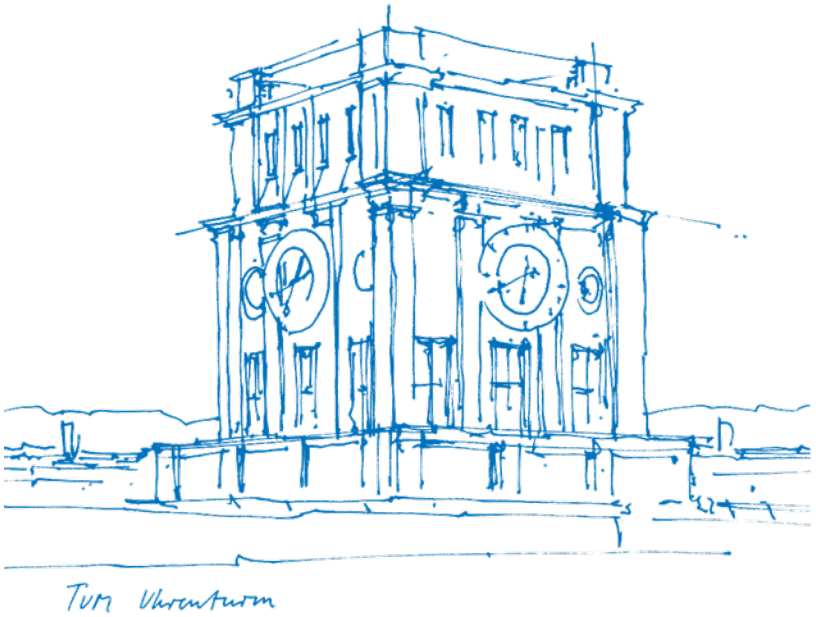
Clustering-Based Sentiment Analysis for Media Agenda Setting

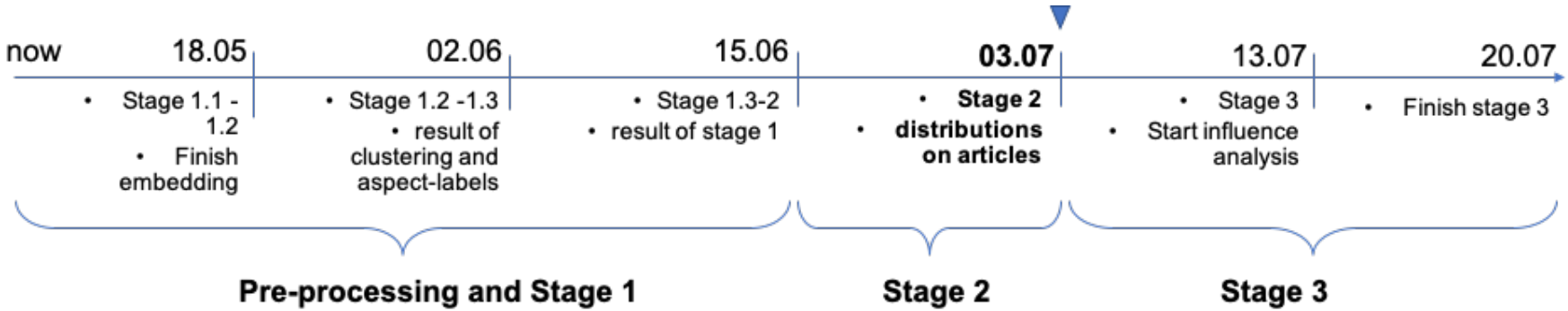
Opinion Lab Group 2.3, presentation 5

Wing Sheung Leung, Qiaoxi Liu

July 3, 2020



Milestones


now
18.05
• Stage 1.1 - 1.2
• Finish embedding
02.06
• Stage 1.2 -1.3
• result of clustering and aspect-labels
15.06
• Stage 1.3-2
• result of stage 1
03.07
• Stage 2
• distributions on articles
13.07
• Stage 3
• Start influence analysis
20.07
• Finish stage 3
Pre-processing and Stage 1
Stage 2
Stage 3

