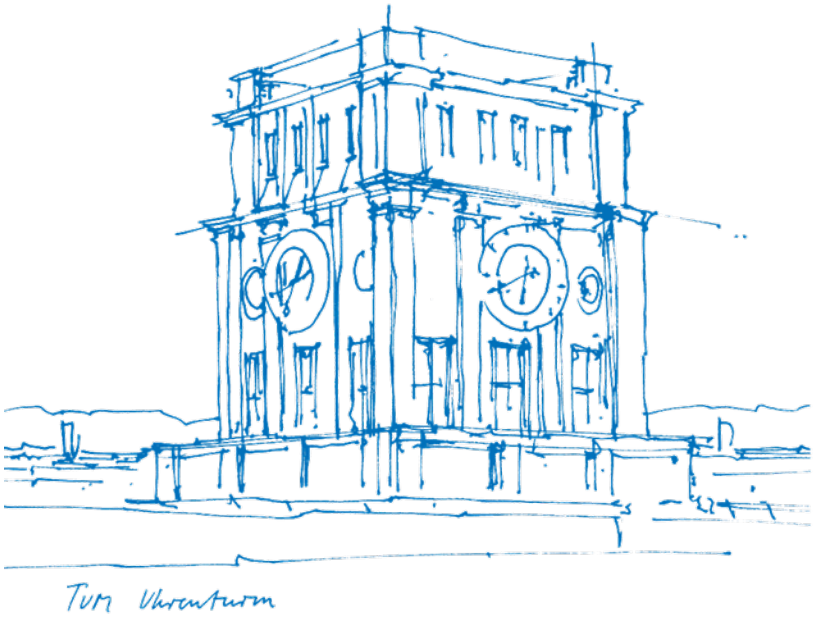
Technical University of Munich

# Clustering-Based Sentiment Analysis for Media Agenda Setting

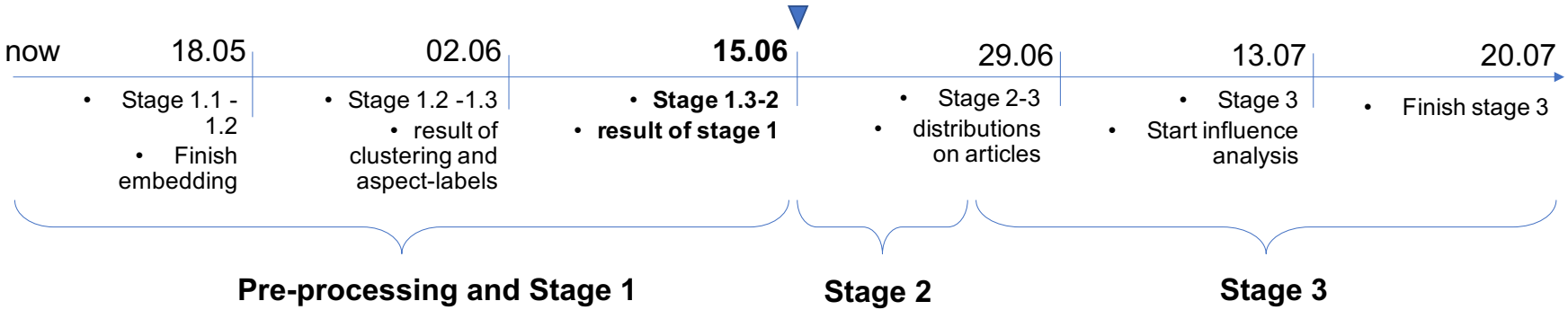
Opinion Lab Group 2.3, presentation 4

Wing Sheung Leung, Qiaoxi Liu

June 16, 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
29.06
• Stage 2-3
• 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

