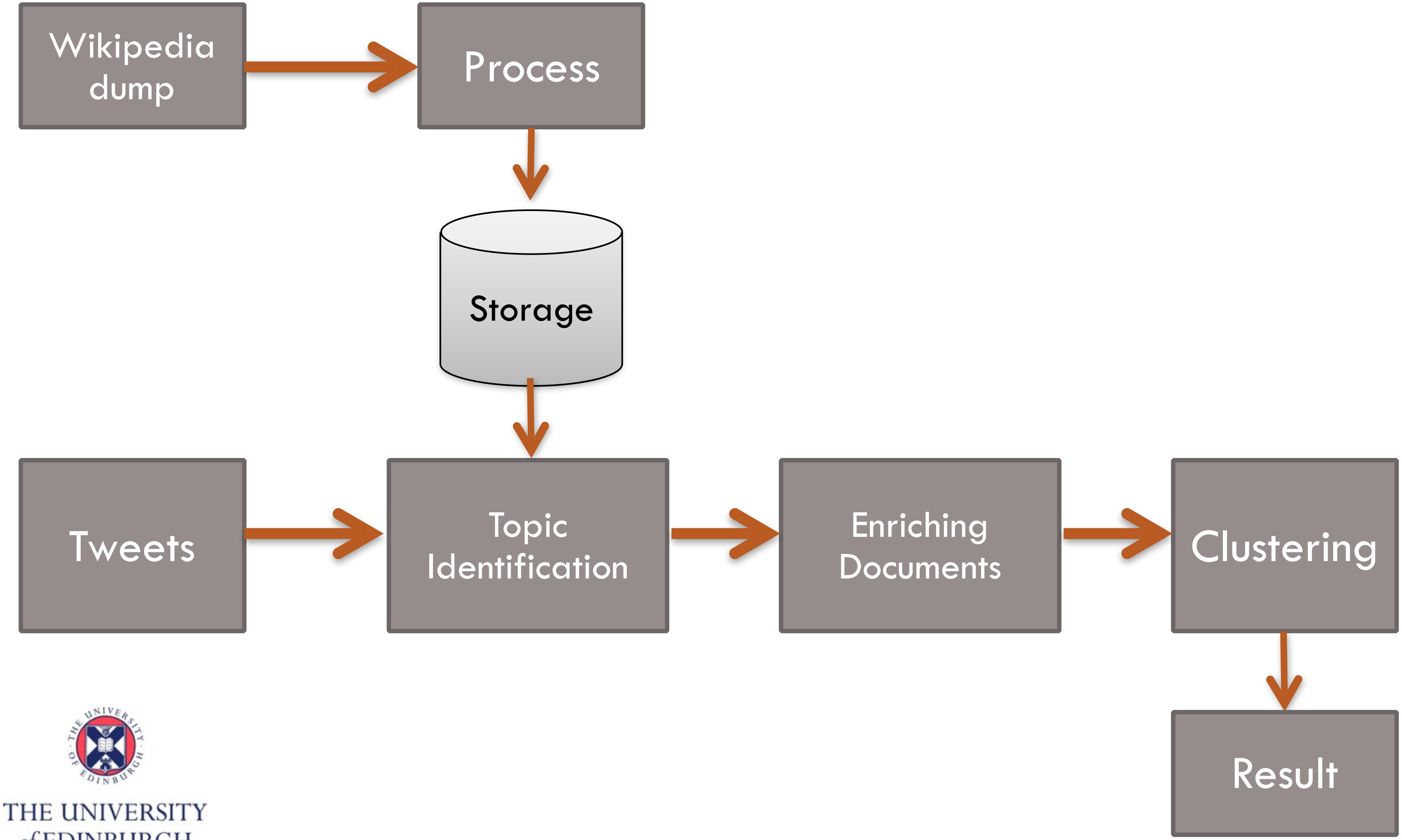
    Apple = {Apple inc., Apple Computer}
- Assign the highest commonness topic to given anchor
  - Apple = Apple Computer

THE UNIVERSITY *of* EDINBURGH

Topic filtering

- Result of disambiguation
  - Steve Jobs={Steve Jobs}
  - CEO = {CEO}
  - Apple = {Apple inc.}
- Finally, check the coherence between selected anchors

Proposed Method Overview



THE UNIVERSITY *of* EDINBURGH

Document Enrichment

- Applying TF-IDF on short text documents such as tweets is usually not able to identify the important terms. Eg:

