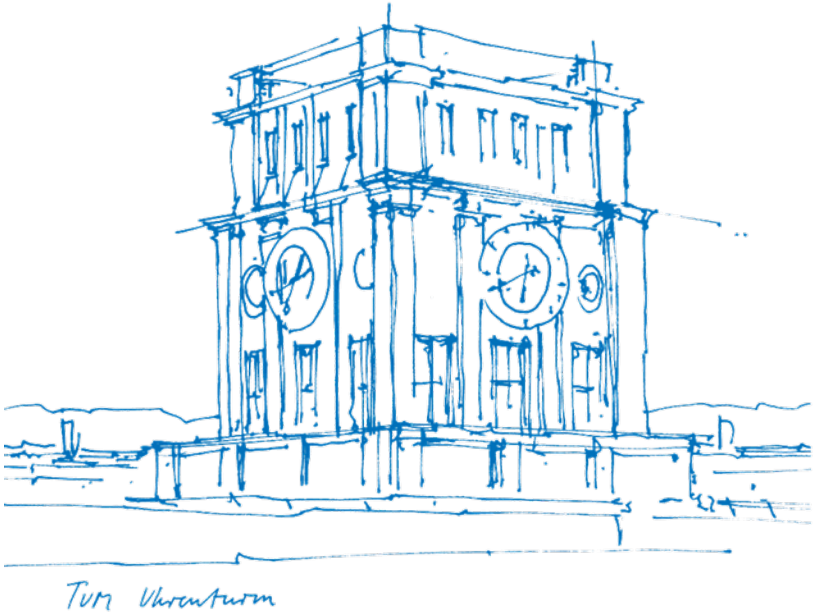
Technical University of Munich

Clustering-Based Sentiment Analysis for Media Agenda Setting

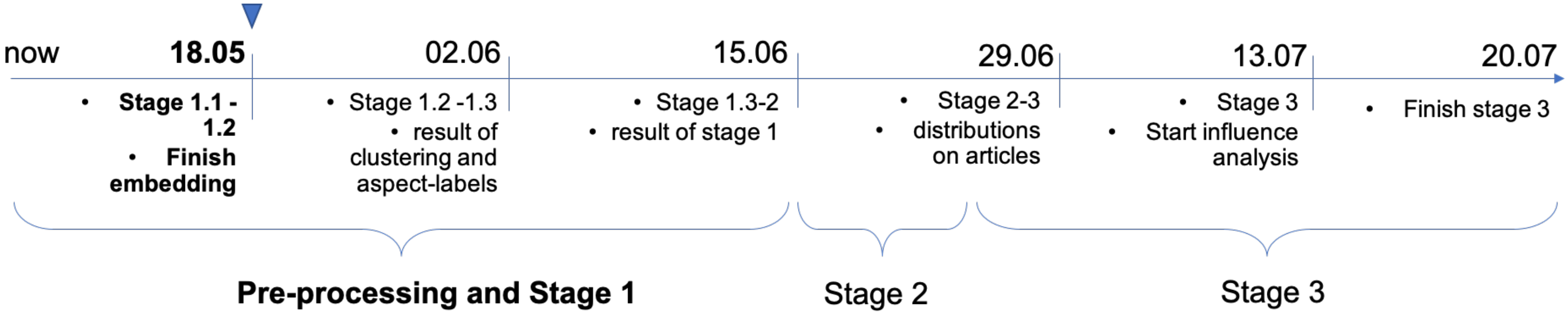
Opinion Lab Group 2.3

Wing Sheung Leung, Qiaoxi Liu

June 1, 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

1.2 Kmeans and Elbow Method

sklearn.cluster.MiniBatchKMeans
Elbow Method for determining optimal k
AIC for determining optimal k
BIC for determining optimal k
Determination of suitable k by looking at top n words in each potential clusters
Clustering results
Clustering wordclouds

Future Plan

Stage 1.2: sklearn.cluster.MiniBatchKMeans

```
class KMeansClustering():
def __init__(self, k, X, is_mini_batch = True, plot_bar_chart = True):
  self.k = k
  self.X = np.array(X).reshape(len(X), 512)
  self.km = MiniBatchKMeans(n_clusters=k, init=’k-means++’, batch_size=3000, compute_labels=True).fit(self.X)
```