# Selecting global minimum in AIC as desired k

k = 16

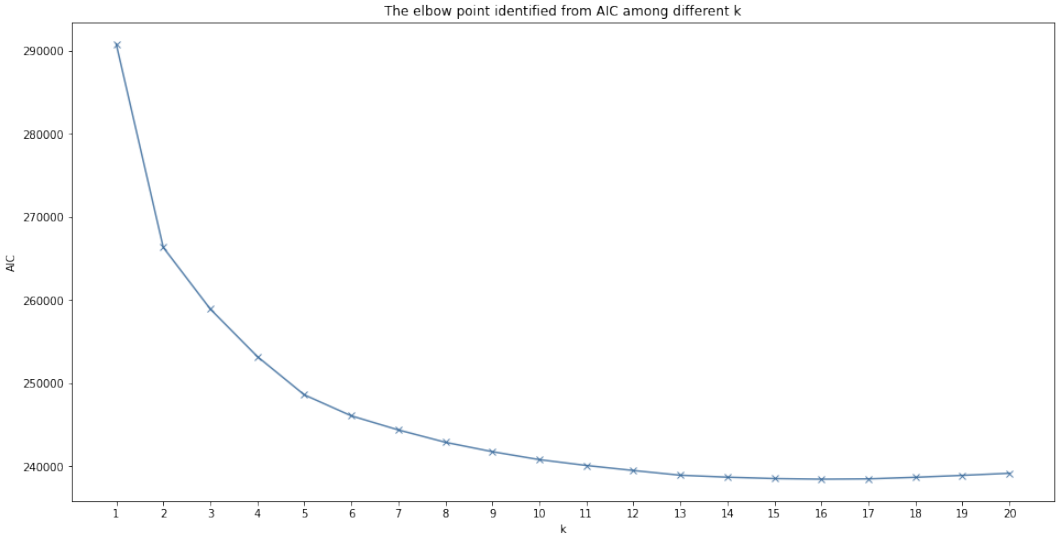


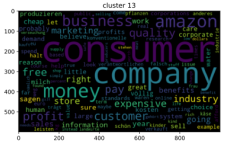

**Figure:** Steeper slopes are found from k = 1 to 5, and the changes becomes gentler from k = 5 to 14. Then, it is almost flat from k = 14 to 16. It reaches **the lowest point when k = 16** and starts going up very slowly.

# Word cloud generated by naive method (Recap) I

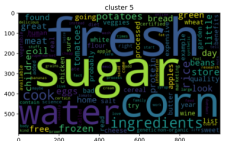
Naive method: sorting 1-gram words by their own frequencies within a cluster with NLTK and customized stopwords and stemming


**7: . Retailers & brands**
Believe it or not, in my neighborhood (Chelsea) the cheapest produce and groceries are at Whole Foods.

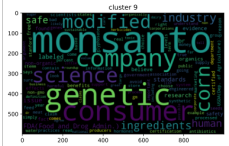

**0: Garbage ?**
- I do not see why so many people are making fun or looking down on this.
- There is no way around that.
- This is a complex issue.



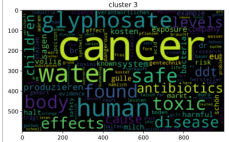

**13. Economy & Costs**
"Bio ist teurer - und das ist gut so", sagt Michael Radau vom Verband der Biosupermärkte.




**5: Garbage ?**
- Entspannt euch, Leute!.
- Na dann guten Apetit mit Hühnchen für 2,99 \x80.




**9. GMO & chemicals**
Said GMOs have not been shown by any credible science to cause any harm to consumers.




**3. Human health & disease**
In fact, under a microscope, cancer cells from a teenager with osteosarcoma are indistinguishable from a any breed dog's bone cancer cells

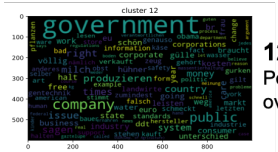

**6. Nutrition & lifestyle**
Conclusion: The published literature lacks strongsignificantly more nutritious than conventional foods.


**8. Evidence**
Now, I will give you an idea of an experiment to show you how little man does know.
Was künstlich ist habe ich bereits definiert.

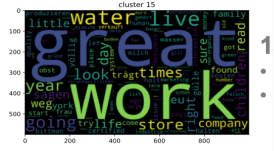


**Figure:** First part of word cloud generated by naive method for 8 clusters with sample sentences

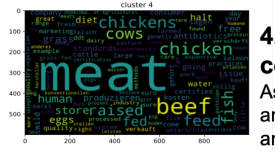
# Word cloud generated by naive method (Recap) II




**12. Politics**
Politicians get in fist fights and there are riots over the issue in So.

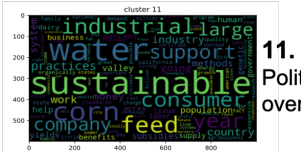



**15. Garbage cluster**
- Auch Kabeljau, den ich gekocht liebe.
- He doesn't care, so I'm trying to maintain happiness in the relationship.



