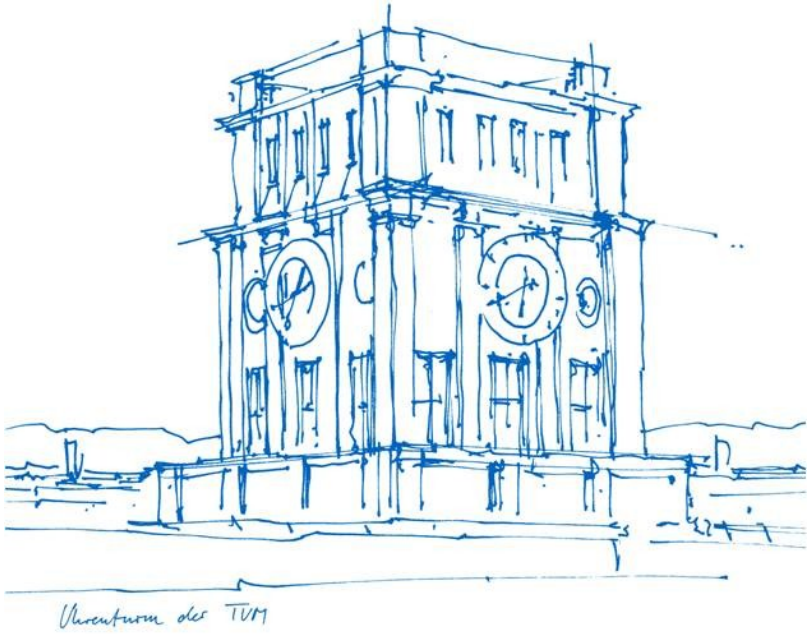
Social Topic Distributions

Mürüvvet Hasanbaşoğlu
Hakan Akyürek

Technical University of Munich

Faculty of Informatics

Munich, 20. July 2020



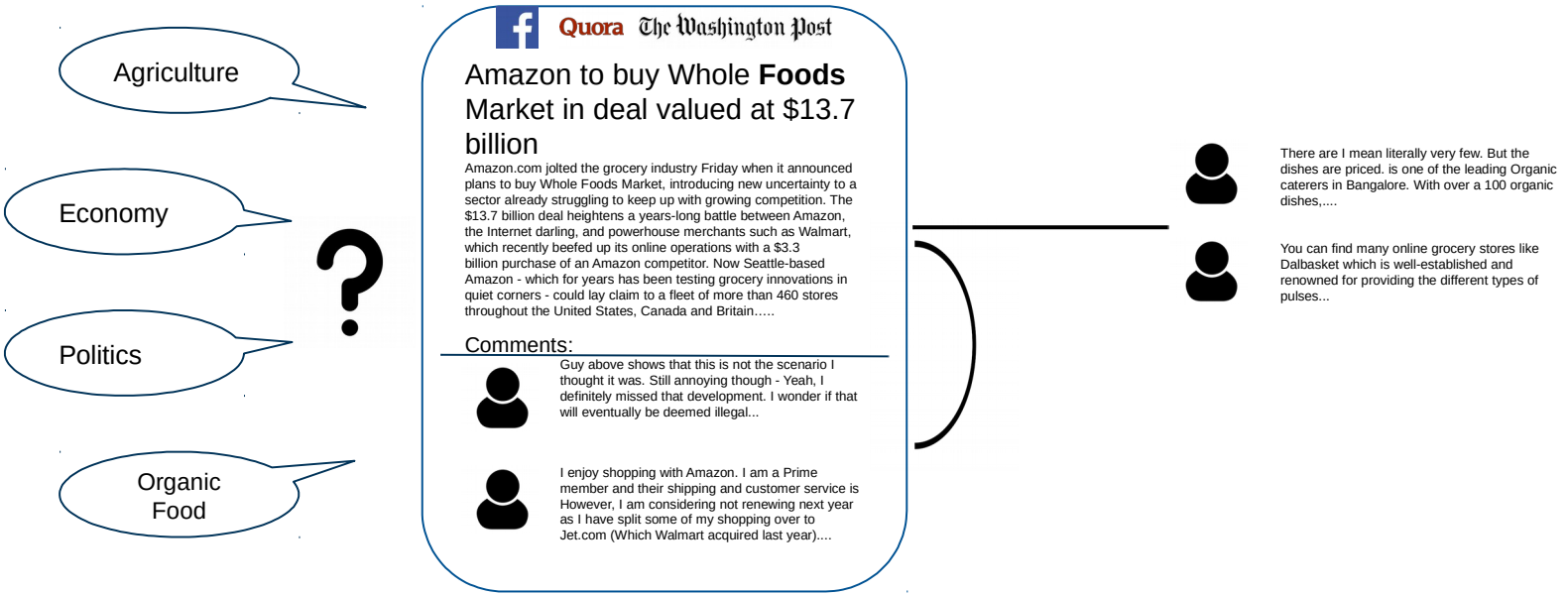
Outline

- Motivation
- Data
- Experiments
- Results

Outline

- **Motivation**
- Data
- Experiments
- Results

Motivation


Agriculture
Economy
Politics
Organic Food
?
Quora
The Washington Post
Amazon to buy Whole Foods Market in deal valued at $13.7 billion
Amazon.com jolted the grocery industry Friday when it announced plans to buy Whole Foods Market, introducing new uncertainty to a sector already struggling to keep up with growing competition. The $13.7 billion deal heightens a years-long battle between Amazon, the Internet darling, and powerhouse merchants such as Walmart, which recently beefed up its online operations with a $3.3 billion purchase of an Amazon competitor. Now Seattle-based Amazon - which for years has been testing grocery innovations in quiet corners - could lay claim to a fleet of more than 460 stores throughout the United States, Canada and Britain.....
Comments:
Guy above shows that this is not the scenario I thought it was. Still annoying though - Yeah, I definitely missed that development. I wonder if that will eventually be deemed illegal...
I enjoy shopping with Amazon. I am a Prime member and their shipping and customer service is However, I am considering not renewing next year as I have split some of my shopping over to Jet.com (Which Walmart acquired last year)....
There are I mean literally very few. But the dishes are priced. is one of the leading Organic caterers in Bangalore. With over a 100 organic dishes,....
You can find many online grocery stores like Dalbasket which is well-established and renowned for providing the different types of pulses...

Outline

- Motivation
- **Data**
- Experiments
- Results

Data

Curated Organic Dataset			# of articles	# of users
Biased		Facebook	5,013	4,705
		Food Babe	15	15
		Food Revolution	78	60
		Organic Authority	66	0
		Organic Consumers	64	0
Unbiased	Forum	Cafe Mom	86	85
		Disqus	36	36
		Quora	567	523
		Reddit	81	78
		US Message Board	0	0
	News sites	Chicago Tribune	2,283	78
		Huffington Post	880	0
		LA Times	1,522	77
		NY Post	106	0
		NY Times	438	137
		USA Today	95	22
		Washington Post	1,563	943

12,893 164,387

1. Tokenization: `r"[a-zA-Z0-9]+|\.|\?|\!"`
2. Lowercase
3. Stop word removal
4. Rare word removal

Tokens	Vocabulary
~19M	88,119

Data

Curated Organic Dataset			# of articles	# of users
Biased		Facebook	5,013	4,705
		Food Babe	15	15
		Food Revolution	78	60
		Organic Authority	66	0
		Organic Consumers	64	0
Unbiased	Forum	Cafe Mom	86	85
		Disqus	36	36
		Quora	567	523
		Reddit	81	78
		US Message Board	0	0
	News sites	Chicago Tribune	2,283	78
		Huffington Post	880	0
		LA Times	1,522	77
		NY Post	106	0
		NY Times	438	137
		USA Today	95	22
		Washington Post	1,563	943

12,893 164,387

1. Tokenization: `r"[a-zA-Z0-9]+|\.|\?|\!"`
2. Lowercase
3. Stop word removal
4. Rare word removal

Tokens	Vocabulary
~19M	88,119

Word embeddings ➜ GloVe

Sentence embeddings ➜ Universal Sentence Encoder TensorFlow Hub

Data

Curated Organic Dataset			# of articles	# of users
Biased		Facebook	5,013	4,705
		Food Babe	15	15
		Food Revolution	78	60
		Organic Authority	66	0
		Organic Consumers	64	0
Unbiased	Forum	Cafe Mom	86	85
		Disqus	36	36
		Quora	567	523
		Reddit	81	78
		US Message Board	0	0
	News sites	Chicago Tribune	2,283	78
		Huffington Post	880	0
		LA Times	1,522	77
		NY Post	106	0
		NY Times	438	137
		USA Today	95	22
		Washington Post	1,563	943

Data

Curated Organic Dataset			# of articles	# of users	
Biased		Facebook	5,013	4,705	➜ 150 Articles
		Food Babe	15	15	
		Food Revolution	78	60	
		Organic Authority	66	0	
		Organic Consumers	64	0	
Unbiased	Forum	Cafe Mom	86	85	
		Disqus	36	36	
		Quora	567	523	➜ 150 Articles
		Reddit	81	78	
		US Message Board	0	0	
	News sites	Chicago Tribune	2,283	78	
		Huffington Post	880	0	
		LA Times	1,522	77	
		NY Post	106	0	
		NY Times	438	137	
		USA Today	95	22	
		Washington Post	1,563	943	➜ 150 Articles

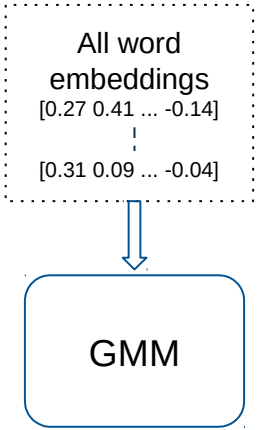
User datasets:

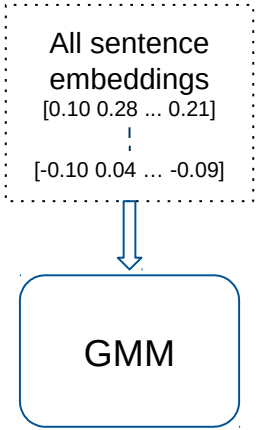
- Answering Users
- Random Users:
  1. General
  2. Biased
  3. Unbiased
  4. Forum
  5. News sites

Outline

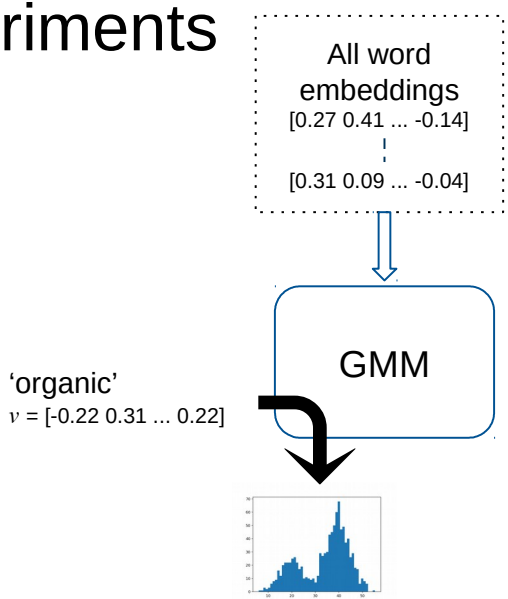
- Motivation
- Data
- **Experiments**
- Results

Experiments

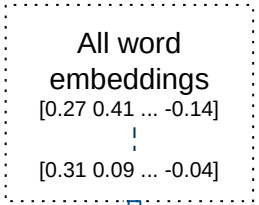


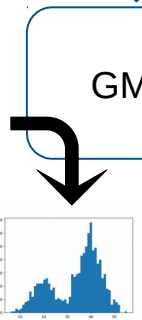


Experiments



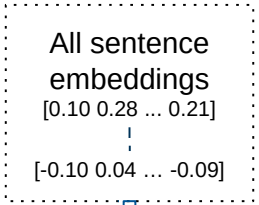
Experiments

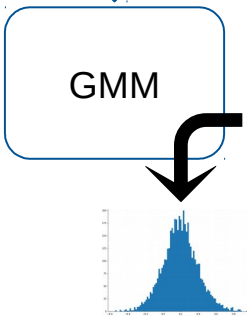




‘organic’
v = [-0.22 0.31 ... 0.22]

$$P(d_l) = \frac{CF * IDF(c_i, d_l, D)}{\sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}} = \frac{\sum_{j=1}^{k} v_{ij} * log\frac{|D|}{|d \in D; c_i \in d|}}{\sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}}$$



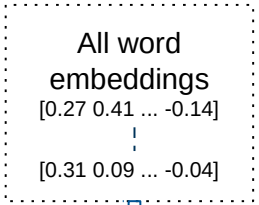


‘enjoy shopping amazon ?’
[0.22 0.26 ... 0.25] = v

$$P(d_l) = \frac{\sum_{j=1}^{k} v_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}}$$

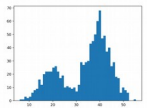
l : # of documents
k : # of words / sentences in a document
n : # of topic clusters

Experiments

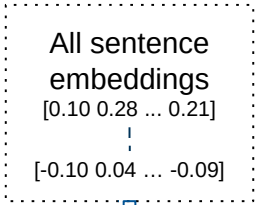


‘organic’
v = [-0.22 0.31 ... 0.22]

GMM

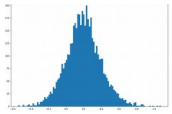


$$P(d_l) = \frac{CF * IDF(c_i, d_l, D)}{\sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}} = \frac{\sum_{j=1}^{k} v_{ij} * log\frac{|D|}{|d \in D; c_i \in d|}}{\sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}}$$



GMM

‘enjoy shopping amazon ?’
[0.22 0.26 ... 0.25] = v



$$P(d_l) = \frac{\sum_{j=1}^{k} v_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}}$$

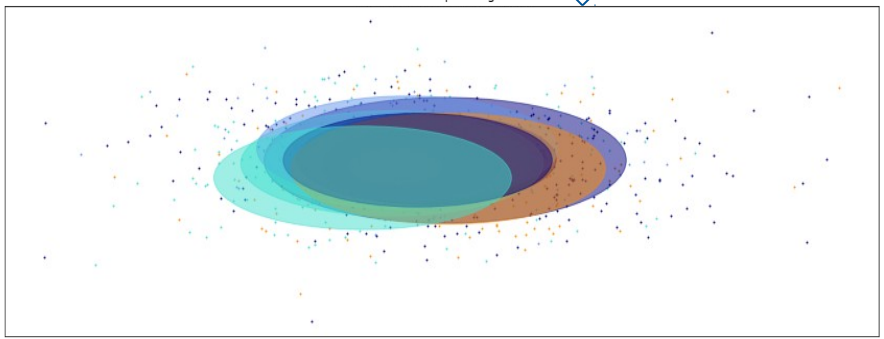
l : # of documents
k : # of words / sentences in a document
n : # of topic clusters

?

Experiments

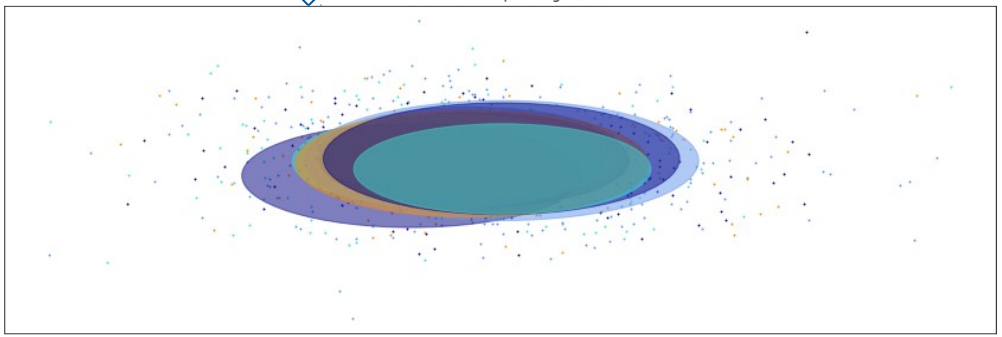
All word embeddings
[0.27 0.41 ... -0.14]
[0.31 0.09 ... -0.04]

10 cluster sample diag GMM



All sentence embeddings
[0.10 0.28 ... 0.21]
[-0.10 0.04 ... -0.09]

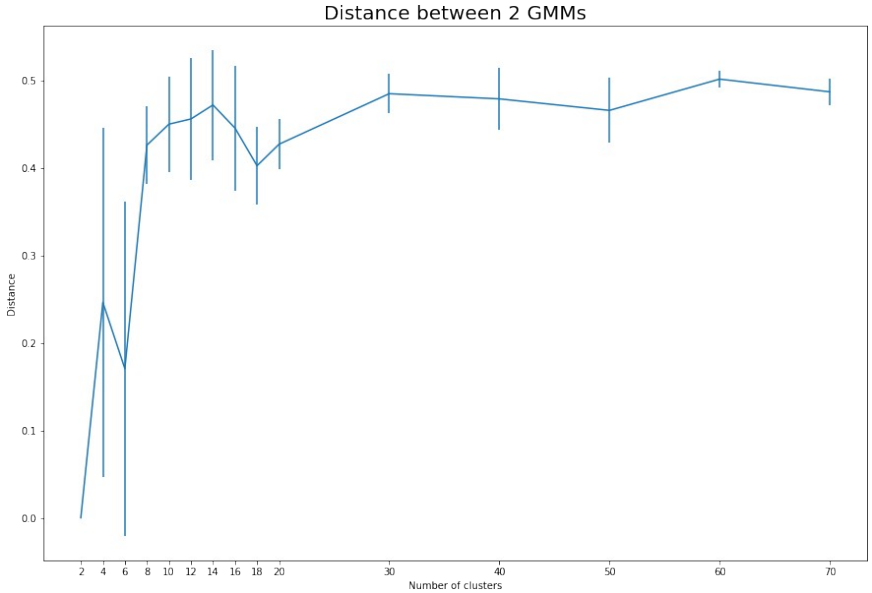
10 cluster sample diag GMM



l : # of documents
k : # of words / sentences in a document
n : # of topic clusters ?