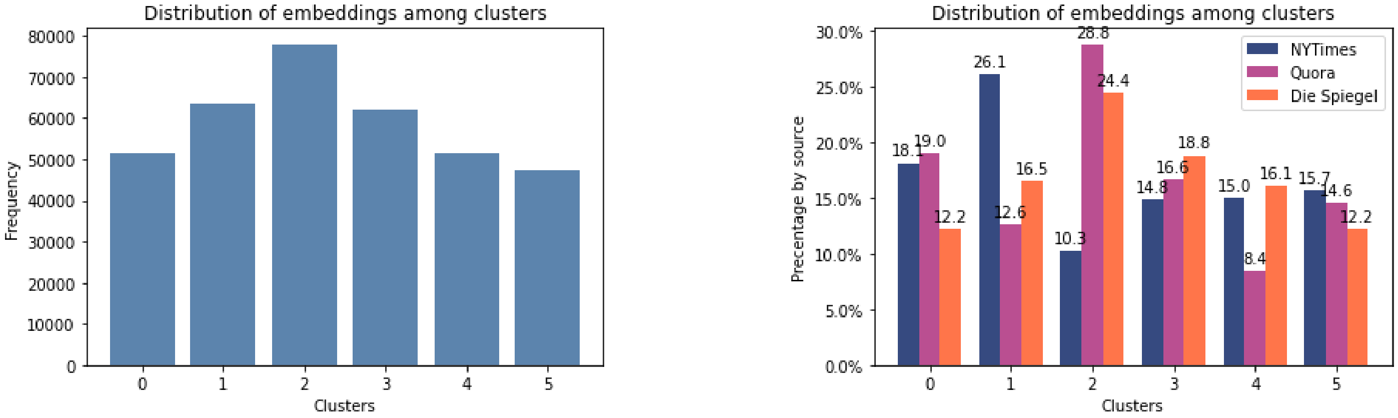
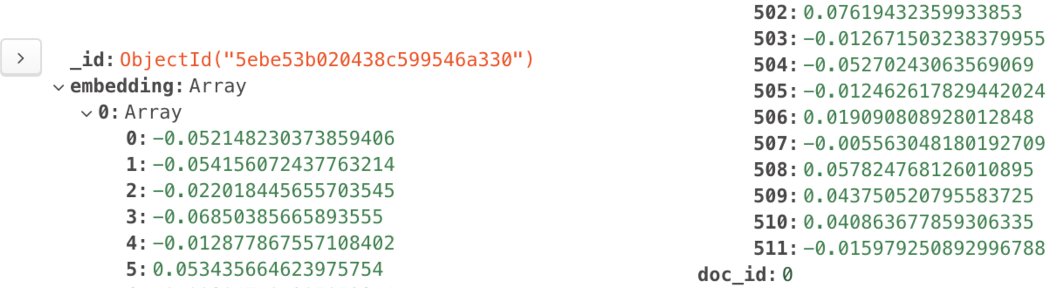


Figure: Example of distribution of embeddings of all tokenized sentences from the three sources among 6 clusters

Stage 1.2: sklearn.cluster.MiniBatchKMeans

Why MiniBatchKmeans instead of original sklearn.cluster.KMeans

XLING sentence level embeddings is generated in 512 dimensions for each tokenized sentence by NLTK.



```
_id: ObjectId("5ebe53b020438c599546a330")
embedding: Array
  0: Array
    0: -0.052148230373859406
    1: -0.054156072437763214
    2: -0.022018445655703545
    3: -0.06850385665893555
    4: -0.012877867557108402
    5: 0.053435664623975754
```

```
    502: 0.07619432359933853
    503: -0.012671503238379955
    504: -0.05270243063569069
    505: -0.012462617829442024
    506: 0.019090808928012848
    507: -0.005563048180192709
    508: 0.057824768126010895
    509: 0.043750520795583725
    510: 0.040863677859306335
    511: -0.015979250892996788
doc_id: 0
```

Figure: XLING embedding output for a sample sentence. Left: First 6 dimensions. Right: Last 10 dimensions

Source	Embedding JSON size	Original corpus size
New York Times	827 MB	55.9 MB
Quora	638 MB	15.9 MB
Die Speigel	2.3 GB	131 MB

Table: Embeddings generated are greatly larger then the original corpus size

Stage 1.2: Elbow Method for determining optimal k

```
K = range(2, 21)
for k in K:
  model = KMeansClustering(k, X)
  distortions.append(model.km.inertia_)
```