"Watching on Youtube is easier and faster"

TF: {watch: 1, youtube: 1, easier: 1, faster: 1}

IDF: {watch: 0.35, youtube: 0.47, easier: 0.56, faster: 0.57}

THE UNIVERSITY of EDINBURGH

Document Enrichment – Method 1

"Watching on Youtube is easier and faster"

- Identified topic: youtube. Add it to the tweet

"Watching on Youtube is easier and faster Youtube"

TF: {watch: 1, youtube: 2, easier: 1, faster: 1}

IDF: {watch: 0.26, youtube: 0.73, easier: 0.44, faster: 0.44}

Document Enrichment – Method 2

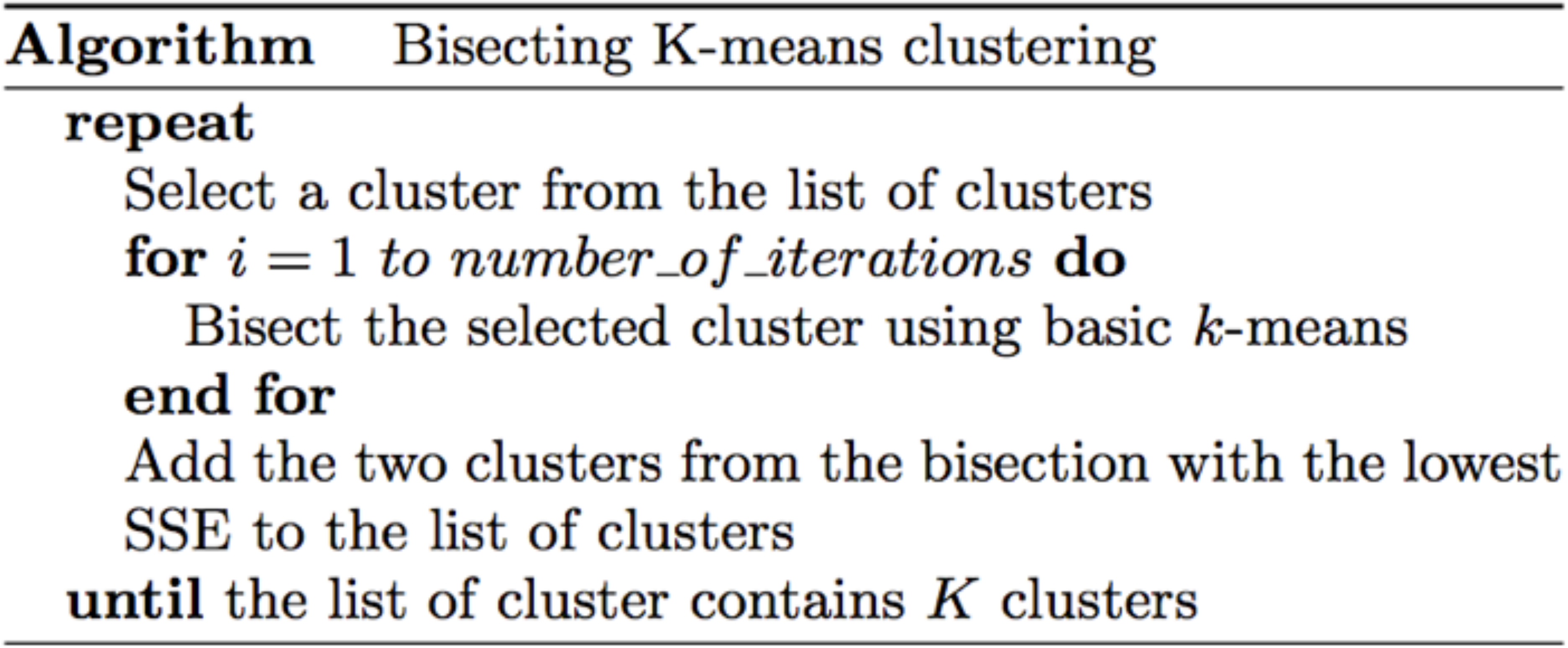
- However, Method 1 ignores that two tweets might have semantic related topics.
  - "Flickr is awesome!" => topic: Flickr

    "Just in love with Shutterfly" => topic: Shutterfly
  - Flickr and Shutterfly are both in "Photo Sharing" category in Wikipedia
- Therefore, adding Wikipedia category to both tweet to increase the cosine similarity

