Social Topic Distributions

Hakan Akyürek

Mürüvvet Hasanbaşoğlu

May 18, 2020

Overview

1. **Data structure**
2. **Preprocessing**
3. **Statistics of the dataset**
4. **Fine-tuning GloVe Embeddings**
5. **Next Tasks -** for the next 2 weeks

1. Data Structure

- Json files that contain some number of articles from a social media website
  - A number of comments for each article
- Relevance flag: 1 or 0


```
{
  "article_link":"https://medium.com/@johnroulac/oxford-study-attacks-regenerative-ag-monsanto-ally-2986ee9918c4",
  "resource_type":"blog",
  "article_title":"",
  "article_url":"https://www.facebook.com/organicconsumers/posts/10155409284844934",
  "article_id":"13341879933_10155409284844934",
  "search_query":"organic consumers",
  "article_text":"Supposed climate independent food research group claims grass-fed beef and CAFOs have the same climate impacts
  "article_source":"fb",
  "article_time":"2017-11-04 13:42:45",
  "comments":[
    {},
    {},
    {},
    {},
    {},
    {
      "comment_text":"Big Ag has strong influential power due to money unfortunately...",
      "comment_id":"10155409284844934_10155410227019934",
      "comment_rating":0,
      "comment_time":"2017-11-04 21:03:05",
      "comment_author":{
        "comment_author_id":"917663604962089",
        "comment_author_name":"Hannah Bessell"
      },
      "processed_comment_text":"Big Ag has strong influential power due to money unfortunately.",
      "custom_processed_comment_text":"big ag strong influential power due money unfortunately"
    },
    {},
    {},
    {},
    {}
  ],
  "article_author":[

  ],
  "relevant":1,
  "processed_article_text":"Supposed climate independent food research group claims grass-fed beef and CAFOs have the same clima
  "custom_processed_article_text":"suppose climate independent food research group claim grass fed beef cafos climate impact thi
}
```

2. Preprocessing

- ➔ Spelling Corrections(it cannot detect all of the spelling errors)
- ➔ Contractions(can’t ---> cannot)
- ➔ Language Curation
  - ◆ There are Hindi and English comments(though most of hindi ones are removed)
- ➔ Some social terms are replaced(btw ---> by the way)

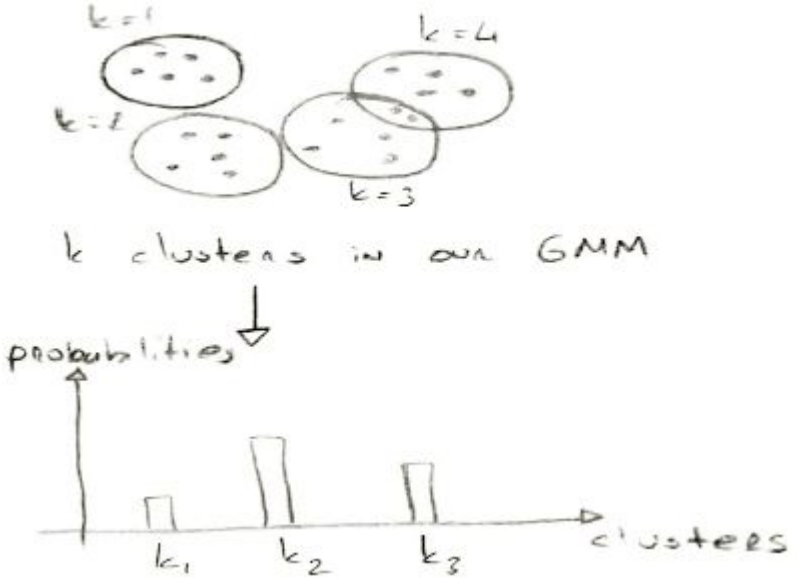
2. Preprocessing

➔ Lemmatization

➔ Stopword and punctuation removal

2 main ideas:

1. Syntactic and semantic features aren't musts for us.
   a. No topic labeling -- Probability distributions
2. Reducing the vocabulary size, more specifically oov words.



3. Statistics of the datasets

Raw Organic Dataset:

- Quora has balanced user distribution
- 4 datasets without comments
- 1 dataset without article

Data set			# of articles	# of articles with comments	# of answering user comments	# of all user comments
Biased		Facebook	5,013	4,705	298,996	299,126
		Food Babe	15	15	3,944	3,944
		Food Revolution	78	60	2,966	2,966
		Organic Authority	66	0	0	0
		Organic Consumers	64	0	0	0
Unbiased	Forum	Cafe Mom	86	85	1,962	1,983
		Disqus	36	36	6,150	7,984
		Quora	567	523	4,196	9,591
		Reddit	81	78	2,371	9,291
		US Message Board	0	0	0	78,044
	News sites	Chicago Tribune	2,283	78	281	281
		Huffington Post	880	0	0	0
		LA Times	1,522	77	374	374
		NY Post	106	0	0	0
		NY Times	438	137	16,128	16,128
		USA Today	95	22	259	259
		Washington Post	1,563	943	84,669	84,669