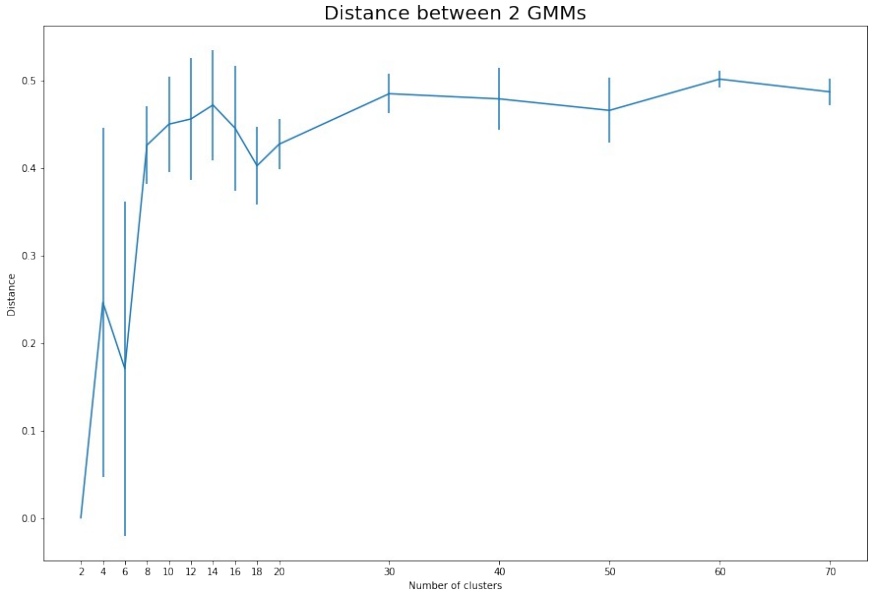
Experiments

Optimal number of clusters for word embeddings:



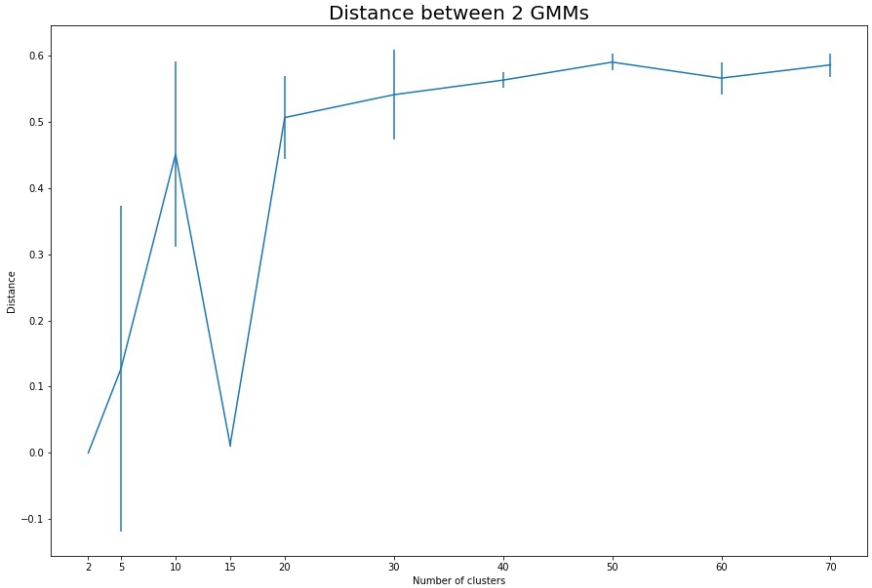
Experiments

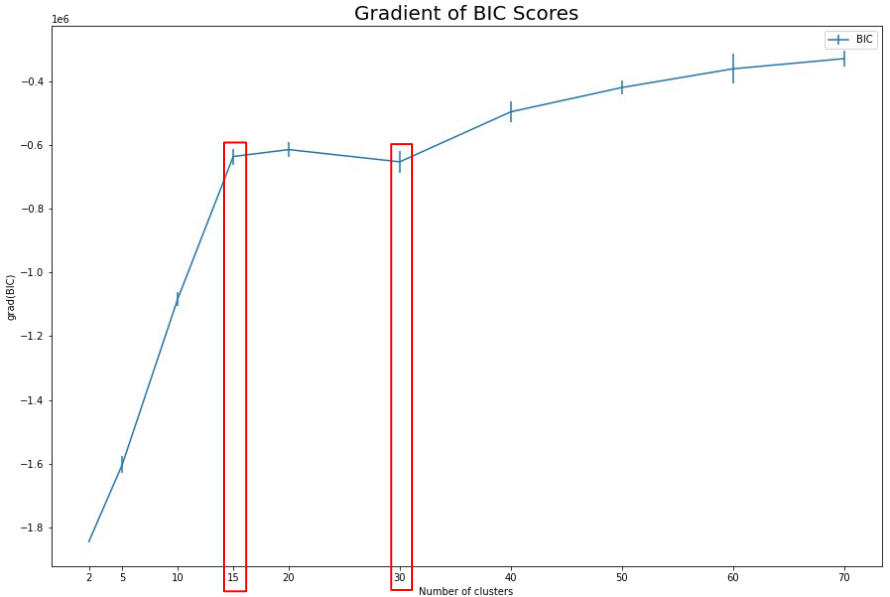
Optimal number of clusters for word embeddings:



Experiments

Optimal number of clusters for sentence embeddings:

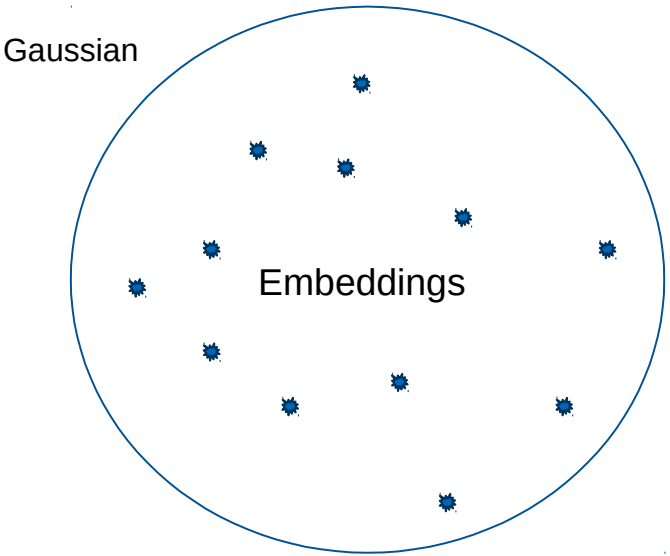




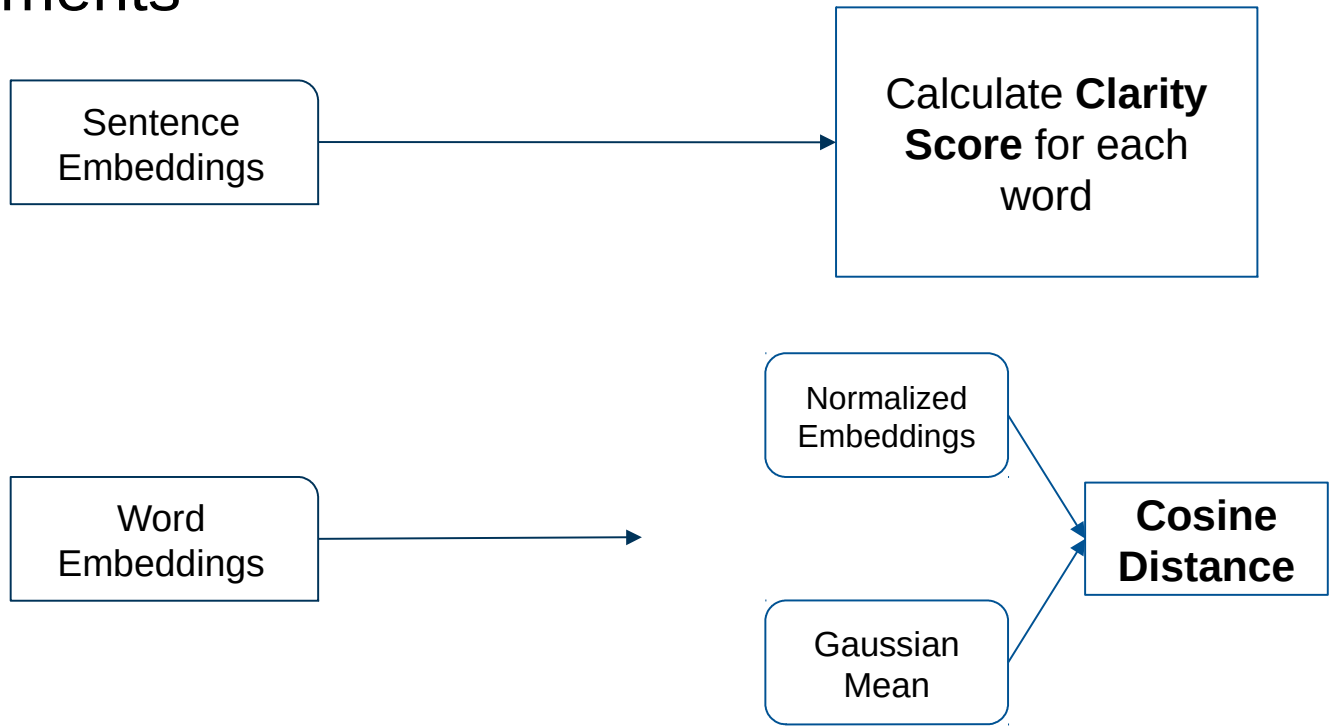
Experiments

How do we label Gaussians in a GMM model?