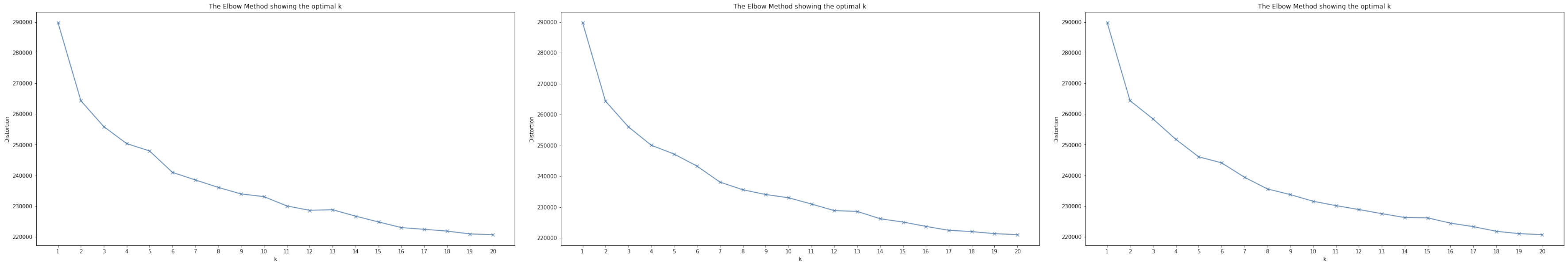


Figure: After few epochs of K-Means clustering, there is no distinguishable elbow of the curve for determination of optimal k

Stage 1.2: sklearn.cluster.MiniBatchKMeans

Why MiniBatchKmeans instead of original sklearn.cluster.KMeans

Just loading all sentence embeddings in Google Colaboratory, 6.36 GB out of the given 12.72 GB RAM had already been used up.

MiniBatchKMeans is faster and helps to prevent the session from crushing, however, gives slightly different results.

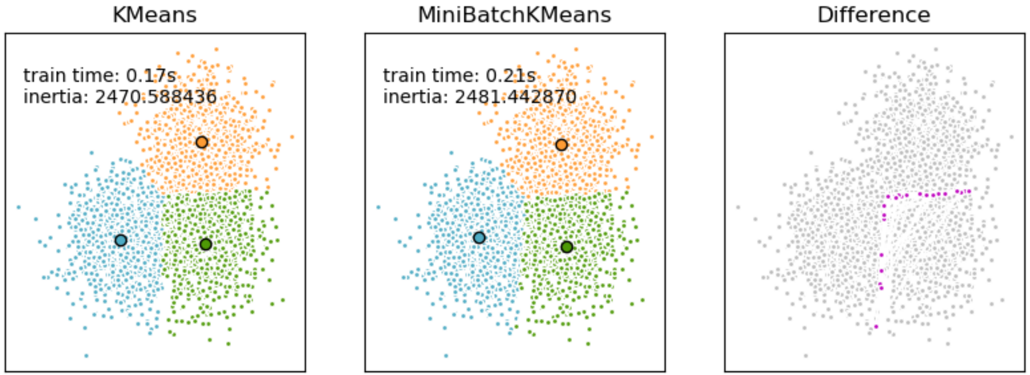


Figure: Extracted from scikit-learn; Data points classified differently are shown as purple points in ’Difference’ block

https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html

Stage 1.2: AIC for determining optimal k

```
def get_AIC(self):
    k, m = self.km.cluster_centers_.shape # dimension of centroids
    D = self.km.inertia_ # within-cluster sum of square distances, residual sum of squares
    AIC = D + 2 * m * k
    return AIC
```

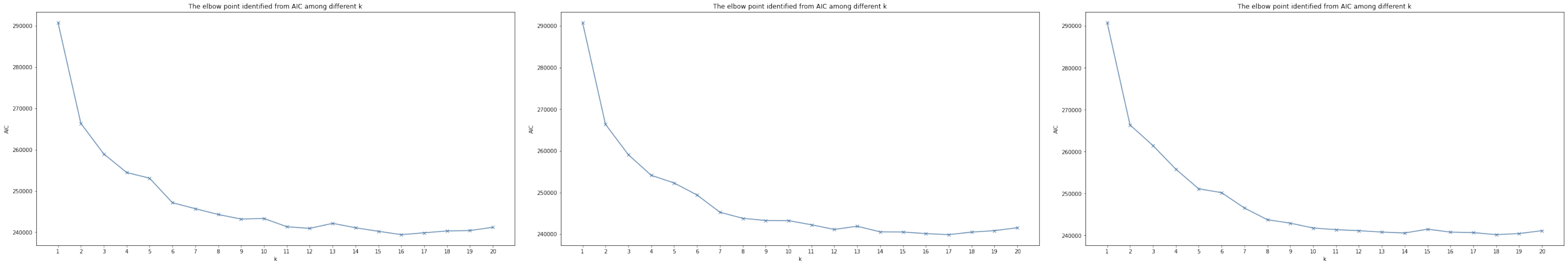


Figure: For those epochs, AIC curves more or less follow the trends in elbow method, but it is more feasible to see that the curves become more steady from k larger than 7 or 8.