THE UNIVERSITY
of EDINBURGH

Clustering tweets

- □ Using Bisecting K-means



Algorithm Bisecting K-means clustering

```
repeat
  Select a cluster from the list of clusters
  for $i = 1$ to $number\_of\_iterations$ do
    Bisect the selected cluster using basic $k$-means
  end for
  Add the two clusters from the bisection with the lowest
  SSE to the list of clusters
until the list of cluster contains $K$ clusters
```

Evaluating the result

- Three testing cases
  - Baselines
  - Adding Wikipedia topics
  - Adding Wikipedia topics and categories
- Datasets
  - Ground (golden) truth - 20 topic groups
    - 20 tweets for each group
  - Testing sets
    - ~ 1.1 million tweets (English only)

THE UNIVERSITY
of EDINBURGH

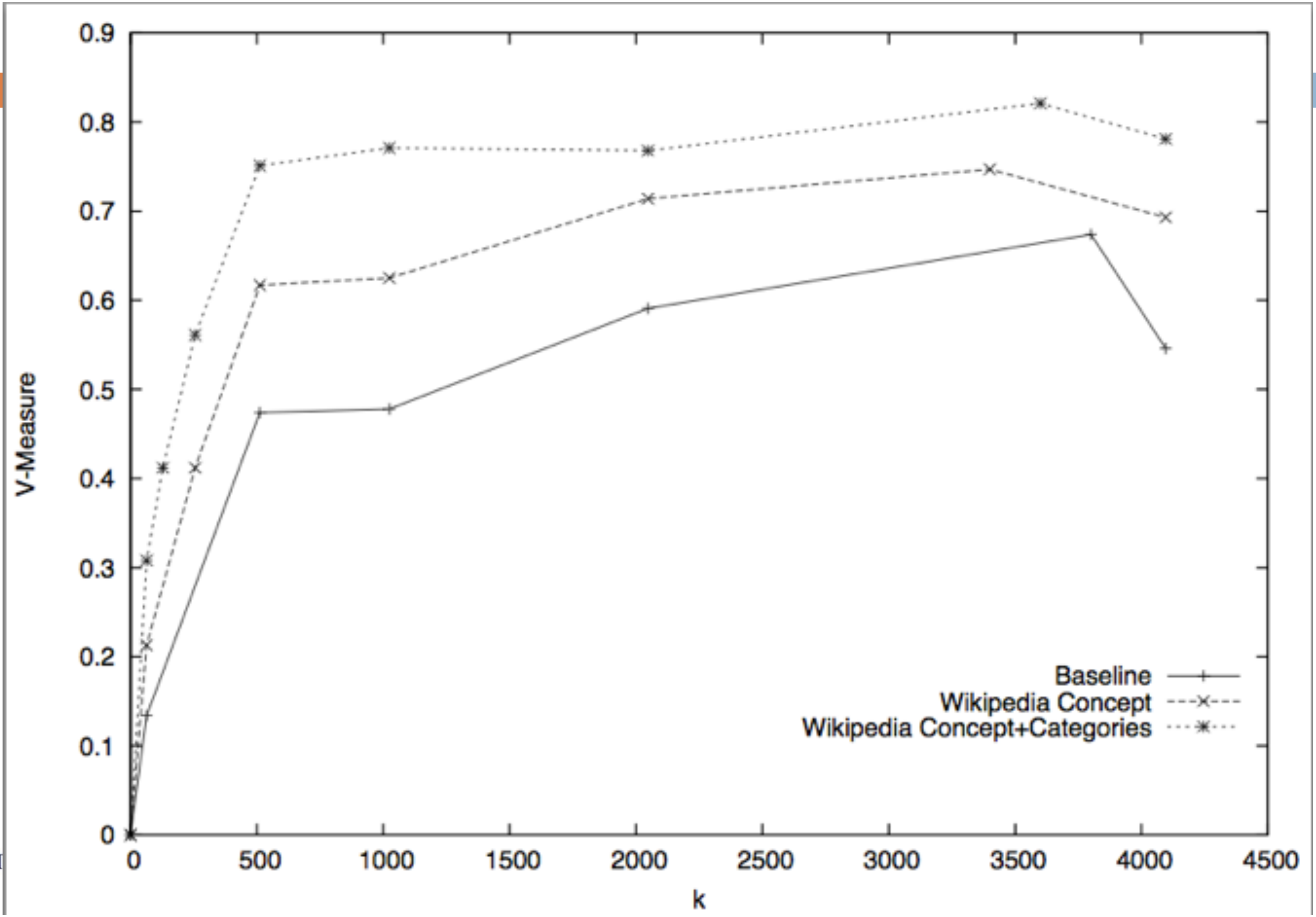
Evaluating the results

- Using V-measure to evaluate the generated clusters
  - V-measure is a evaluation functions which considers both homogeneity and completeness

$$V_\beta = \frac{(1+\beta) * h * c}{(\beta + h) + c}$$

  - homogeneity: each cluster contains only members of a single class
  - completeness: all members of a given class are assigned to the same cluster