- Find the most representative words
- Manually select a topic



Experiments



Outline

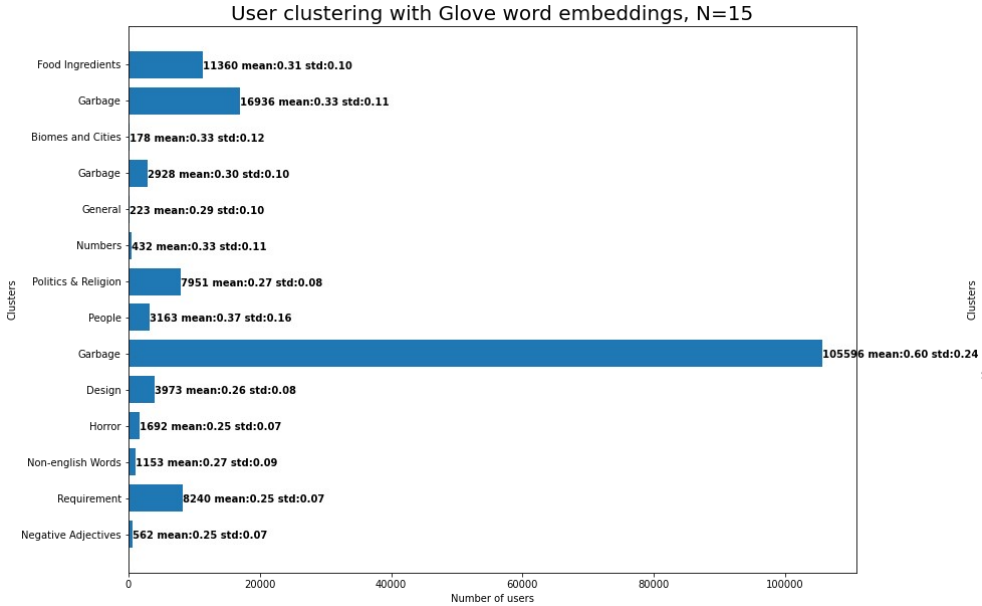
- Motivation
- Data
- Experiments
- **Results**

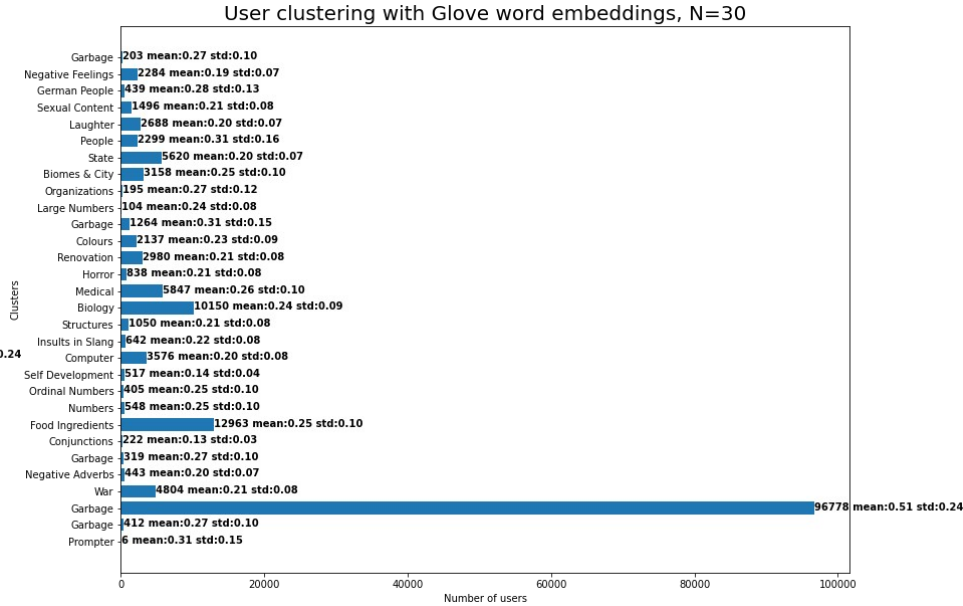
Results

How the users are distributed?

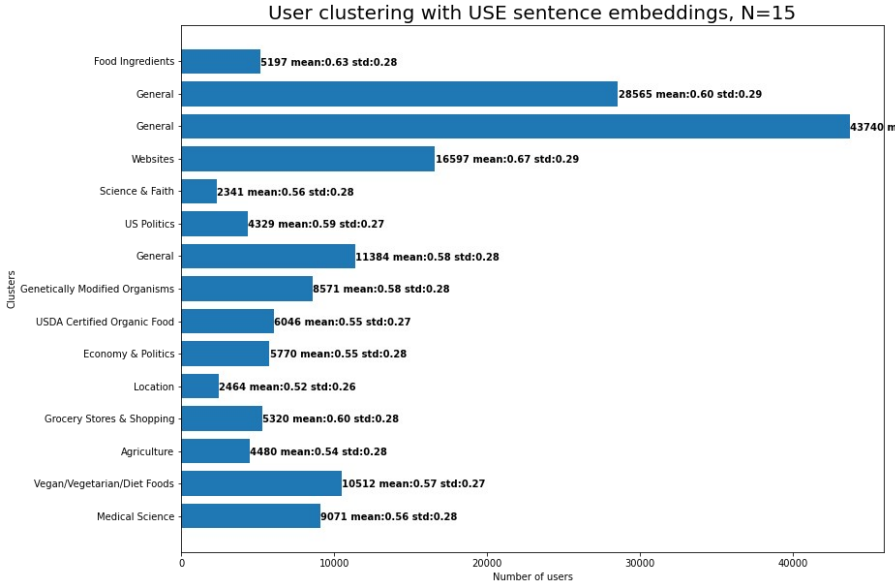
- in Word embedding space
- in Sentence embedding space

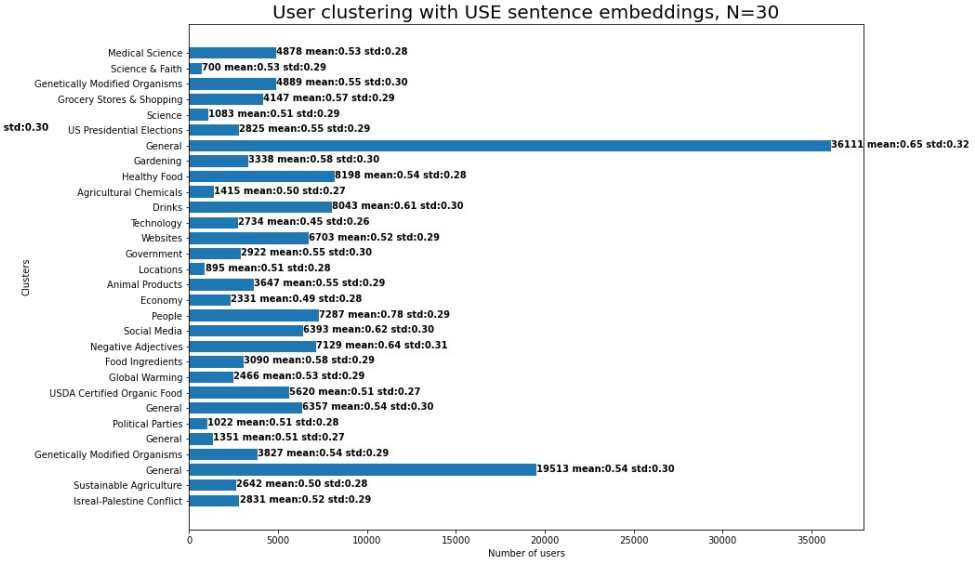
Results


User clustering with Glove word embeddings, N=15
Food Ingredients
Garbage
Biomes and Cities
Garbage
General
Numbers
Politics & Religion
People
Garbage
Design
Horror
Non-english Words
Requirement
Negative Adjectives
11360 mean:0.31 std:0.10
16936 mean:0.33 std:0.11
178 mean:0.33 std:0.12
2928 mean:0.30 std:0.10
223 mean:0.29 std:0.10
432 mean:0.33 std:0.11
7951 mean:0.27 std:0.08
3163 mean:0.37 std:0.16
105596 mean:0.60 std:0.24
3973 mean:0.26 std:0.08
1692 mean:0.25 std:0.07
1153 mean:0.27 std:0.09
8240 mean:0.25 std:0.07
562 mean:0.25 std:0.07
Clusters
0
20000
40000
60000
80000
100000
Number of users


User clustering with Glove word embeddings, N=30
Garbage
Negative Feelings
German People
Sexual Content
Laughter
People
State
Biomes & City
Organizations
Large Numbers
Garbage
Colours
Renovation
Horror
Medical
Biology
Structures
Insults in Slang
Computer
Self Development
Ordinal Numbers
Numbers
Food Ingredients
Conjunctions
Garbage
Negative Adverbs
War
Garbage
Garbage
Prompter
203 mean:0.27 std:0.10
2284 mean:0.19 std:0.07
439 mean:0.28 std:0.13
1496 mean:0.21 std:0.08
2688 mean:0.20 std:0.07
2299 mean:0.31 std:0.16
5620 mean:0.20 std:0.07
3158 mean:0.25 std:0.10
195 mean:0.27 std:0.12
104 mean:0.24 std:0.08
1264 mean:0.31 std:0.15
2137 mean:0.23 std:0.09
2980 mean:0.21 std:0.08
838 mean:0.21 std:0.08
5847 mean:0.26 std:0.10
10150 mean:0.24 std:0.09
1050 mean:0.21 std:0.08
642 mean:0.22 std:0.08
3576 mean:0.20 std:0.08
517 mean:0.14 std:0.04
405 mean:0.25 std:0.10
548 mean:0.25 std:0.10
12963 mean:0.25 std:0.10
222 mean:0.13 std:0.03
319 mean:0.27 std:0.10
443 mean:0.20 std:0.07
4804 mean:0.21 std:0.08
96778 mean:0.51 std:0.24
412 mean:0.27 std:0.10
6 mean:0.31 std:0.15
Clusters
0
20000
40000
60000
80000
100000
Number of users

