Clustering documents

WUM 2

Dawid Płudowski   Antoni Zajko

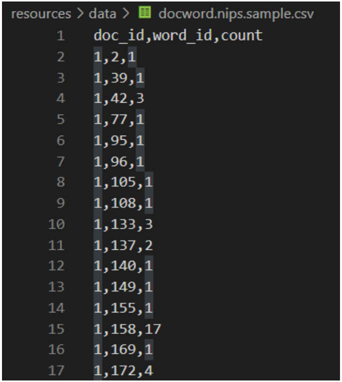
Wasaw University of Technology

Presentation Overview

1. Dataset
   - Data introduction
   - Preprocessing

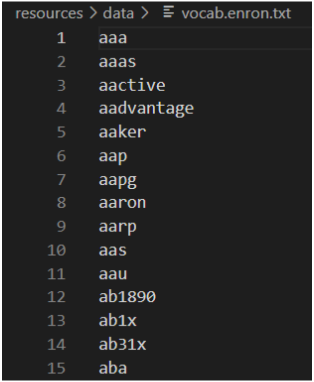
2. Models
   - Visualisation
   - Baseline
   - Final model

About the data



```
doc_id,word_id,count
1,2,1
1,39,1
1,42,3
1,77,1
1,95,1
1,96,1
1,105,1
1,108,1
1,133,3
1,137,2
1,140,1
1,149,1
1,155,1
1,158,17
1,169,1
1,172,4
```

Figure: Docword dataset example



```
aaa
aaas
aactive
aadvantage
aaker
aap
aapg
aaron
aarp
aas
aau
ab1890
ab1x
ab31x
aba
```

Figure: Vocab dataset example

Preprocessing methods

1. Sample 1500 documents from each dataset.
2. Convert raw dataframes into dictionaries with *Bags of Words* (*BoW*).
3. Split into train and test samples.
4. Filter out rare and type-wise tokens (such as links or stop words).
5. Generating additional features (LDA topics, statistics).
6. Encode BoW using ***T**erm **F**requency - **I**nverse **D**ocument **F**requency* (*tf-idf*).
7. Reduce dimensions with truncated SVD.
8. Standardize final dataset.

Models

1. Agglomerative;
2. DBScan;
3. Gaussian Mixture;
4. KMeans;
5. KMedioids;
6. LDA;

Visualisation

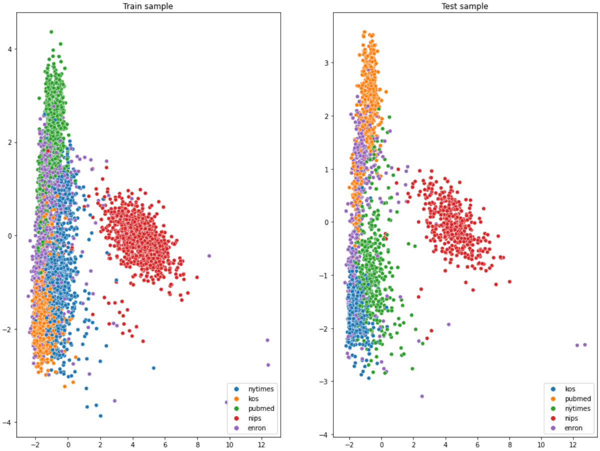


Figure: Visualisation of data with original labels

Baseline

Assign labels randomly.

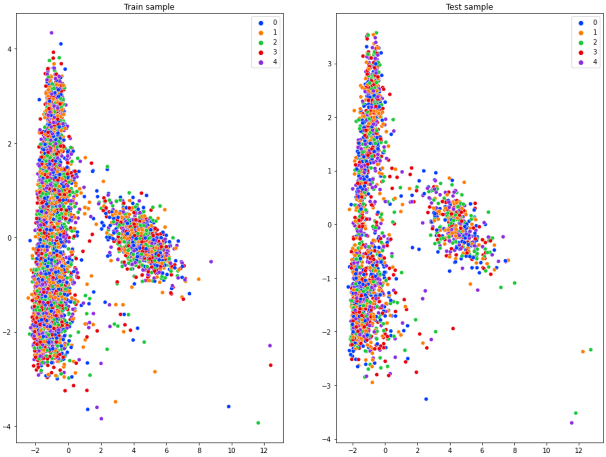


Figure: Visualisation of data with random model

Gaussian Mixture model

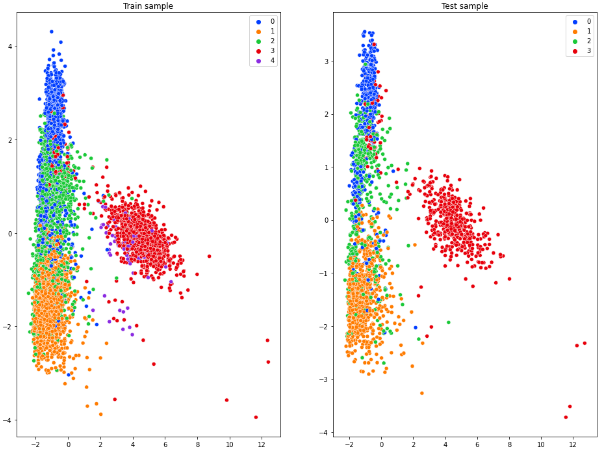
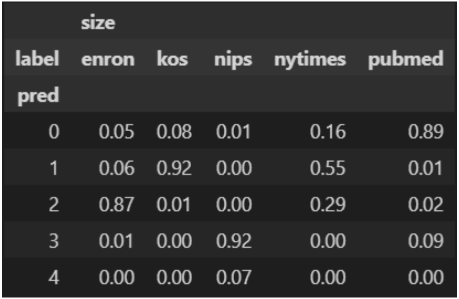


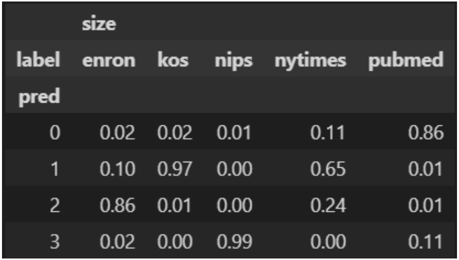
Figure: Visualisation of data with GMM model

Real labels comparison



	size				
label	enron	kos	nips	nytimes	pubmed
pred					
0	0.05	0.08	0.01	0.16	0.89
1	0.06	0.92	0.00	0.55	0.01
2	0.87	0.01	0.00	0.29	0.02
3	0.01	0.00	0.92	0.00	0.09
4	0.00	0.00	0.07	0.00	0.00

Figure: Comparison in train data



	size				
label	enron	kos	nips	nytimes	pubmed
pred					
0	0.02	0.02	0.01	0.11	0.86
1	0.10	0.97	0.00	0.65	0.01
2	0.86	0.01	0.00	0.24	0.01
3	0.02	0.00	0.99	0.00	0.11

Figure: Comparison in test data

Thanks for your attention