Overview

Stage 1: Generate sentence embeddings with our corpus

- 1.1 Embeddings
  - XLING sentence-level embeddings
  - Indexing sentences
- 1.2 Kmeans and Elbow Method
  - sklearn.cluster.MiniBatchKMeans
  - Elbow Method for determining optimal k
- 1.3 Identifying topics and setting sentiment
  - Generate topword list
  - Extracting topics form clustering results
  - Assigning sentiment by pre-trained model

Stage 2: Distribution

Representing sentiment distribution for single article

Re-examination on topwords and Kmeans

Output from stage 1

	global_id	corpus_name	doc_id	com_id	date	cluster	sentiment
0	0	nytimes	0	NaN	2005-11-01	7	0.12500
1	1	nytimes	0	NaN	2005-11-01	7	-0.12500
2	2	nytimes	0	NaN	2005-11-01	9	0.07500
3	3	nytimes	0	NaN	2005-11-01	7	0.50000
4	4	nytimes	0	NaN	2005-11-01	7	-0.03125

Stage 2: Distribution of articles and comments Overview

Comments from NYTimes and Spiegel

NYtimes

99 out of 327 news have comments.

Number of sentences among comments from each news:

- minmax = (3, 2535)
- mean = 481.7
- median= 216.0

Spiegel

61 out of 152 news have comments.

Number of sentences among comments from each news:

- minmax=(8, 23255)
- mean = 2309.0
- median= 697.0

Sampling

To analyse the correlations between news and its comments, we select the news with more comments.

- NYtimes: selected 52 out of 99 news with comments (each news has 200+ sentences of comments)
- Spiegel: selected 34 out of 61 news with comments (each news has 500+ sentences of comments)

Statsitc on distirbution Approach 1

Accumulating and Normalizing (separate positive and negative)

1. accumulate positive/negative sentences from each tuple

```
cluster_group = doc.groupby('cluster')
for (_,cluster) in cluster_group:
  label = cluster.cluster.unique()[0]
  doc_dic[label] = sum([abs(x) for x in scores]), sum([x for x in scores if x>0])
```

2. normalize the accumulation score

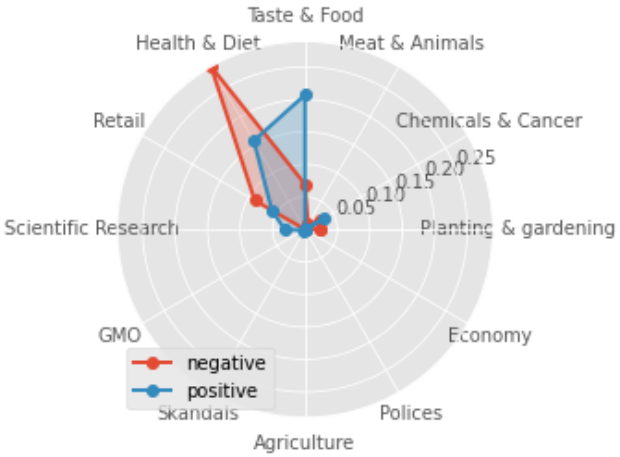
```
df_new_t['normalized_pos'] = (df_new_t[1])/all_abs_sum
df_new_t['normalized_neg'] = (df_new_t[2])/all_abs_sum
```

Example:

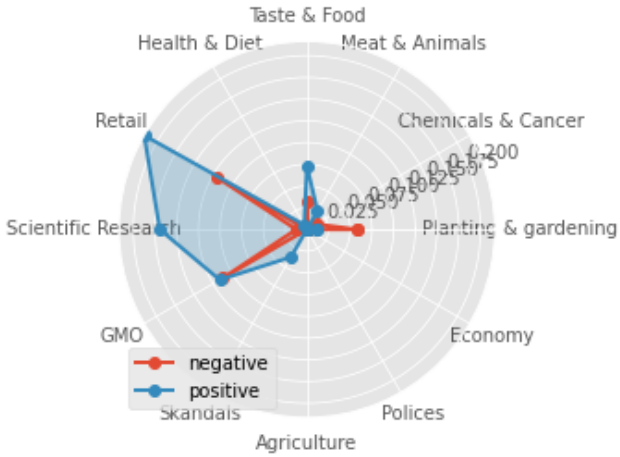
	normalized_pos	normalized_neg
Chemicals & Cancer	0.000000	-0.023515
Meat & Animals	0.032921	-0.028218
Health & Diet	0.000000	-0.009406
Retail	0.206930	-0.069604
GMO	0.157471	-0.289232
Agriculture	0.058858	-0.089356
Polices	0.031353	0.000000
Economy	0.003135	0.000000