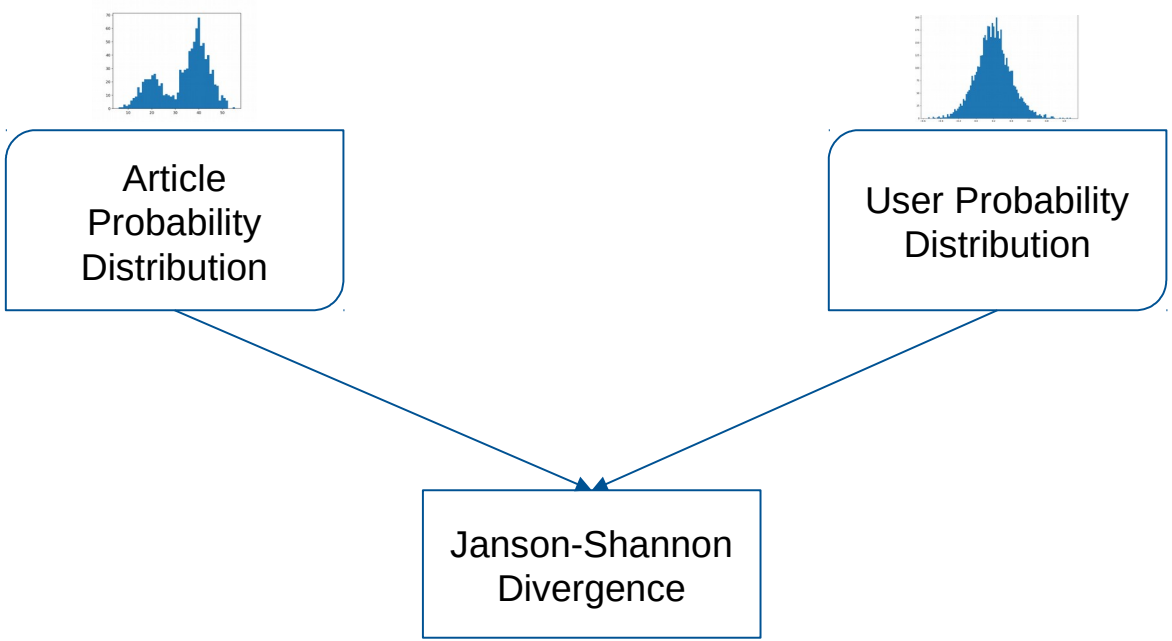
Results


User clustering with USE sentence embeddings, N=15
Clusters
Food Ingredients
General
General
Websites
Science & Faith
US Politics
General
Genetically Modified Organisms
USDA Certified Organic Food
Economy & Politics
Location
Grocery Stores & Shopping
Agriculture
Vegan/Vegetarian/Diet Foods
Medical Science
5197 mean:0.63 std:0.28
28565 mean:0.60 std:0.29
43740 mean:0.68 std:0.30
16597 mean:0.67 std:0.29
2341 mean:0.56 std:0.28
4329 mean:0.59 std:0.27
11384 mean:0.58 std:0.28
8571 mean:0.58 std:0.28
6046 mean:0.55 std:0.27
5770 mean:0.55 std:0.28
2464 mean:0.52 std:0.26
5320 mean:0.60 std:0.28
4480 mean:0.54 std:0.28
10512 mean:0.57 std:0.27
9071 mean:0.56 std:0.28
0
10000
20000
30000
40000
Number of users

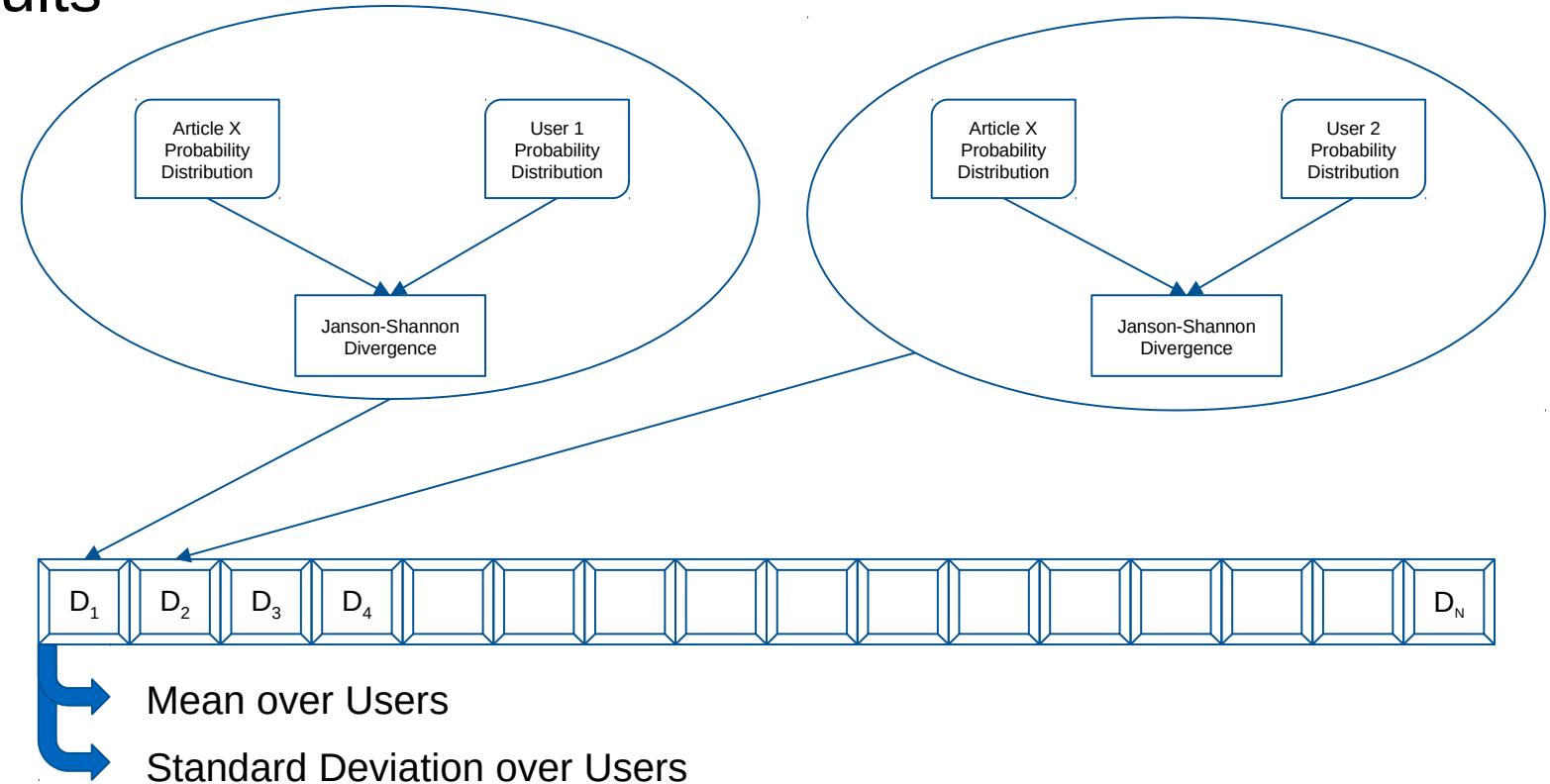

User clustering with USE sentence embeddings, N=30
Clusters
Medical Science
Science & Faith
Genetically Modified Organisms
Grocery Stores & Shopping
Science
US Presidential Elections
General
Gardening
Healthy Food
Agricultural Chemicals
Drinks
Technology
Websites
Government
Locations
Animal Products
Economy
People
Social Media
Negative Adjectives
Food Ingredients
Global Warming
USDA Certified Organic Food
General
Political Parties
General
Genetically Modified Organisms
General
Sustainable Agriculture
Isreal-Palestine Conflict
4878 mean:0.53 std:0.28
700 mean:0.53 std:0.29
4889 mean:0.55 std:0.30
4147 mean:0.57 std:0.29
1083 mean:0.51 std:0.29
2825 mean:0.55 std:0.29
36111 mean:0.65 std:0.32
3338 mean:0.58 std:0.30
8198 mean:0.54 std:0.28
1415 mean:0.50 std:0.27
8043 mean:0.61 std:0.30
2734 mean:0.45 std:0.26
6703 mean:0.52 std:0.29
2922 mean:0.55 std:0.30
895 mean:0.51 std:0.28
3647 mean:0.55 std:0.29
2331 mean:0.49 std:0.28
7287 mean:0.78 std:0.29
6393 mean:0.62 std:0.30
7129 mean:0.64 std:0.31
3090 mean:0.58 std:0.29
2466 mean:0.53 std:0.29
5620 mean:0.51 std:0.27
6357 mean:0.54 std:0.30
1022 mean:0.51 std:0.28
1351 mean:0.51 std:0.27
3827 mean:0.54 std:0.29
19513 mean:0.54 std:0.30
2642 mean:0.50 std:0.28
2831 mean:0.52 std:0.29
0
5000
10000
15000
20000
25000
30000
35000
Number of users

Results

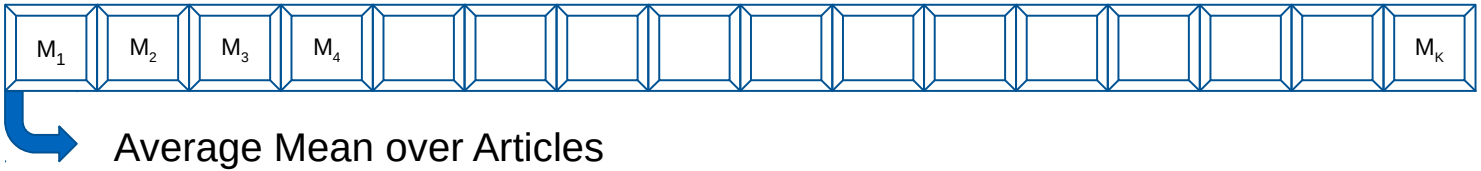
Hypothesis: Answering users should be closer to the article than random users.



