Social Topic Distributions

Hakan Akyürek

Mürüvvet Hasanbaşoğlu

May 04, 2020

Overview

1. **Project Goal**
2. **Outcomes**
3. **Methodology**
4. **Evaluation**
5. **Next Tasks -** for the next 2 weeks

1. Project Goal: What are we going to do?

Topic Modeling: aims to extract topics/clusters from a corpus of texts.

1. **Project Goal:** What are we going to do?

Topic Modeling: aims to extract topics/clusters from a corpus of texts.

➔ Derive social topic distributions of:

1- news and social media articles

2- social media users

using their word or sentence vector representations.

1. Project Goal: What are we going to do?

Topic Modeling: aims to extract topics/clusters from a corpus of texts.

➔ Derive social topic distributions of:

1- news and social media articles

2- social media users

using their word or sentence vector representations.

➔ Analyze the topic distributions of different sets of articles and users

2. **Outcomes:** What will we analyze?

- Analyze if news articles attract the users who have similar topics of interest
- Analyze if **relevant users** have more similar topic distributions compared to the **random users**
- Repeat and report the analysis among **different news and social media resources**

2. Outcomes: What will we analyze?

- Analyze if news articles attract the users who have similar topics of interest
- Analyze if **relevant users** have more similar topic distributions compared to the **random users**
- Repeat and report the analysis among **different news and social media resources**

Additionally:

- Group individual social media users that have similar topics of interest
- Group social media and news articles that have similar topics of interest

3. **Methodology:** How are we going to do it?

Raw Organic Dataset:

- Forums, blogs, news sites, social media
- Biased : more strict opinions towards organic food
- Unbiased : less strict opinions towards organic food

	English	German
Number of sentences	441895	487794
Number of comments	140119	94442
Portion of biased (%)	0.465	0.026
Number of tokens	7198582	7752885
Size of vocabulary	141579	262672

Table: Statistics of raw organic dataset

3. Methodology: How are we going to do it?

Raw Organic Dataset:

- Forums, blogs, news sites, social media
- Biased : more strict opinions towards organic food
- Unbiased : less strict opinions towards organic food

	English	German
Number of sentences	441895	487794
Number of comments	140119	94442
Portion of biased (%)	0.465	0.026
Number of tokens	7198582	7752885
Size of vocabulary	141579	262672

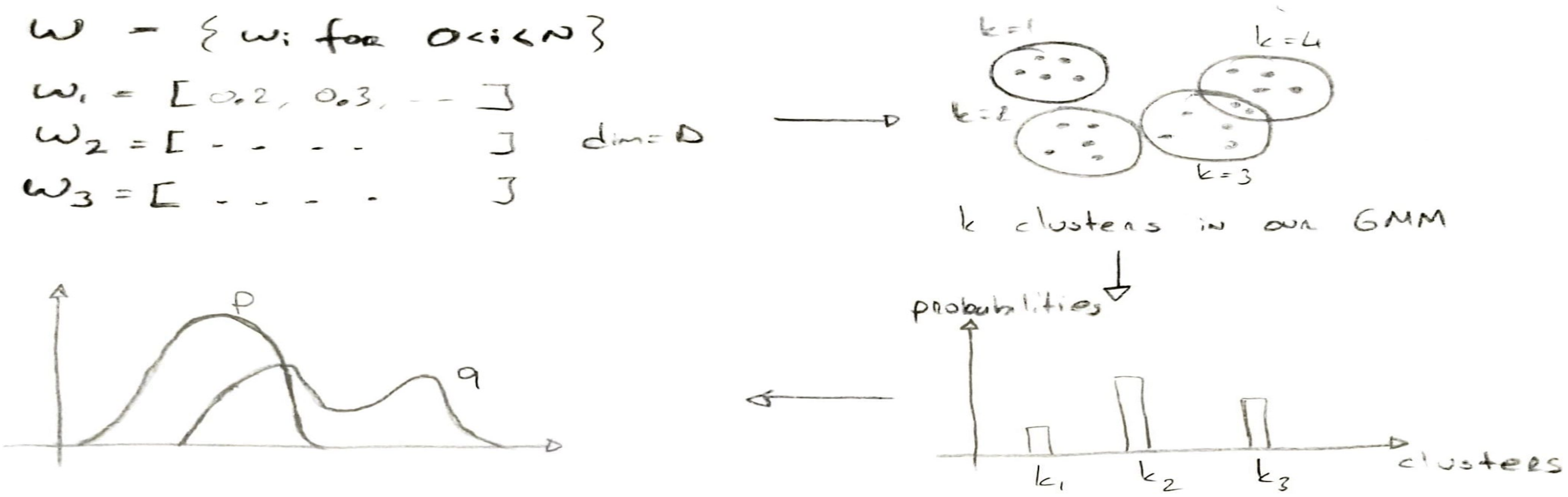
Table: Statistics of raw organic dataset

Features:

- Fine-tune the language model from BERT/GloVe on the whole dataset
- Word and sentence embeddings

3. **Methodology:** How are we going to do it?