Representing NYTimes articles

Sentiment Distribution across 12 aspects for article 0 in nytimes



Sentiment Distribution across 12 aspects for article 5 in nytimes



Statsitc on distirbution Approach 2

Summing up all sentiment scores and average

1. instead of accumulating separately, we sum up sentences from each cluster and use weights

```
for (_,cluster) in cluster_group:
  label = cluster.cluster.unique()[0]
  scores = np.array(cluster['sentiment'])
```

2. calculate weight and mean for each cluster

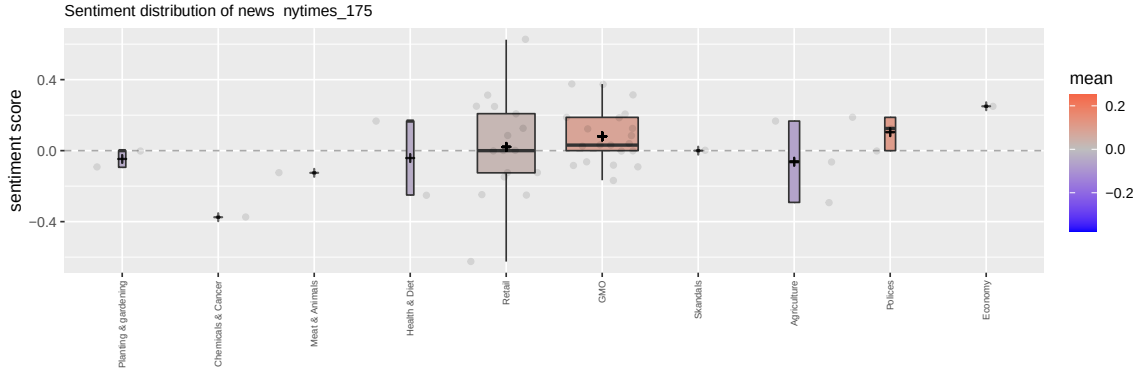
```
df = pd.DataFrame(scores,columns=['scores'])
df['weight'] = length_per_cluster/total_num_sen
df['mean'] = scores.mean()
```

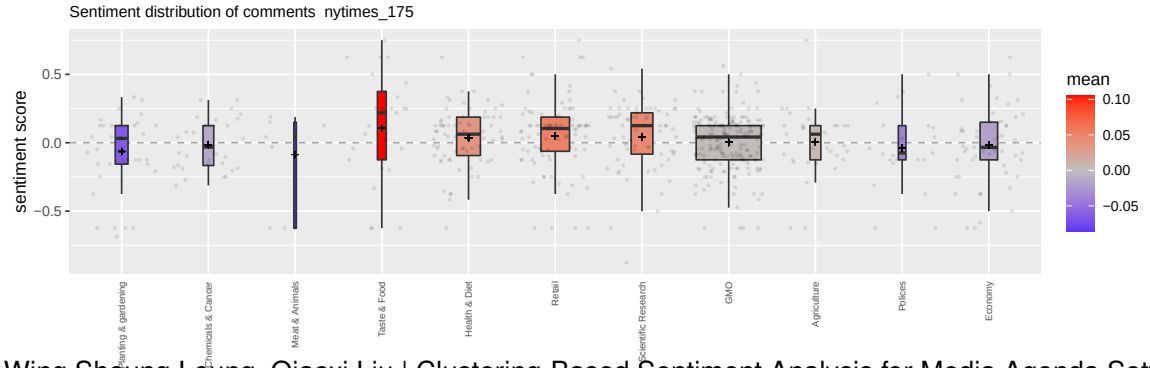
3. remove 0 score sentences in German comments ("Zitat von")
4. boxplot

```
ggplot(aes(y=scores,factor(cluster),fill=mean,weight=weight))
```