Stage 1.2: BIC for determining optimal k

```
def get_BIC(self):
  k, m = self.km.cluster_centers_.shape # dimension of centroids
  n = self.n
  D = self.km.inertia_ # within-cluster sum of square distances, residual sum of squares
  BIC = D + 0.5 * m * k * np.log(n)
  return BIC
```

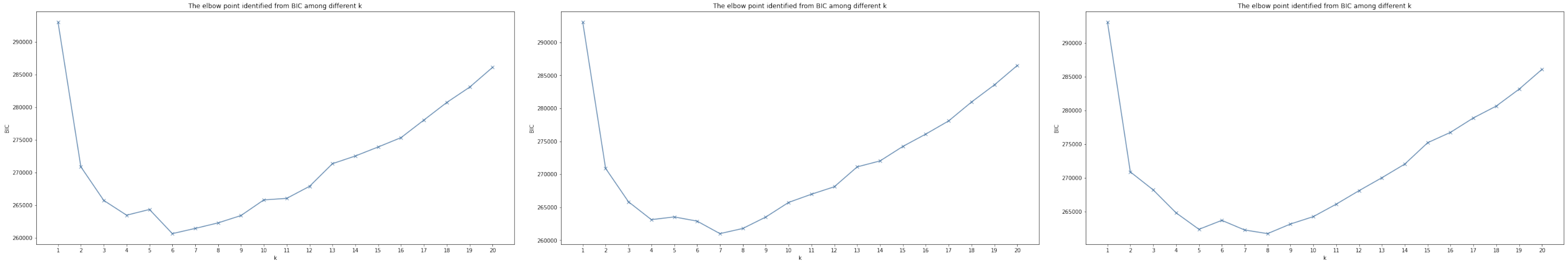


Figure: For those epochs, 5 to 8 are the potential candidates for optimal k

Stage 1.2: Determination of suitable k by looking at top n words in each potential clusters

Clustering results (different k)

Under each k, there is a cluster without NYtimes sentence.

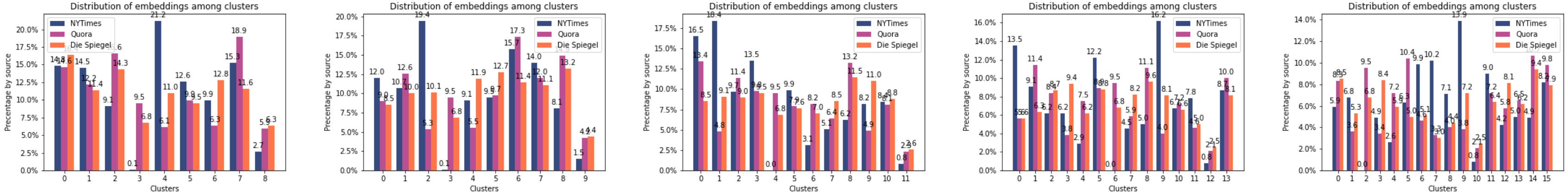
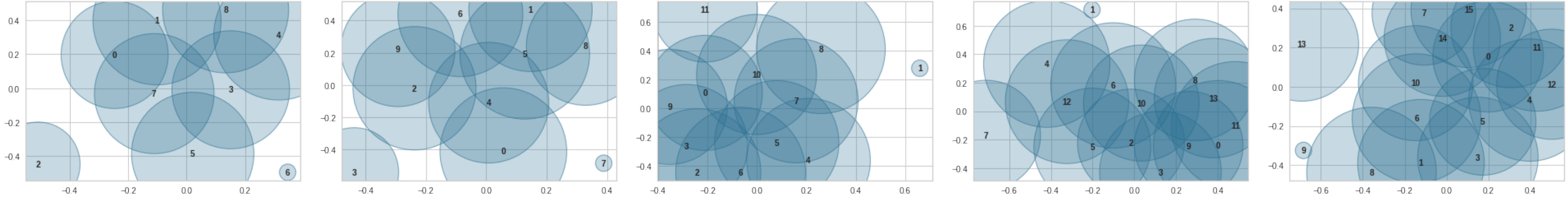