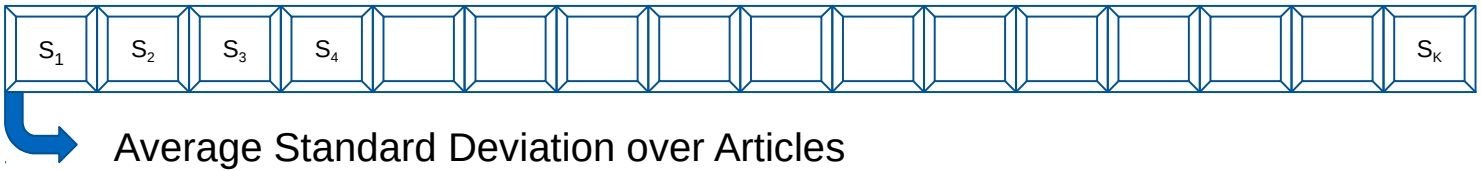
Results


Article X Probability Distribution
User 1 Probability Distribution
Janson-Shannon Divergence
Article X Probability Distribution
User 2 Probability Distribution
Janson-Shannon Divergence
For one Article
D_1
D_2
D_3
D_4
D_N
Mean over Users
Standard Deviation over Users

Results

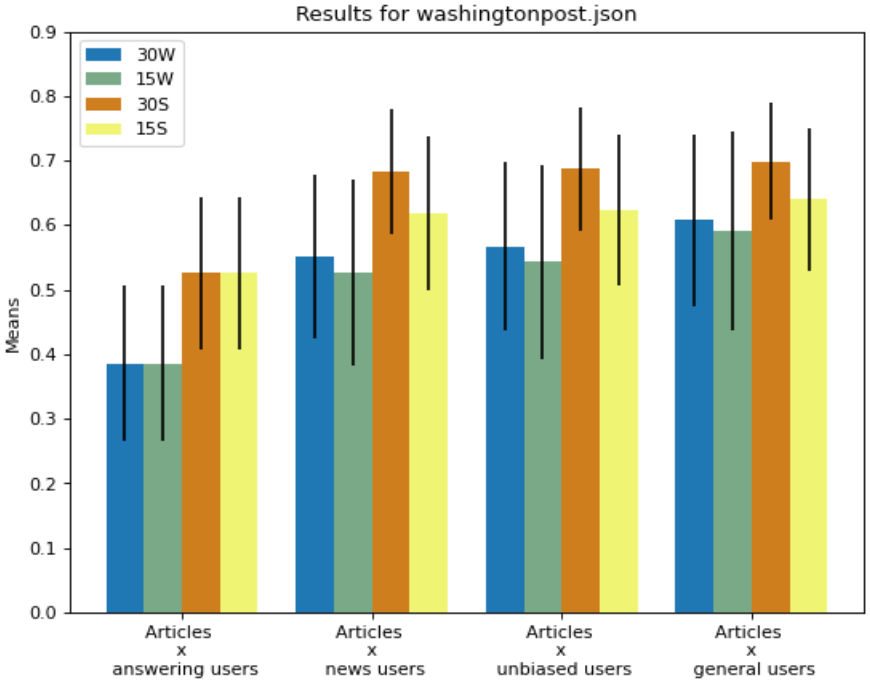


Average Mean over Articles

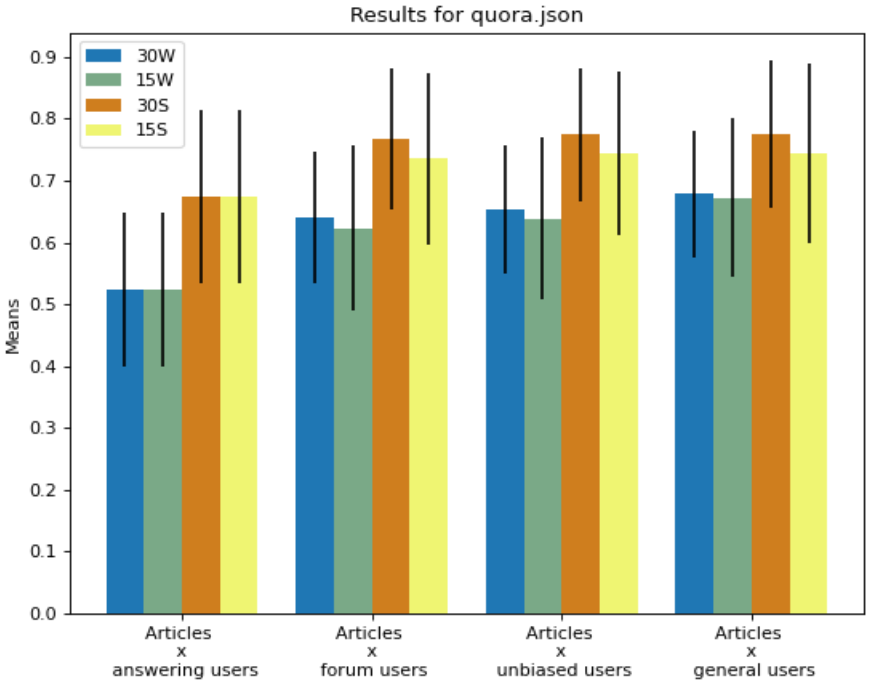


Average Standard Deviation over Articles

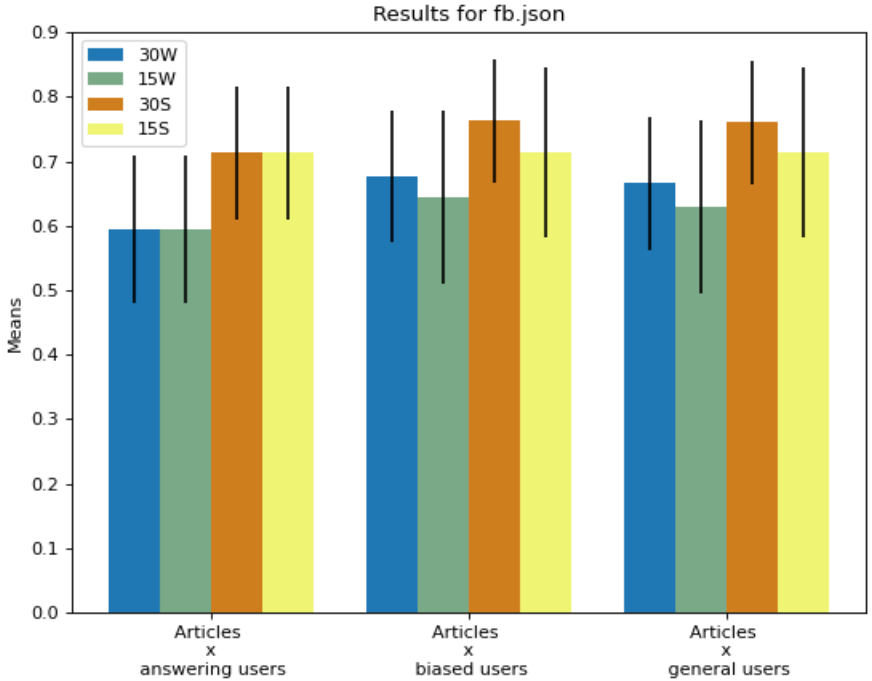
Results


Results for washingtonpost.json
30W
15W
30S
15S
Means
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Articles x answering users
Articles x news users
Articles x unbiased users
Articles x general users

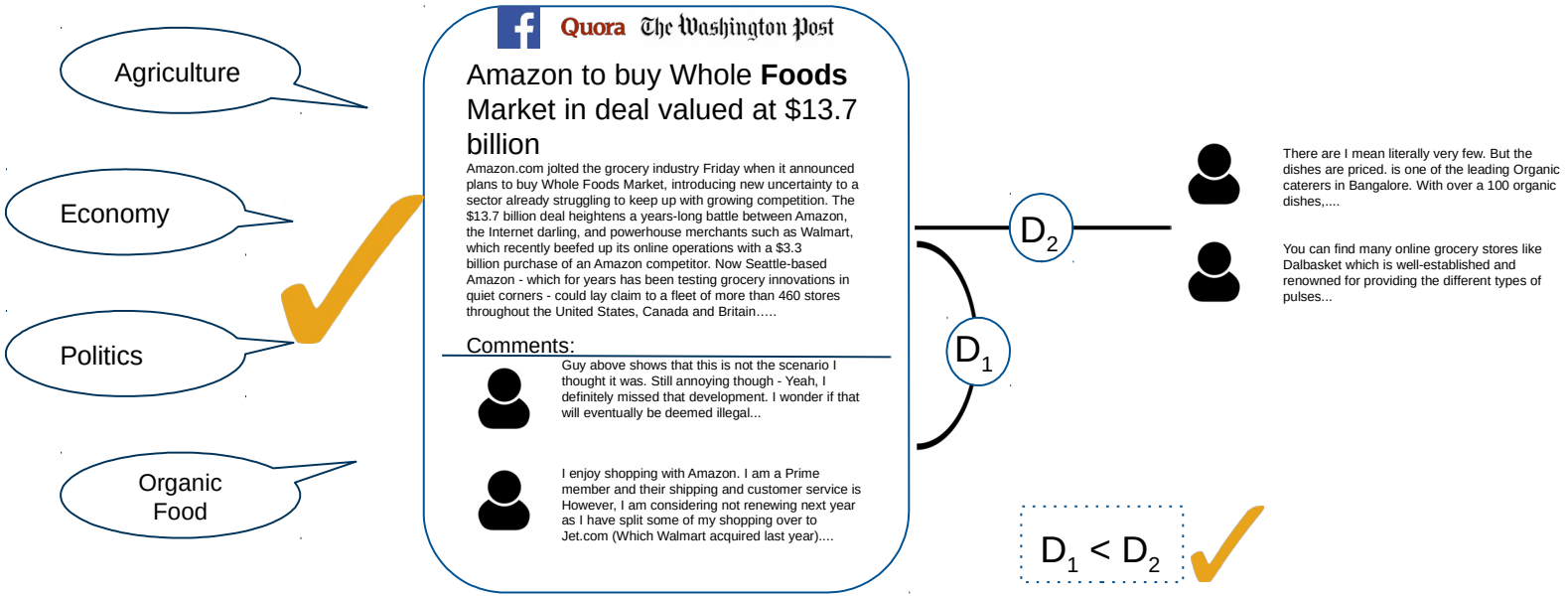
Results


Results for quora.json
30W
15W
30S
15S
Means
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Articles x answering users
Articles x forum users
Articles x unbiased users
Articles x general users