- color = mean of sentiment scores
- wider = more weight (thinner = less weight)
- higher = larger variance (flatter = smaller variance)

NYTimes: Labeling GMO





NYTimes: Labeling GMO

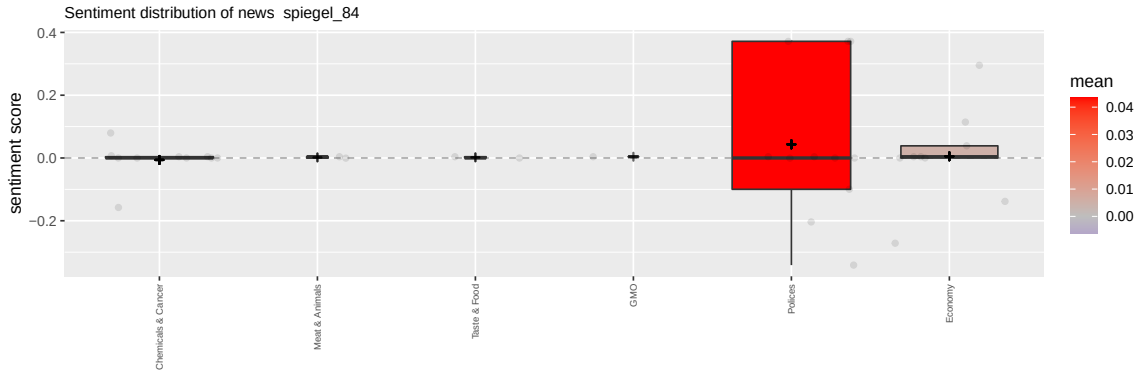
article:

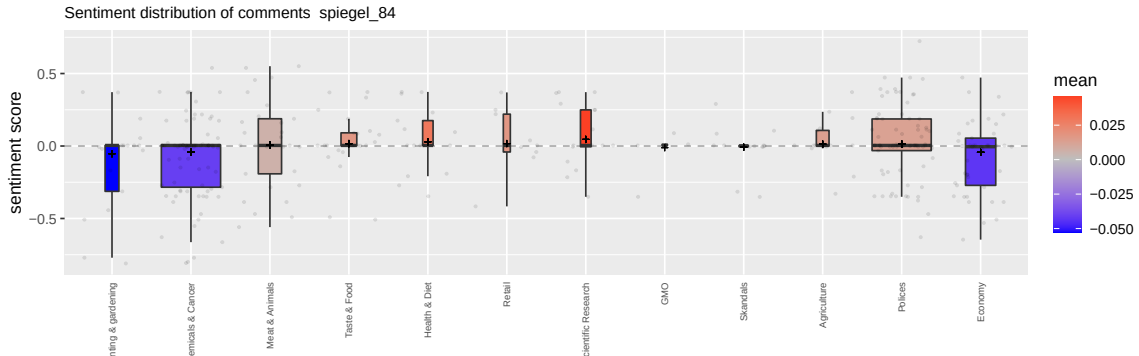
Whole Foods Market, the grocery chain, on Friday became the first retailer in the United States to require labeling of all genetically modified foods sold in its stores, a move that some experts said could radically alter the food industry. ... Whole Foods, which specializes in organic products, tends to be favored by those types of consumers, and it enjoys strong sales of its private-label products, whose composition it controls. ... He said Whole Foods looked forward to working with suppliers on the labeling.

comments:

- Sometimes genetically modifying food has unintended consequences like bad tasting tomatoes (or worse)
- I have a cousin who works at Whole Foods. He is a happy employee and loves it. Thinks it is a great company.
- I am not sure if non-GMO foods are healthier to eat but they are certainly better for the environment.
- I applaud Whole Foods for at least taking a stand.
- consuming red meat is an emotionally charged issue for many people.
- In March 2012, Harvard Public Health published the results of a study on red meat, which followed 120,000 people over 20 years. The conclusion: "Red meat consumption is associated with an increased risk of total, heart, and cancer mortality"
- Since apples are apparently the most pesticide-ridden fruit, I have gotten to like the more expensive but sweeter ones.