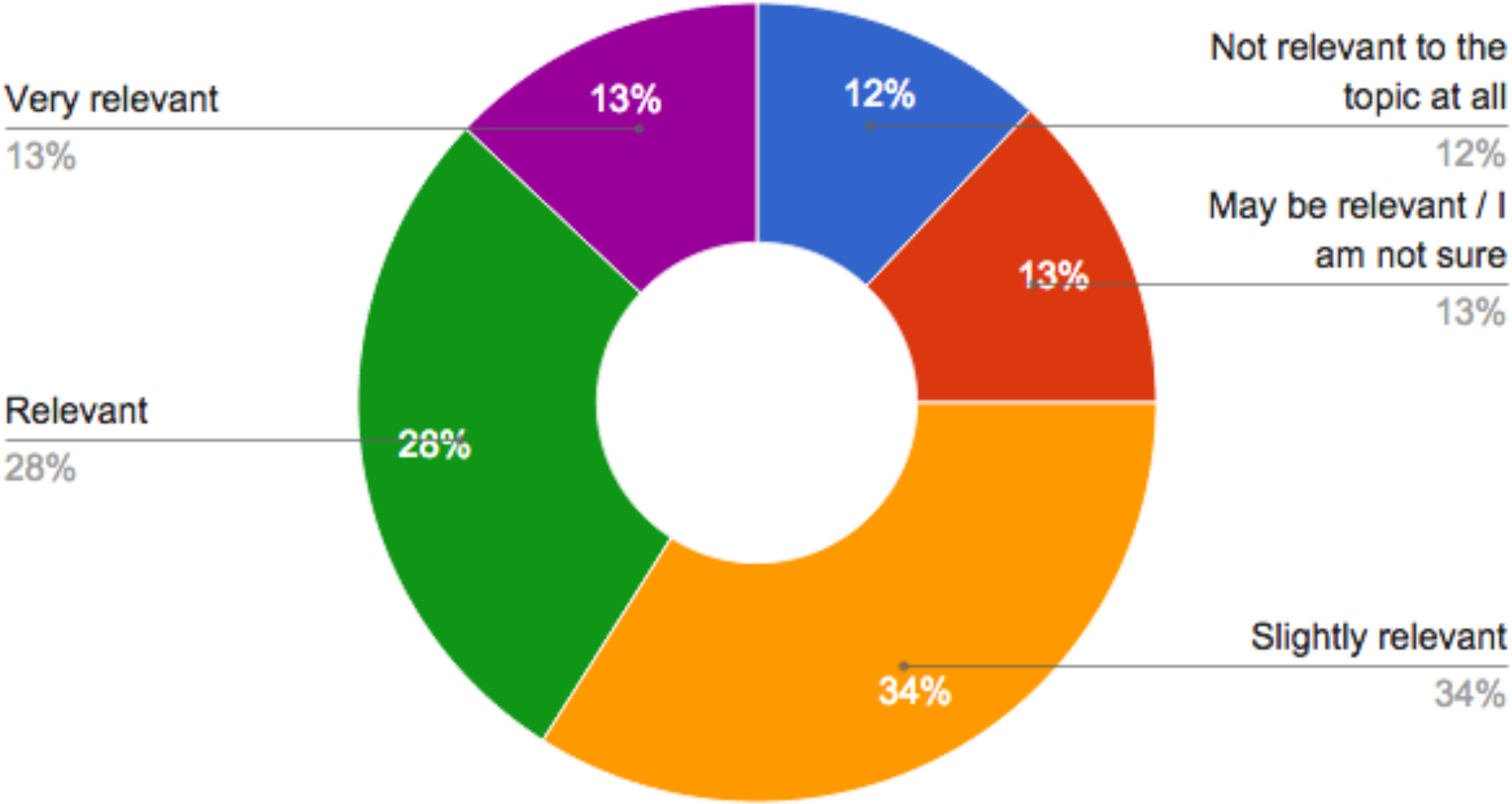
Results


V-Measure
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
k
Baseline
Wikipedia Concept
Wikipedia Concept+Categories