Classifier: Gaussian Mixture Model



3. **Methodology:** How are we going to do it?

Histogram Generation:

- Calculate the probability that an article is about topic k
- We are able to get necessary probabilities from our GMM model

$$k^* = \arg\max_{\theta_k} p(k|w'_1, \cdots, w'_N)$$

$$= \arg\max_{\theta_k} p(w'_1, \cdots, w'_N|k)p(k)$$

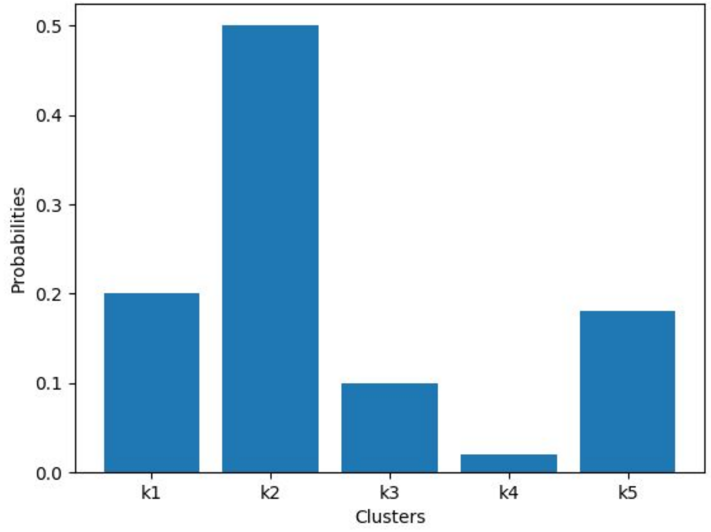
$$k^* = \arg\max_{\theta_k} p(k) \prod_{i=1}^{N} p(w'_i|k)$$

3. **Methodology:** How are we going to do it?

Histogram Generation:

- Calculate the probability that an article is about topic k
- We are able to get necessary probabilities from our GMM model
- Generate a histogram

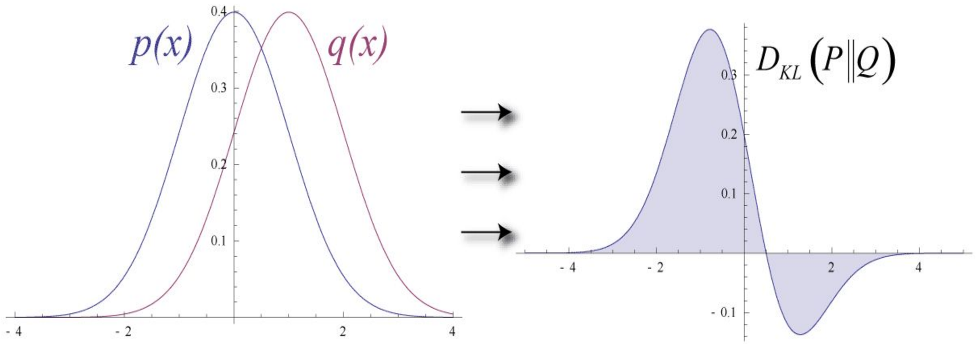
$$k^* = \arg\max_{\theta_k} p(k|w'_1, \cdots, w'_N)$$
$$= \arg\max_{\theta_k} p(w'_1, \cdots, w'_N|k)p(k)$$
$$k^* = \arg\max_{\theta_k} p(k) \prod_{i=1}^{N} p(w'_i|k)$$



4. Evaluation: How are we going to measure our results?

Performance metrics:

- Jensen Shannon or KL divergence to calculate similarity between distributions



Original Gaussian PDF's KL Area to be Integrated

How similar are p and q distributions?
In our case p would be an article and q, one of it's comments.

5. Next Tasks:

- Set up Colab environment
- Prepare the dataset:
    - cleaning
    - preprocessing
    - create random user dataset
- Get the vector representations for our data
- Further reading about the task

References

- Document Classification with Distributions of Word Vectors
- Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words