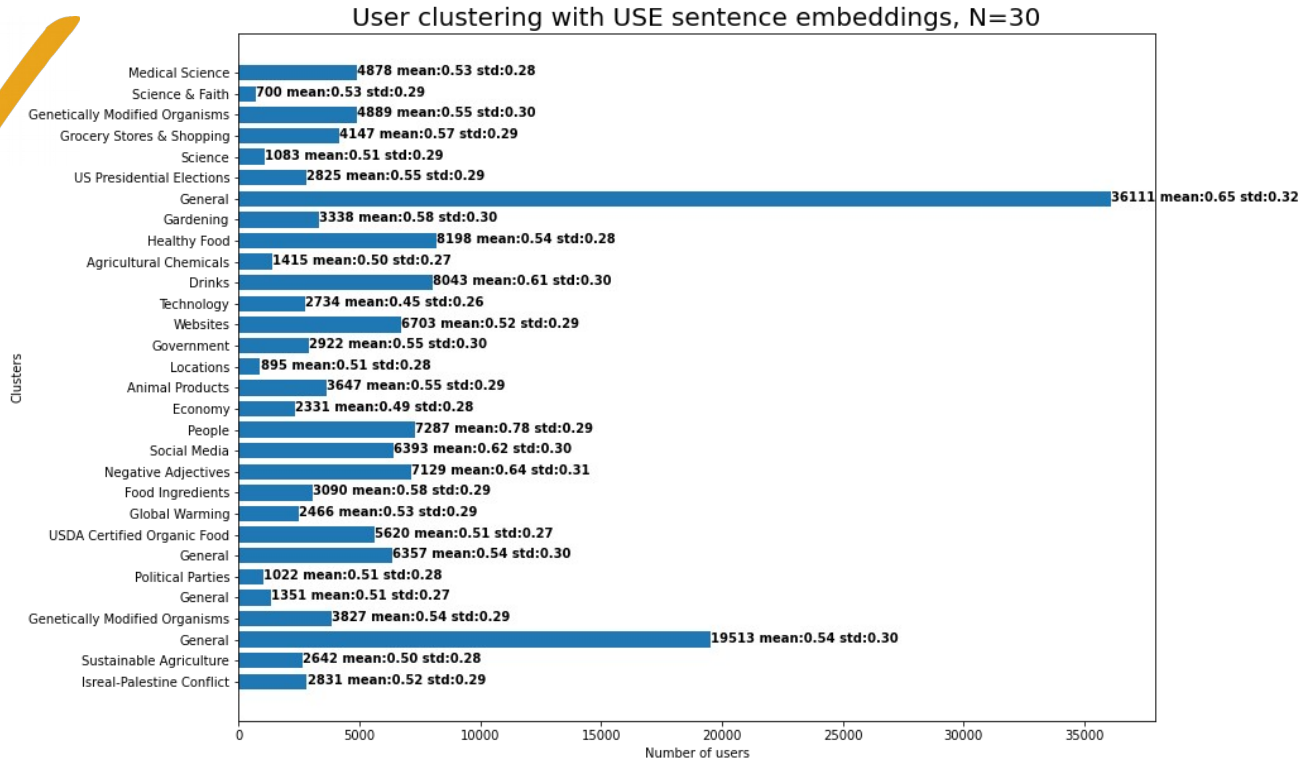
Results


Results for fb.json
30W
15W
30S
15S
Means
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Articles x answering users
Articles x biased users
Articles x general users

Results


Agriculture
Economy
Politics
Organic Food
Quora
The Washington Post
Amazon to buy Whole Foods Market in deal valued at $13.7 billion
Amazon.com jolted the grocery industry Friday when it announced plans to buy Whole Foods Market, introducing new uncertainty to a sector already struggling to keep up with growing competition. The $13.7 billion deal heightens a years-long battle between Amazon, the Internet darling, and powerhouse merchants such as Walmart, which recently beefed up its online operations with a $3.3 billion purchase of an Amazon competitor. Now Seattle-based Amazon - which for years has been testing grocery innovations in quiet corners - could lay claim to a fleet of more than 460 stores throughout the United States, Canada and Britain.....
Comments:
Guy above shows that this is not the scenario I thought it was. Still annoying though - Yeah, I definitely missed that development. I wonder if that will eventually be deemed illegal...
I enjoy shopping with Amazon. I am a Prime member and their shipping and customer service is However, I am considering not renewing next year as I have split some of my shopping over to Jet.com (Which Walmart acquired last year)....
D2
D1
There are I mean literally very few. But the dishes are priced. is one of the leading Organic caterers in Bangalore. With over a 100 organic dishes,....
You can find many online grocery stores like Dalbasket which is well-established and renowned for providing the different types of pulses...
D1 < D2

Results


User clustering with USE sentence embeddings, N=30
Clusters
Medical Science 4878 mean:0.53 std:0.28
Science & Faith 700 mean:0.53 std:0.29
Genetically Modified Organisms 4889 mean:0.55 std:0.30
Grocery Stores & Shopping 4147 mean:0.57 std:0.29
Science 1083 mean:0.51 std:0.29
US Presidential Elections 2825 mean:0.55 std:0.29
General 36111 mean:0.65 std:0.32
Gardening 3338 mean:0.58 std:0.30
Healthy Food 8198 mean:0.54 std:0.28
Agricultural Chemicals 1415 mean:0.50 std:0.27
Drinks 8043 mean:0.61 std:0.30
Technology 2734 mean:0.45 std:0.26
Websites 6703 mean:0.52 std:0.29
Government 2922 mean:0.55 std:0.30
Locations 895 mean:0.51 std:0.28
Animal Products 3647 mean:0.55 std:0.29
Economy 2331 mean:0.49 std:0.28
People 7287 mean:0.78 std:0.29
Social Media 6393 mean:0.62 std:0.30
Negative Adjectives 7129 mean:0.64 std:0.31
Food Ingredients 3090 mean:0.58 std:0.29
Global Warming 2466 mean:0.53 std:0.29
USDA Certified Organic Food 5620 mean:0.51 std:0.27
General 6357 mean:0.54 std:0.30
Political Parties 1022 mean:0.51 std:0.28
General 1351 mean:0.51 std:0.27
Genetically Modified Organisms 3827 mean:0.54 std:0.29
General 19513 mean:0.54 std:0.30
Sustainable Agriculture 2642 mean:0.50 std:0.28
Isreal-Palestine Conflict 2831 mean:0.52 std:0.29
0
5000
10000
15000
20000
25000
30000
35000
Number of users

Thanks for listening!