Human Experts Examination

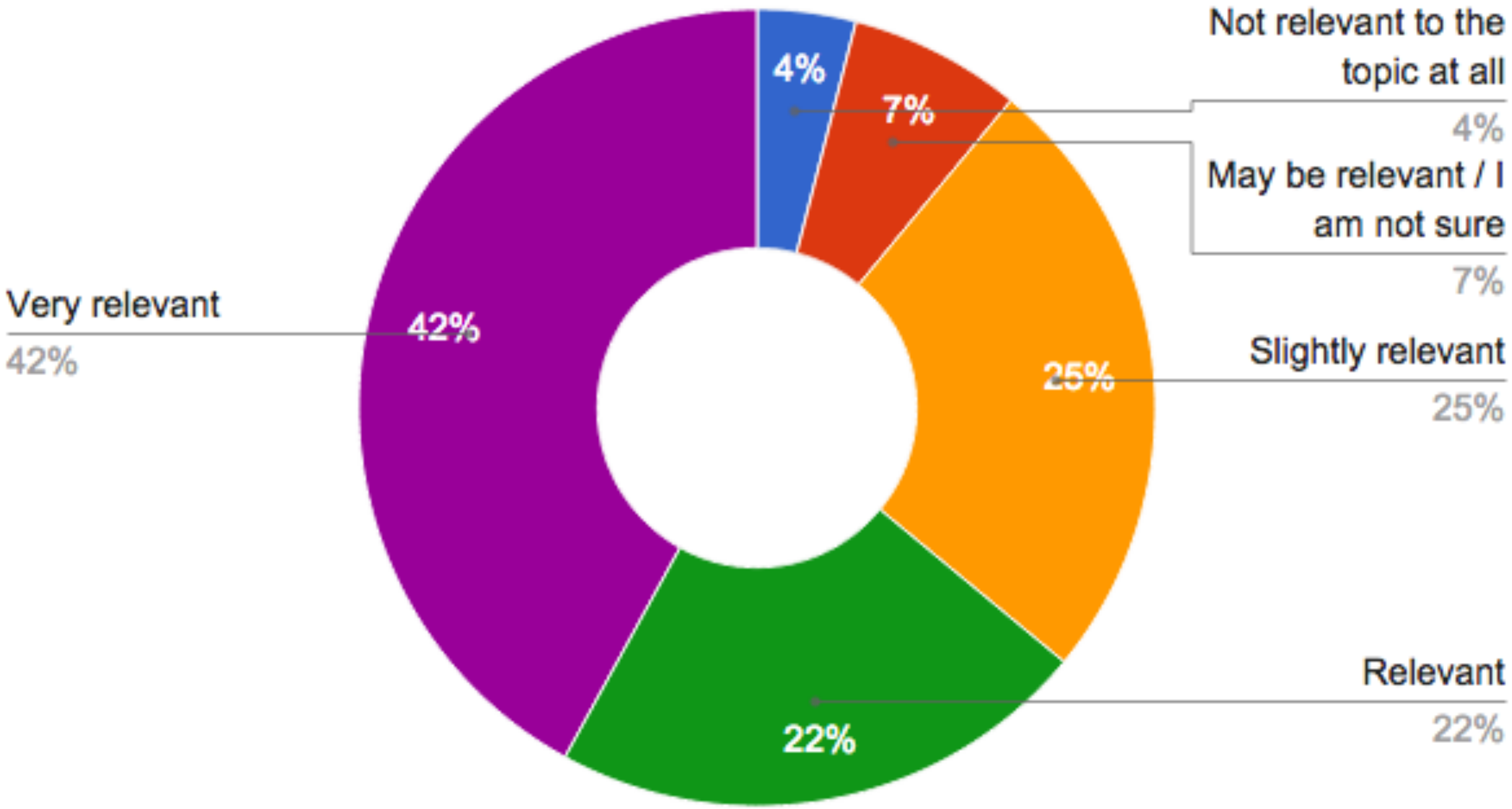
- 10 human examiners
  - 5 groups for each examiner and 10 tweets for each group
  - Given a generated cluster and ask the expert to rate the relevance from 1 ~ 5
    - 1 - Not relevant at all
    - 2 - Maybe relevant or I'm not quite sure
    - 3 - Slightly relevant
    - 4 - Relevant
    - 5 - Very relevant