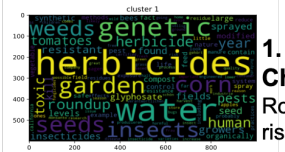

**4. Animal welfares & meat consumption**
As long as you dont prevent me from eating animals, I wont prevent you from not eating animals.


**10. Garbage cluster**
- Die legen bei der Kälte keine Eier.
- Isaac Bozeman VT please advise the peer review scientific papers you are reciting.

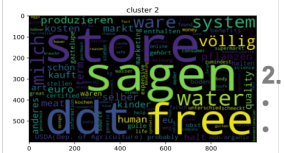



**11. Production & Agriculture**
Politicians get in fist fights and there are riots over the issue in So.


**14. Garbage cluster**
- Na viel Spaß.
- And raw does not mean better.




**1. Environment & Gardening & Chemicals**
Roundup is according to my knowledge a less risky herbicide which deteriorates in little time.




**2. Garbage cluster**
- traurig aber wahr!
- Thanks for your answer.

**Figure:** Second part of word cloud generated by naive method for 8 clusters with sample sentences

# Word cloud generated with clarity scoring I

Clarity scoring: measuring how much more likely to observe a word w in topic k

The more positive the clarity score, the more important the word to be existed in topic k


**Figure:** Word cloud generated with clarity scores with stopwords defined by *NLTK, sklearn and some terms existing in all clusters*

# Word cloud generated with clarity scoring II

With other stopwords

Mentioned in Saif, Fernandez, and Alani 2014 *"On stopwords, filtering and data sparsity for sentiment analysis of Twitter"*, based on Zipf's law, stopwords can be selected among

1. most frequent words (TF-High)
2. words occured once, i.e. singleton words (TF1)
3. words with low inverse document frequency (IDF)

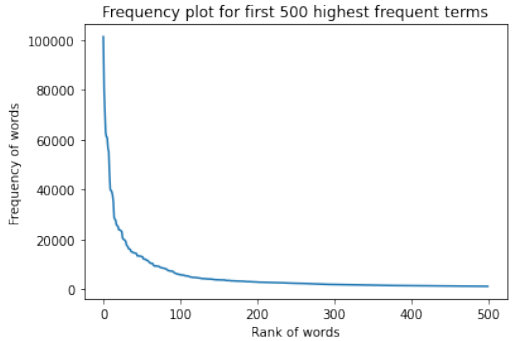



**Figure:** The term frequency plot fits Zipf's law, which states the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.

# Word cloud generated with clarity scoring III

Stopwords: the first 430 most frequent terms (TF-High) except 73 potential topic related terms (e.g. pesticides, farmer and GMO)


**Figure:** Word cloud generated with clarity scores with *TF-High* stopwords

# Word cloud generated with clarity scoring IV

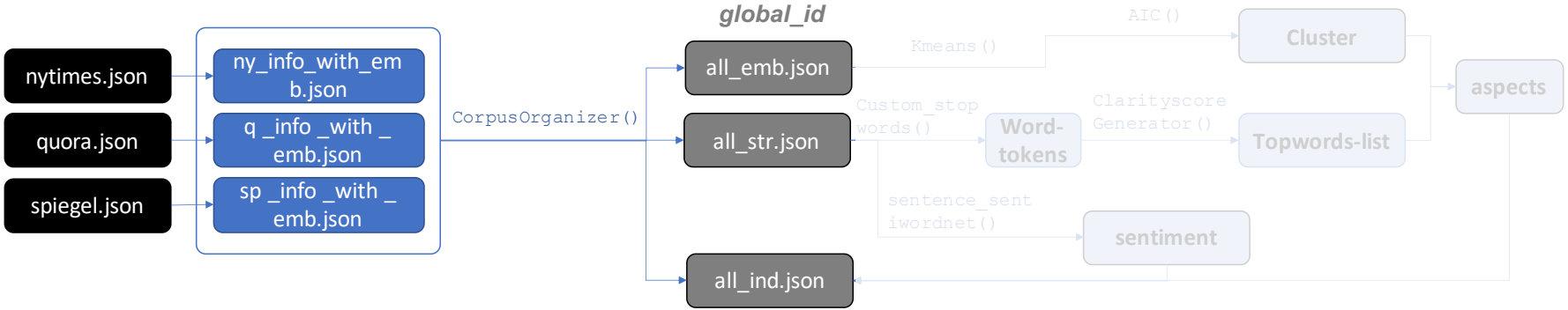
Stopwords: singleton words (TF1)