Spiegel: 5% Spielraum von Bionade



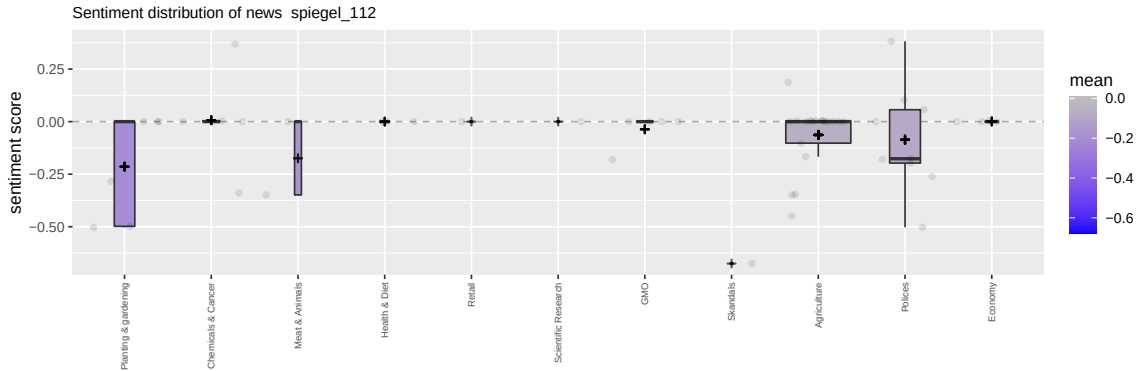


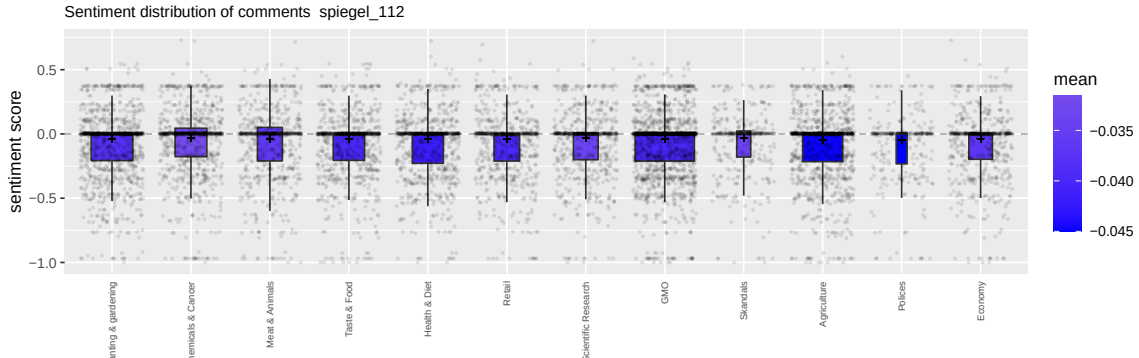
Spiegel: 5% Spielraum von Bionade

article:
Es ist eine erstaunliche Erfolgsgeschichte: Kaum ein Produkt hat den Markt in kürzester Zeit so verändert wie Bionade, das Brausegetränk aus Bayern. Jetzt aber hat das Unternehmen Probleme, Bio-Rohstoffe zu bekommen - und zeigt damit, dass sich Öko-Anspruch und Wachstum nicht immer vertragen...."Bionade nutzt Spielraum aus". Möglich ist das, weil der Gesetzgeber den Produzenten beim Bio-Siegel eine Lücke gelassen hat. Wer das grün-schwarze EU-Zertifikat auf seine Produkte drucken will, der muss nur bei 95 Prozent der Zutaten nachweisen, dass sie biologisch angebaut sind. "Diesen Spielraum von fünf Prozent nutzt Bionade aus", sagt Markwardt.
comments:

- Wie sollen Menschen in Deutschen Staedten Litschi anbauen ?...das Thema Biolebensmittel und Biolandwirtschaft viel zu komplex ist, ... Nur Städter am Schreibtisch können sich dafür begeistern,welche von Tierhaltung,Biologie,Ökologie nur Theoriewissen haben... Kaum jemand hat selbst mal Hühner gehalten geschweige denn Großtiere. ...Wer hat z.B.schon mal tatsächlich in der Landwirtschaft gearbeitet,Gemüse angebaut,Nutztiere gehalten ?doch kaum jemand! ...daß hier mal wirklich über WAS IST BIO diskutiert wird .
- Ich selbst habe mich maßlos über die unverschämte Preiserhöhung von Bionade geärgert.
- Möglichst billige Rohstoffe, möglichst billige Angestellte.
- Arbeitswirtschaftlich sind Bestände über 500 Tier ...bei den niedrigen Erzeugerpreisen nicht von Hand zu füttern. Zudem gibt es sehr viele (unabhängige) Untersuchungen die gerade in der Geflügelhaltung technische Hilfsmittel mit Blick auf die Tiergesundheit sowie die Tierbedürfnisse positiv bewerten.

Spiegel: Der Skandal um Dioxin





Spiegel: 5% Spielraum von Bionade

article:

Es ist einer der größten Giftskandale der vergangenen Jahre: Bis zu 3000 Tonnen dioxinverseuchtes Fett wurden laut Bundeslandwirtschaftsministerium an 25 Futtermittelhersteller in mindestens vier Bundesländern geliefert. Wo das Gift von dort aus hingelangte und welche Mengen an Nahrungsmitteln belastet sind, ist weitgehend unklar. Verbraucher reagieren zunehmend verunsichert: Der Verkauf von Hühnereiern ist "spürbar" gesunken, teilte die landwirtschaftliche Marktberichterstattungsstelle MEG mit. Welche Gefahren drohen durch die Einnahme von Dioxin? Welche Vorsichtsmaßnahmen können getroffen werden? SPIEGEL ONLINE gibt Antworten auf sieben Fragen.

comments:

- Der Dioxin-Skandal war hoffentlich nicht der letzte. Es sollten so viele wie möglich vorkommen. Am besten aber wäre, wenn ein paar Konsumenten nachweislich an solchen oder anderen Giftstoffen in Lebensmitteln sterben.
- 3000 Tonnen verseuchtes Tierfutter - das ist ein Terroranschlag.
- Ich kann und will niemandem verbieten Fleisch von deutschen Rindern zu essen.
- den gesetzlichen Vorgaben ist so eine Sache. Fahren Sie mal mit einem Fiat 500 mit 80 km/h frontal gegen eine S-Klasse!. Ihr Vergleich hinkt doch wohl. Wer sich mit Bio-Artikeln überfrisst, stirbt auch. Also sind Bio-Lebensmittel auch lebensgefährlich, wenn man falsch damit umgeht. Guten Appetit.

Manually observed Features

- Spiegel
  - news: Kritical (suspection)
  - comments: Sarkasmus, irony, not direct, more philosophic, spreading among more topics ,
- NYtimes
  - news: More like an advertisment for stakeholder
  - comments: simpler, clear statement (against or for), benefit or not

Next Plan

- To improve the visualization and capture more patterns
- To measure the influence of articles onto commenters using Granger coefficients and cross-lagged correlations

Re-examination on topwords and Kmeans

Original methodology

1. Kmean clustering on embeddings of both English and German text all together
2. Find topwords with naive method, simple term-frequency count with NLTK and self-defined stopwords
3. Find topwords with calculating clarity scoring

$$score_a(w) = t_a(w) log_2 \frac{t_a(w)}{t(w)}$$

, where $t_a(w)$ and $t(w)$ are the l1-normalized tf-idf scores of w in the segments annotated with aspect a (i.e. topic / cluster) and in all annotated segments