Result - Baseline



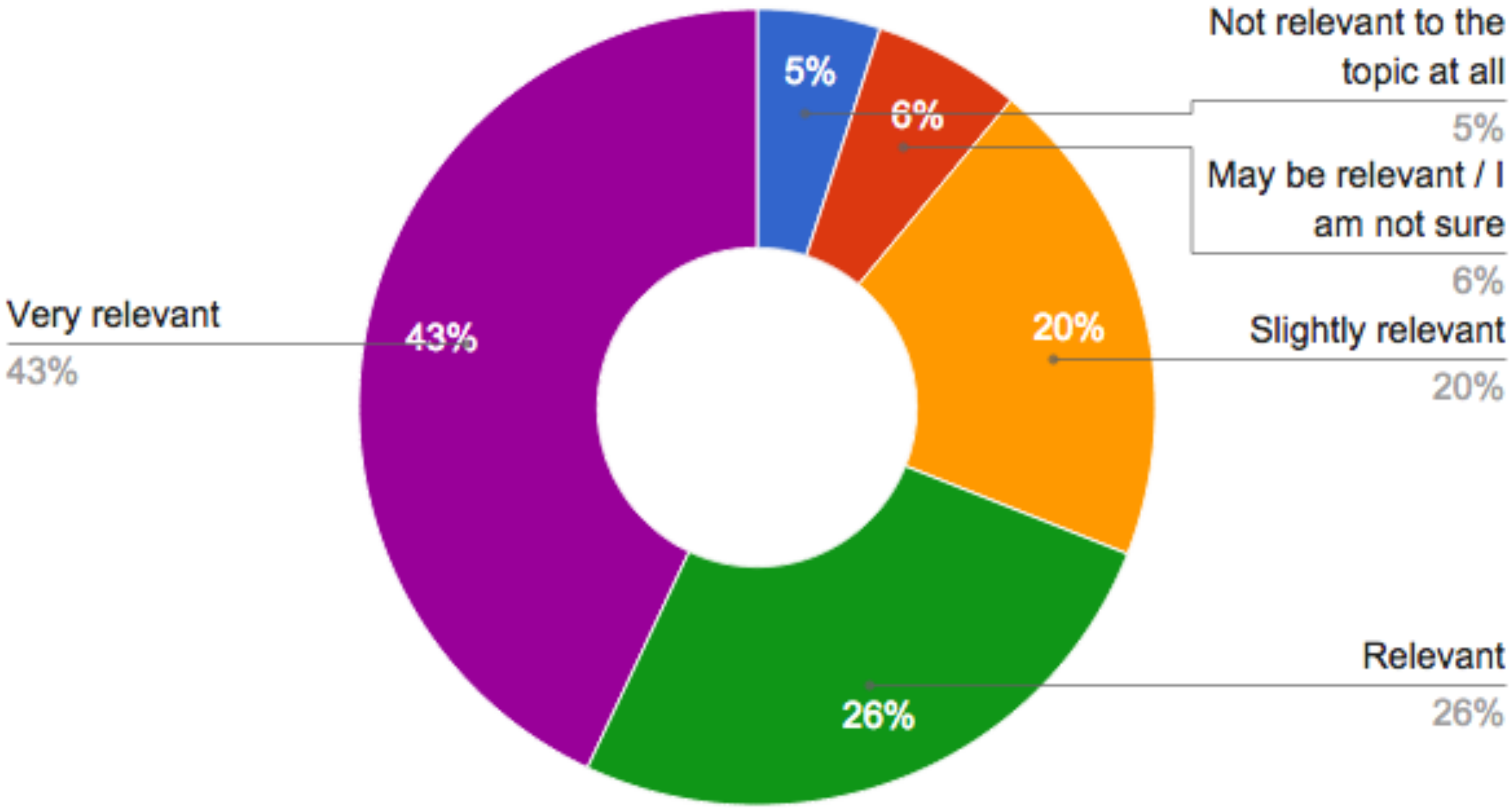
THE UNIVERSITY
of EDINBURGH

Result - Baseline + Topics



THE UNIVERSITY
of EDINBURGH

Result - Baseline + Topics + Categories



THE UNIVERSITY *of* EDINBURGH