**Figure:** Word cloud generated with clarity scores with *TF1* stopwords

# Topwords and garbage clusters

| Cluster | Topic name | Top words sorted by clarity scores |
|---|---|---|
| 0 | Garbage | - |
| 1 | Planting and gardening | pesticide, plant, soil, crop, grow, fruits, water, fertilizer |
| 2 | Garbage | - |
| 3 | Chemicals and cancer | cancer, chemical, safe, water, dioxin, gar, verbraucher |
| 4 | Meat and animals | meat, animal, feed, farm, landwirtschaft, farmer, gar |
| 5 | Taste and food | taste, fresh, milk, fruit, corn, vegetables, ingredients |
| 6 | Health and diet | eating, health, diet, vegetables, real, fruits, life |
| 7 | Retail | store, grocery, company, market, price, fresh, buy |
| 8 | Scientific research | science, studies, research, article, human, life, read |
| 9 | GMO | gmo, monsanto, genetically, label, products, modified, corn |
| 10 | Skandals | welt, produkte, skandal, dioxin, schuld, milch, politik |
| 11 | Agriculture | farmer, agriculture, crop, grow, soil, water, conventional |
| 12 | Polices | government, fleisch, verbraucher, tiere, landwirtschaft, times, problem |
| 13 | Economy | money, consumers, price, company, business, product, marketing |
| 14 | Garbage | - |
| 15 | Garbage | - |

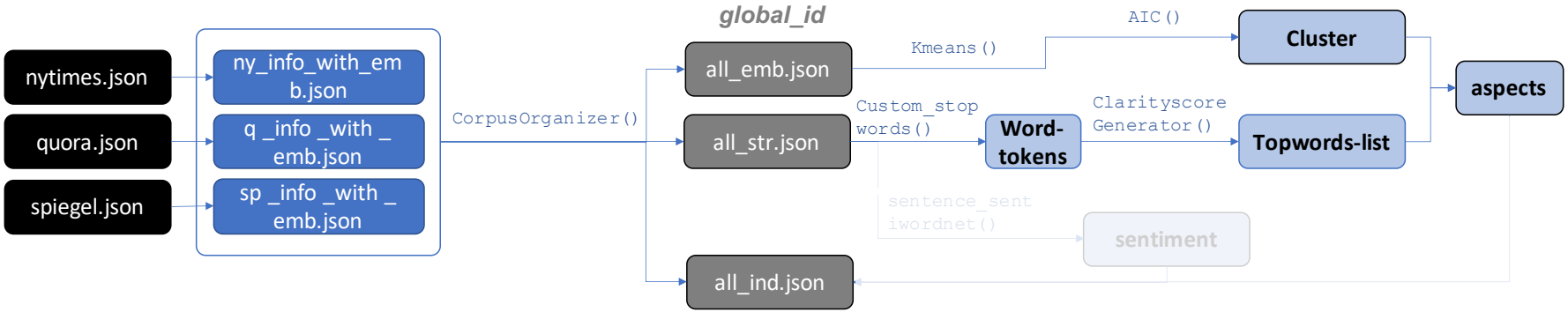
# Recap our journey




Example sentence: *"Like most grocery chains, Whole Foods does not release sales data on individual stores."*

| global_id | corpus_name | doc_id | com_id |
|---|---|---|---|
| 30120 | nytimes | 191 | NaN |

# We are now here!




Example sentence: *"Like most grocery chains, Whole Foods does not release sales data on individual stores."*

| global_id | corpus_name | doc_id | com_id | **cluster** |
|---|---|---|---|---|
| 30120 | nytimes | 191 | NaN | **7** |

# Sentiment Analysis

first try: Vader Sentiment[1]

Features

- working directly on sentence-level
- pre-trained on social media like tweets, nytimes, movie reviews courpus

Drawbacks

- No multilingual
- only work for simple comments on social media, limited by their pre-training corpus

Example

"... McDonald’s will soon be serving a coffee that comes from organic beans and is certified ..."

result: {’compound’: 0.3182, ’neg’: 0.0, ’neu’: 0.943, ’pos’: 0.057}

[1] Hutto and Gilbert 2014.