Re-examination on topwords and Kmeans

Second methodology

1. Kmean clustering on embeddings of both English and German text all together
2. Calculate clarity scoring ***separately for English and German content***

$$score_a(w_{language}) = t_a(w_{language}) log_2 \frac{t_a(w_{language})}{t(w_{language})}$$

, where $t_a(w_{language})$ and $t(w_{language})$ are the l1-normalized tf-idf scores of w in the segments annotated with aspect a (i.e. topic / cluster) for a certain language and in all annotated segments in that langauage

3. Pick 10 topwords with highest clarity score in each cluster of each language, total 20 words for a cluster
4. Sort the words by their relative clarity scores

$$RelativeClarityScore_{language} = \frac{score_{language,i}}{\sum_{i=1}^{n} score_{language,i}}$$

Re-examination on topwords and Kmeans

Additional try

1. Kmean clustering on embeddings of German text and English separately
2. Find topwords with calculating clarity scoring ***separately for German content and English content***

$$score_a(w_{language}) = t_a(w_{language}) log_2 \frac{t_a(w_{language})}{t(w_{language})}$$

, where $t_a(w_{language})$ and $t(w_{language})$ are the l1-normalized tf-idf scores of w in the segments annotated with aspect a (i.e. topic / cluster) for a certain language and in all annotated segments in that langauage

Appendix

(a) Original methodology: naive method via simple term-frequency count

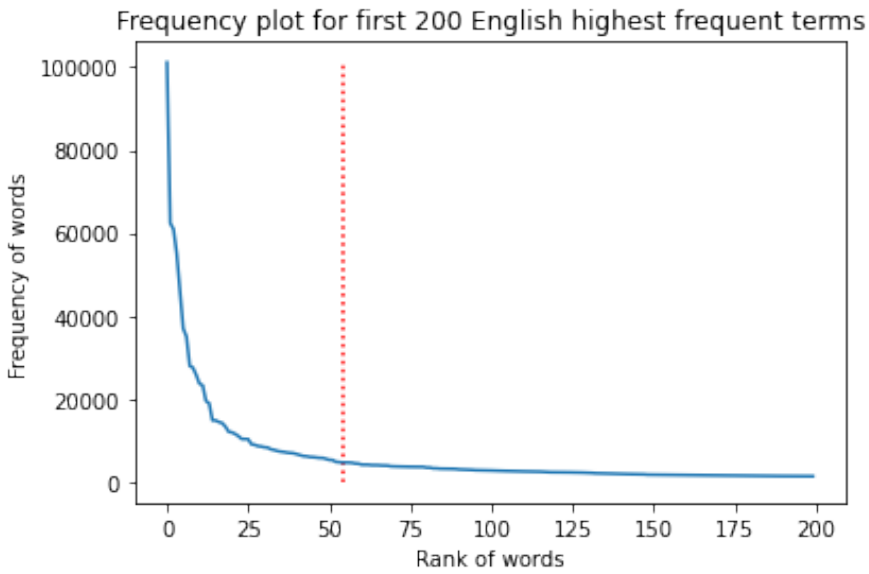
(b) Original methodology: clarity scores with NLTK and sklearn stopwords