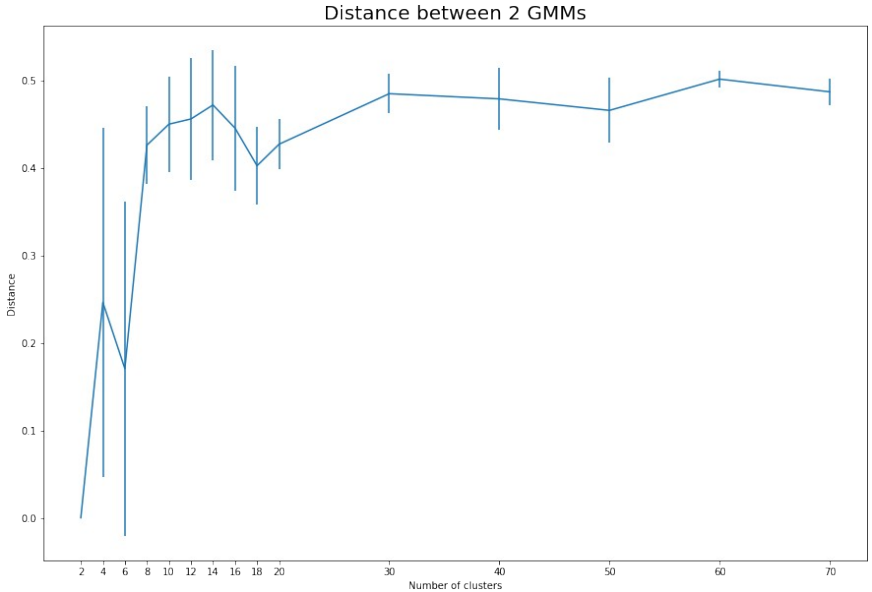
Questions **?**

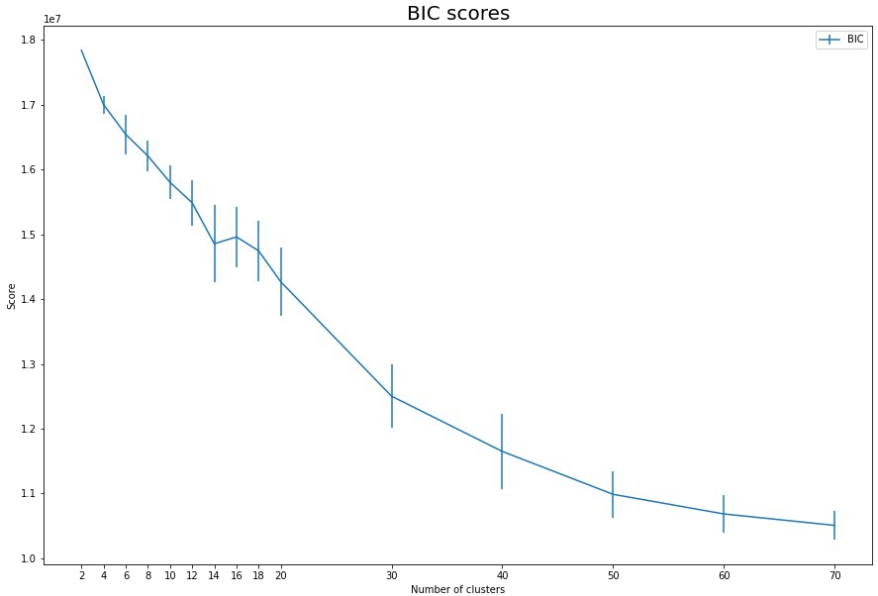
References

- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
- Sridhar, Vivek Kumar Rangarajan. "Unsupervised topic modeling for short texts using distributed representations of words." *Proceedings of the 1st workshop on vector space modeling for natural language processing*. 2015.
- Fine tune GloVe embeddings using Mittens
- A fast implementation of GloVe, with optional retrofitting
- Basics of using pre-trained glove vectors in python
- Sklearn Gaussian Mixture Model
- Gaussian Mixture Model Clusterization how to select the number of components clusters
- Kl divergence of two gmms
- How to find the similarity between two probability distributions using python
- Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Sung, Y. H. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Angelidis, Stefanos, and Mirella Lapata. "Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised." *arXiv preprint arXiv:1808.08858* (2018).
- Kim, Han Kyul, Hyunjoong Kim, and Sungzoon Cho. "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation." *Neurocomputing* 266 (2017): 336-352.

Experiments

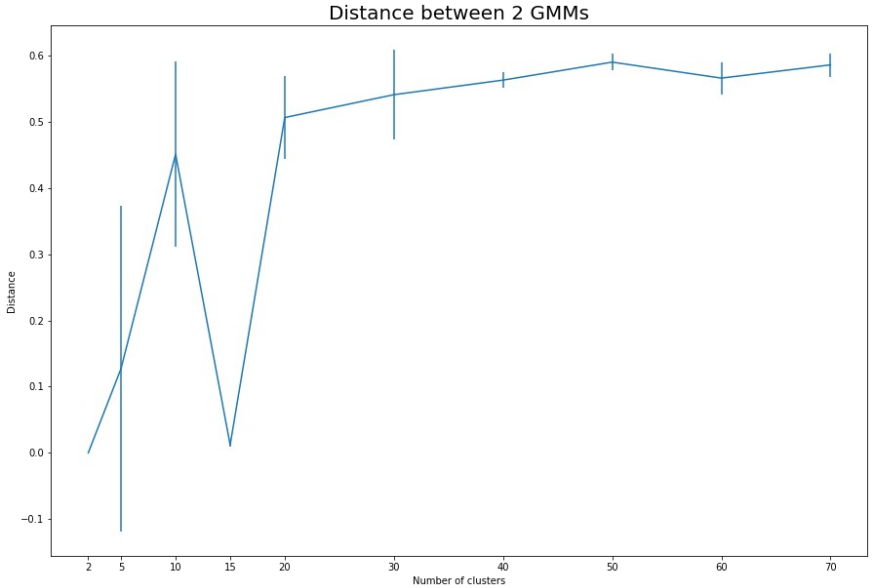
Optimal number of clusters for word embeddings:

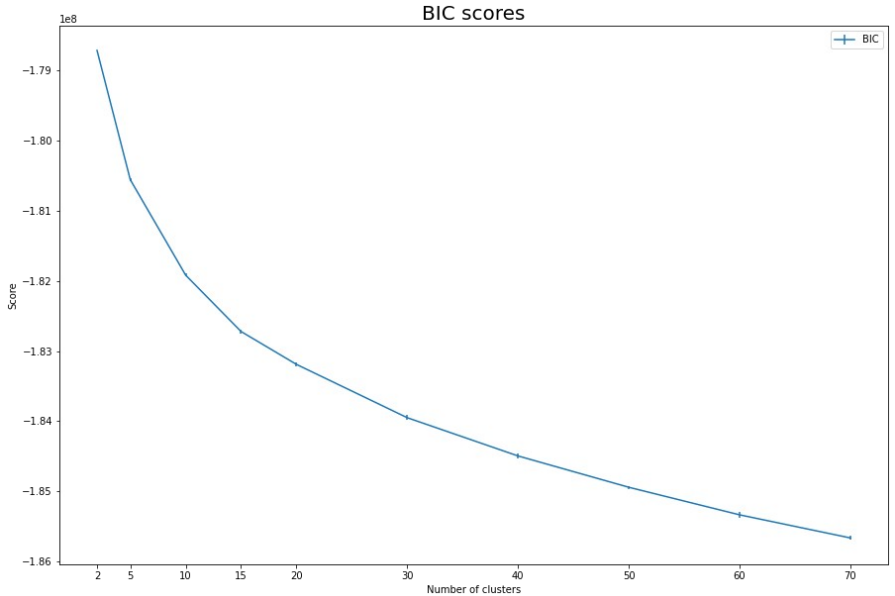




Experiments

Optimal number of clusters for sentence embeddings:





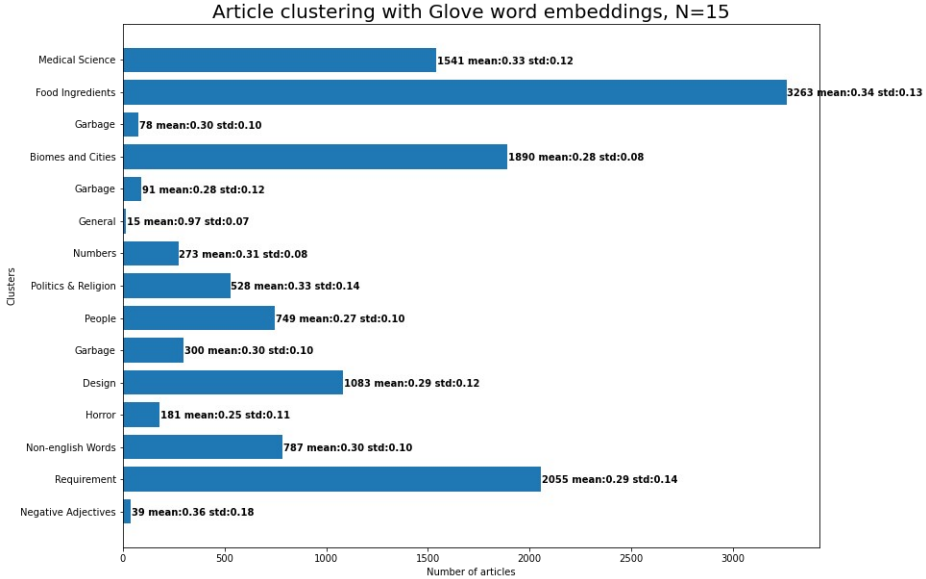
Clarity Score

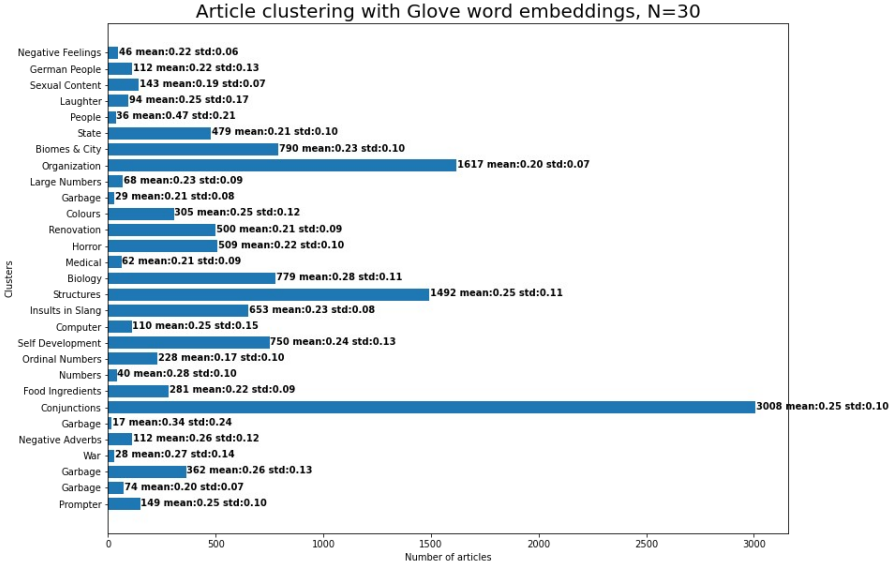
Definition: Clarity score measures how much more likely it is to observe word w in the subset of segments(sentences in our case) that discuss aspect(cluster, gaussian) a, compared to the corpus as a whole.

$$\text{score}_a(w) = t_a(w) \log_2 \frac{t_a(w)}{t(w)}$$

$t_a(w)$: l1-normalized tf-idf score of w in the segments annotated with aspect a.
t(w): l1-normalized tf-idf score of w in all of the segments.

Article Clustering





Article Clustering