3. Statistics of the datasets

Raw Organic Dataset:

- Quora has balanced user distribution
- 4 datasets without comments
- 1 dataset without article

- Spiegel is in German

Data set			# of articles	# of articles with comments	# of answering user comments	# of all user comments
Biased		Facebook	5,013	4,705	298,996	299,126
		Food Babe	15	15	3,944	3,944
		Food Revolution	78	60	2,966	2,966
		Organic Authority	66	0	0	0
		Organic Consumers	64	0	0	0
Unbiased	Forum	Cafe Mom	86	85	1,962	1,983
		Disqus	36	36	6,150	7,984
		Quora	567	523	4,196	9,591
		Reddit	81	78	2,371	9,291
		US Message Board	0	0	0	78,044
	News sites	Chicago Tribune	2,283	78	281	281
		Huffington Post	880	0	0	0
		LA Times	1,522	77	374	374
		NY Post	106	0	0	0
		NY Times	438	137	16,128	16,128
		USA Today	95	22	259	259
		Washington Post	1,563	943	84,669	84,669

3. Statistics of the datasets

Raw Organic Dataset:

Data sets		# of words	Vocabulary size	# of rare words
All datasets		27,902,480	881,644	549,304
After custom preprocessing	**All datasets**	16,807,519	226,734	119,693
	Sub-datasets: Facebook, Quora, Reddit, NYTimes	6,039,245	106,579	56999

Rare words examples: `'lettuse', 'qoxt3iyqip', 'featherweight', 'cafeviennachicago', 'والخنازير', 'farmhaus'`

3. Fine-tuning GloVe Embeddings

Available GloVe embeddings:

Data sets		# of words	Vocabulary size	# of rare words	Out of Vocabulary size		
					Twitter 1.2M	Common Crawl 1.9M	Common Crawl 2.2M
All datasets		27,902,480	881,644	549,304	-	-	-
After custom preprocessing	All datasets	16,807,519	226,734	119,693	142,906	82,801	109,751
	Sub-datasets: Facebook, Quora, Reddit, NYTimes	6,039,245	106,579	56,999	53,232	30,585	39,195

3. Fine-tuning GloVe Embeddings

Available GloVe embeddings:

Data sets		# of words	Vocabulary size	# of rare words	Out of Vocabulary size		
					Twitter 1.2M	Common Crawl 1.9M	Common Crawl 2.2M
All datasets		27,902,480	881,644	549,304	-	-	-
After custom preprocessing	All datasets	16,807,519	226,734	119,693	142,906	**82,801**	109,751
	Sub-datasets: Facebook, Quora, Reddit, NYTimes	6,039,245	106,579	56,999	53,232	**30,585**	39,195

➔ Common Crawl: 42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB

3. Fine-tuning GloVe Embeddings

Available GloVe embeddings:

Data sets		# of words	Vocabulary size	# of rare words	Out of Vocabulary size		
					Twitter 1.2M	Common Crawl 1.9M	Common Crawl 2.2M
All datasets		27,902,480	881,644	549,304	-	**-**	-
After custom preprocessing	**All datasets**	16,807,519	226,734	119,693	142,906	**82,801 / 16,675**	109,751
	Sub-datasets: Facebook, Quora, Reddit, NYTimes	6,039,245	106,579	56,999	53,232	**30,585 / 5,777**	39,195

➔ Common Crawl: 42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB

3. Fine-tuning GloVe Embeddings

Fine-tuning:

1. Load pre-embeddings
2. Build co-occurrence matrix:
    a. Compute word list and find oov's
    b. Construct a corpus from the word list
    c. Compute the co-occurrence matrix based on the corpus with batches of oov's
3. Train new embeddings using the co-occurrence matrix, pre-embeddings and oov's

3. Next Tasks

- Figure out how to connect to the docker container
- Probably rerun the fine-tuning based on the feedback

- Train our GMM model
- Get probability distributions for each article and comments

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.
- Fine tune GloVe embeddings using Mittens
- roamanalytics/mittens: A fast implementation of GloVe, with optional retrofitting

Questions

General:

1. As a preprocessing step, does it make sense to remove the rare words in the entire dataset?
2. When constructing the co-occurrence matrix, should we specify the corpus per dataset? Would this produce more sensible embeddings?
3. Is it relevant for us for the moment to consider ‘relevant’ ‘irrelevant’ information of articles?

Team Specific:

4. If random users are all users, then it includes also the relevant users. How do we define the random users?
    a. Should all users and relevant users be from the same dataset?
5. Any idea why we get connection error on VPN when trying to connect to docker container?

6. **Distribution graph can be observed for**:
    1. all comments of a given user among different articles
    2. all users(random users)
    3. answering users(relevant users)