(c) Original methodology: clarity scores with TF-High stopwords

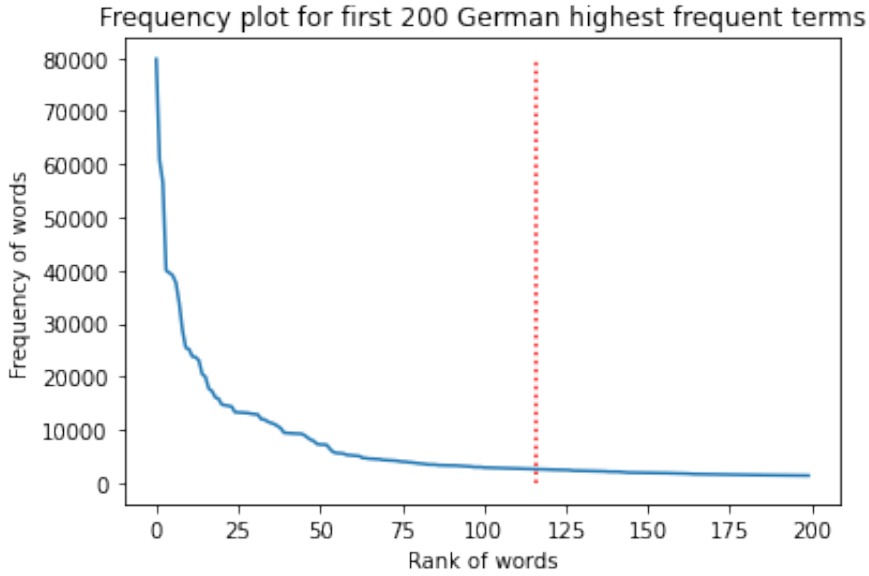
(a) Second methodology on English corpus: clarity scores with NLTK and sklearn stopwords

(b) Second methodology on German corpus: clarity scores with NLTK and sklearn stopwords

(c) Top 20 words from two languages sorted by their relative clarity scores



(a) Second methodology on English corpus: term-frequency distribution



(b) Second methodology on German corpus: term-frequency distribution

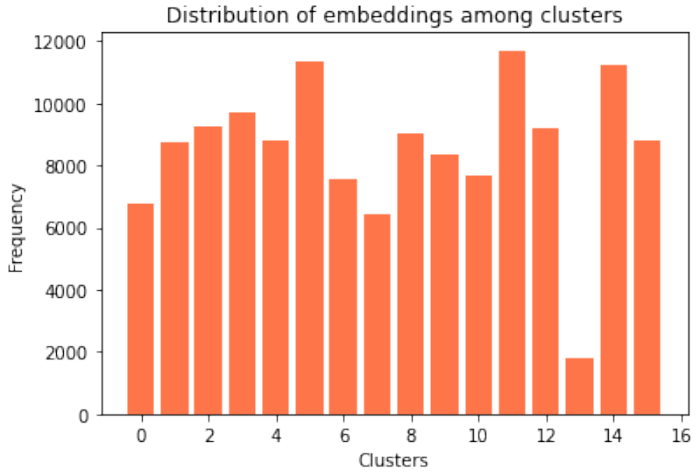
(a) Second methodology on English corpus: clarity scores with TF-high stopwords

(b) Second methodology on German corpus: clarity scores with TF-High stopwords

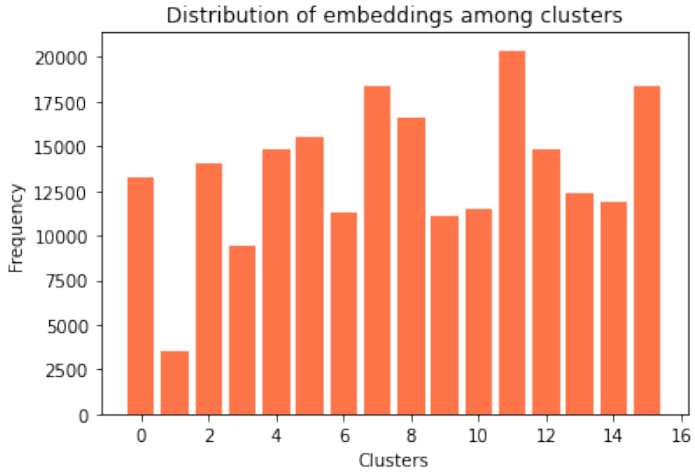
(c) Top 20 words from two languages sorted by their relative clarity scores

(a) Additional try for English corpus only: clarity scores with TF-high stopwords
Remarks: The cluster number is not match as shown before

(b) Additional try for German corpus only: clarity scores with TF-high stopwords
Remarks: The cluster number is not match as shown before



(a) Additional try for English corpus only: embedding distribution
Remarks: The cluster number is not match as shown before



(b) Additional try for German corpus only:embedding distribution
Remarks: The cluster number is not match as shown before