# Sentiment Analysis

second try: Sentiwordnet[3] & SentiWS[4]

Features

- based on POS (part-of-speech) tagging: ADJ,NOUN,ADV,VERB
- EN: using synset (synonym set) to compute average sentiment score
- DE: semantic orientation based on PMI[2]
- sentiment strength score: $\in [-1, 1]$

Example:

"Like most grocery chains, Whole Foods does not release sales data on individual stores."

('Like', 'IN'), ('most', 'JJS'), ('grocery', 'NN'), ('chains', 'NNS'), (',', ','), ('Whole', 'NNP'),...

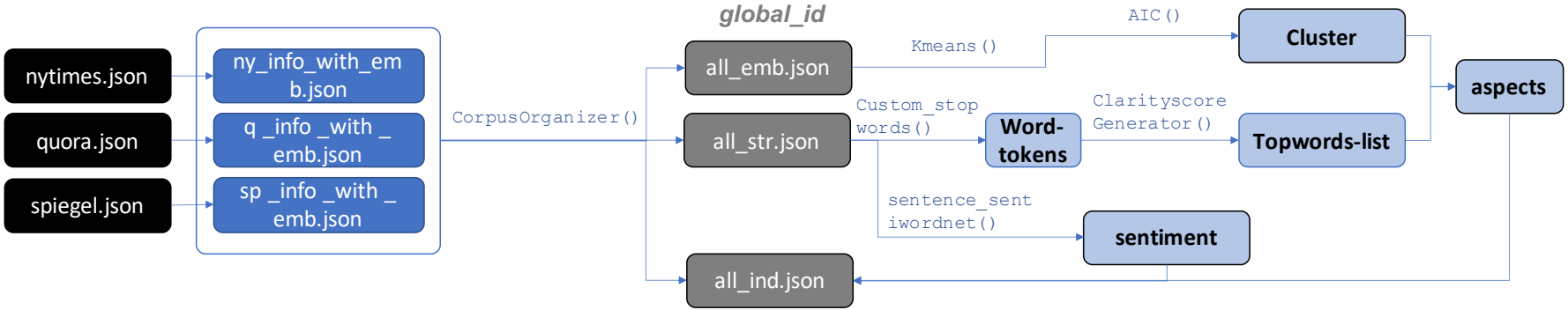
result: -0.625

[2]Remus, Quasthoff, and Heyer 2010.

[3]**minemyopinion**.

[4]Remus, Quasthoff, and Heyer 2010.

# Back to our journey




Example sentence: *"Like most grocery chains, Whole Foods does not release sales data on individual stores."*

| global_id | corpus_name | doc_id | com_id | **cluster** | **sentiment** |
|---|---|---|---|---|---|
| 30120 | nytimes | 191 | NaN | **7** | **-0.625** |

# Accumulating and Normalizing (naive approach)

accumulate all sentences tuple **(<cluster>,<sentiment>)**

```
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])
```

normalize

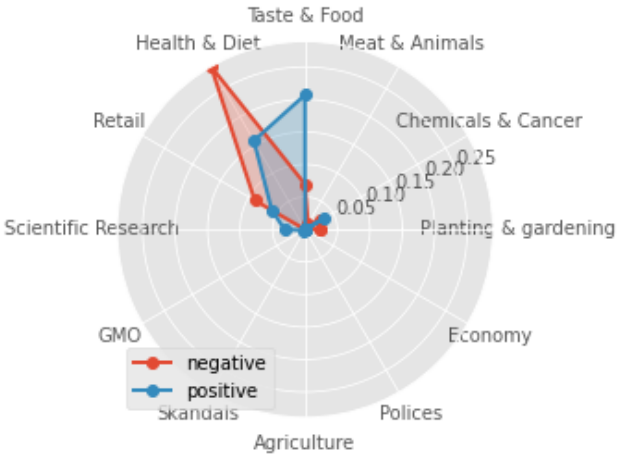
```
df_new_t['normalized_pos'] = (df_new_t[1])/all_abs_sum
df_new_t['normalized_neg'] = (df_new_t[2])/all_abs_sum
```

Example:

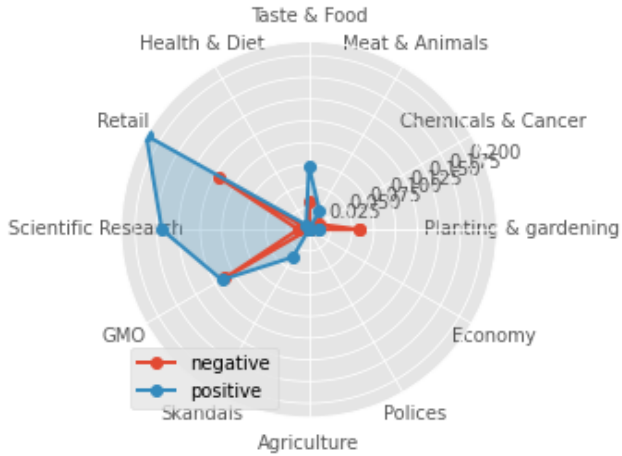
| | normalized_pos | normalized_neg |
|---|---|---|
| Chemicals & Cancer | 0.000000 | -0.023515 |
| Meat & Animals | 0.032921 | -0.028218 |
| Health & Diet | 0.000000 | -0.009406 |
| Retail | 0.206930 | -0.069604 |
| GMO | 0.157471 | -0.289232 |
| Agriculture | 0.058858 | -0.089356 |
| Polices | 0.031353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

# Representing NYTimes articles

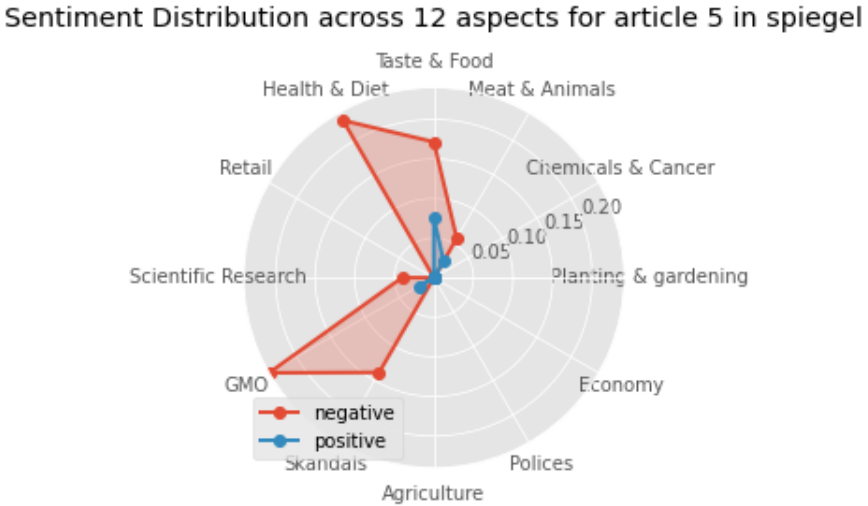
Sentiment Distribution across 12 aspects for article 0 in nytimes

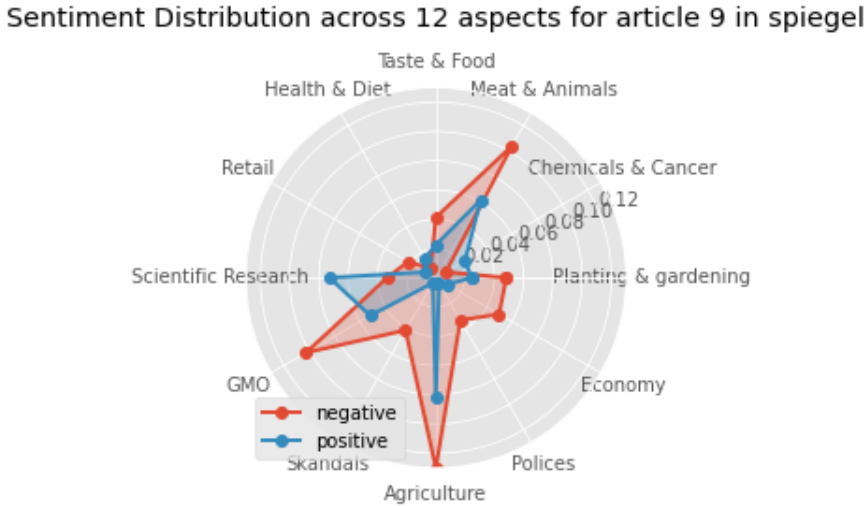



Sentiment Distribution across 12 aspects for article 5 in nytimes



# Representing Spiegel articles



# Distribution on whole corpus (without comments)

In total there are 875 documents. nytimes has 327 documents. quora has 396 documents. (there are more articles that only have comments) spiegel has 152 documents.