Figure: distribution for k=9, 10, 12, 14, 16



tokenizing, stemming and stopwords

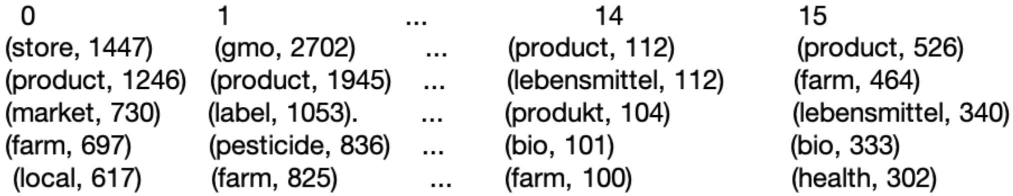
1. tokenizer chosen from ...
2. stemming : (using existing libraries) customized list for both languages

for example: farmer/farming/farm/ product/produce/production/

3. stopwords : also using libararies + customized list for both languages

Top words (frequencies / tfidf)

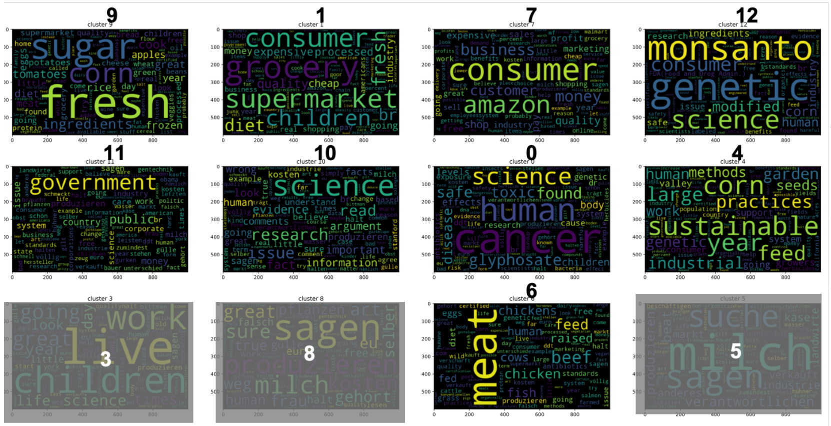
strategy 1: delete the repeated one which appear more than a ratio (0.5)

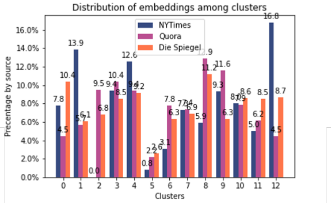


0	1	...	14	15
(store, 1447)	(gmo, 2702)	...	(product, 112)	(product, 526)
(product, 1246)	(product, 1945)	...	(lebensmittel, 112)	(farm, 464)
(market, 730)	(label, 1053).	...	(produkt, 104)	(lebensmittel, 340)
(farm, 697)	(pesticide, 836)	...	(bio, 101)	(bio, 333)
(local, 617)	(farm, 825)	...	(farm, 100)	(health, 302)

strategy 2: clarity scores

Wordclouds k = 13





0: human disease
The long term effects of accumulated pesticide exposure may well include more dementia, cancer, immune disorders, and other chronic conditions

1: Lifestyle and Economy
It is hard to get people to eat healthy foods, when the profits are with the junk food products that can be sold to consumers with massive advertising.

2.Garbage (Hallo,./Eben.)
3.Garbage (I love this blog. /Können Sie das?.)

4.Farmer farming
The opportunity for confusion is of enormous concern to many farmers in the New York region

5:Garbage
Zitat von Habenichts./Achja:.

6.Meat consumption
Manure produced by organically raised animals wreaks less havoc on the environment, but the meat may still wreak havoc on arteries.

7. Retailers and brands
Obst und Gemüse aber nicht aus dem Supermarkt und schon garnicht von den Dicountern.

8.Garbage (Zitat von MarkH.' /'Please!')

9. Food quality and nutritions
'Darum leidet auch die Qualität was sie zu reinen Konsumartikeln und nicht zu Lebensmitteln macht.'

10. Evidence
This article blows my mind in it is lack of research. 'The whole argument is mislabeled.'

11. Politic
Sollen wir uns noch mehr von dieser korrupten Regierung, und diesem meiner Meinung nach fiesen Staatsorgan gefallen lassen.'

12: Chemicals, GMO
Imagine my surprise when I learned that organic farming actually does allow pesticide use, and the pesticides,…

Next steps...

1. confirm the ideal k
2. sentiment classification