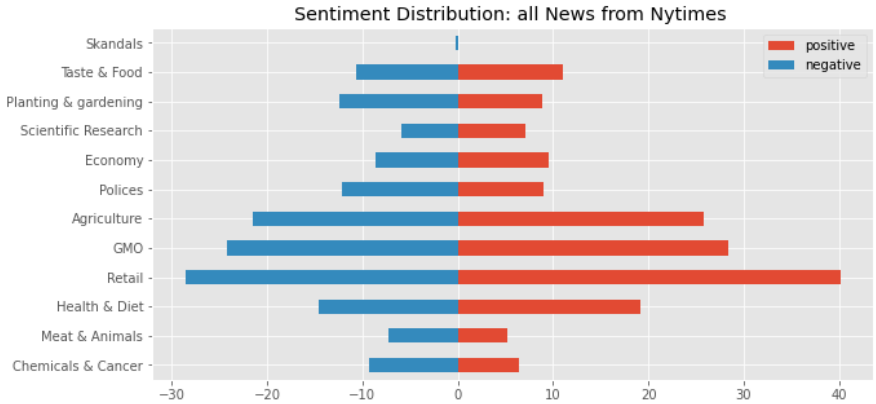



**Figure:** All articles from NYTimes

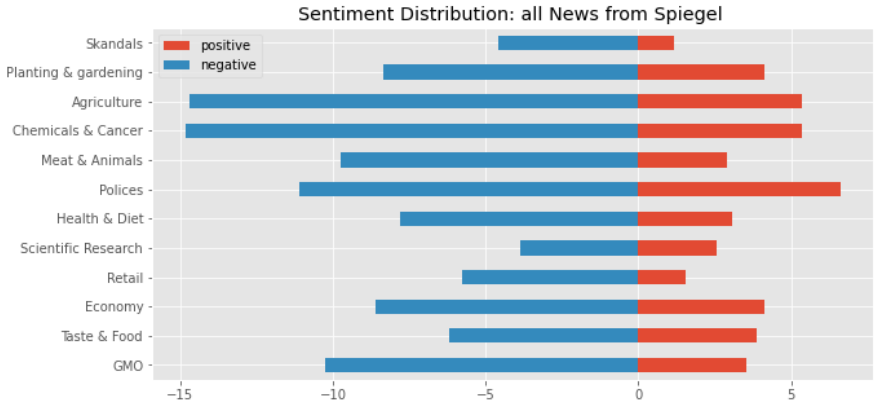
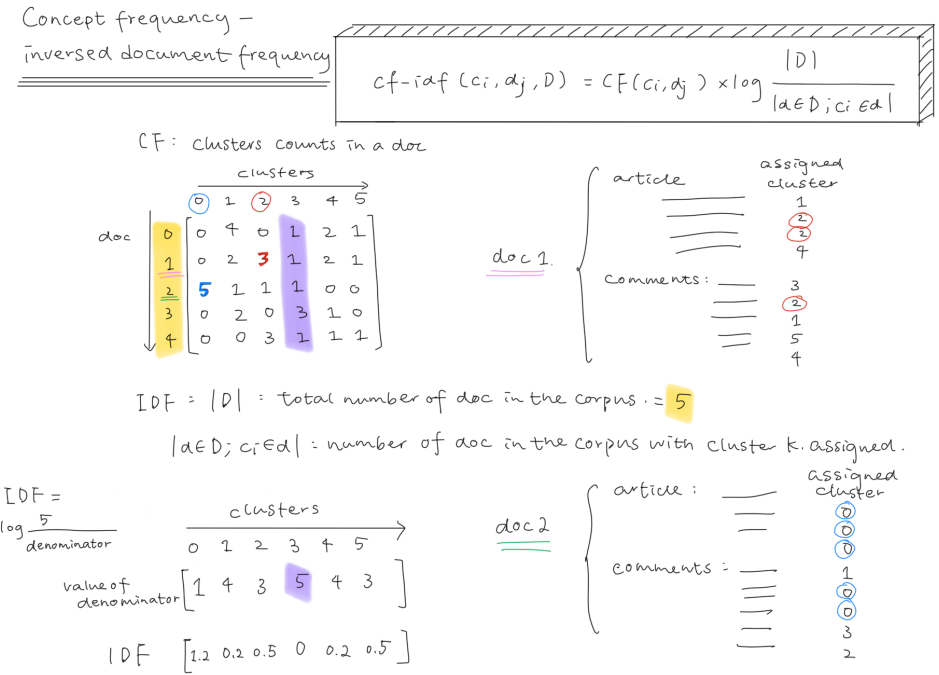



**Figure:** All articles from Spiegel

# Concept frequency - inverse document frequency (CF-IDF)



# CF-IDF results

```
IDF = np.log(TOTAL_NUM_OF_DOC / num_of_doc_per_cluster)
CFIDF = CF * IDF
print(CF[0:3], IDF, CFIDF[0:3])
```

```
CF:
[[ 1.  0.  0.  2.  3.  0.  1. 15.  0. 26.  0. 13.  4.  2.  1.  1.]
 [ 1.  9.  0.  1.  0.  4. 12. 14.  3. 12.  0.  6.  0.  1.  0.  2.]
 [ 0.  0.  0.  0.  0.  1.  0.  2.  1.  0.  0.  3.  0.  0.  0.  2.]]

IDF:
[0.35330728 0.80764074 0.45355535 0.70697101 0.56209002 0.4187323
 0.54964371 0.626757   0.71591962 0.66766464 0.97480679 0.57762431
 0.53735041 0.43302426 0.3059702  0.26291356]

CF-IDF:
[[ 0.35330728  0.          0.          1.41394202  1.68627005  0.
   0.54964371  9.40135496  0.         17.35928068  0.          7.50911602
   2.14940164  0.86604853  0.3059702   0.26291356]
 [ 0.35330728  7.26876666  0.          0.70697101  0.          1.6749292
   6.59572451  8.77459797  2.14775885  8.0119757   0.          3.46574585
   0.          0.43302426  0.          0.52582713]
 [ 0.          0.          0.          0.          0.          0.4187323
   0.          1.253514    0.71591962  0.          0.          1.73287293
   0.          0.          0.          0.52582713]]
```

**Figure:** Sample output for first 3 documents and IDF for the whole corpus

# CF-IDF radar plot example

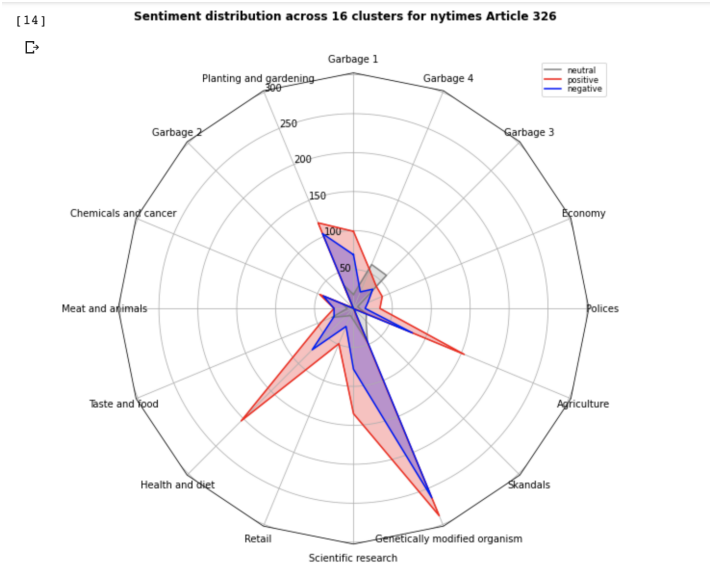



**Figure:** value $=$ CF-IDF $\times$ number of sentiments

# Next Plan

- To represent the newspaper article and its comments (separately?) using CF-IDF
- To analysis the correlations between articles of NYTimes and comments of NYTimes

# References


- Hutto, C. J. and E. Gilbert (2014). “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. In: *Eighth international AAAI conference on weblogs and social media*.
- Remus, R., U. Quasthoff, and G. Heyer (2010). “SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.”. In: *LREC*. Citeseer.
- Saif, H., M. Fernandez, and H. Alani (2014). “On stopwords, filtering and data sparsity for sentiment analysis of twitter”. In: *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC’14)*, pp. 810–817.