NATURAL LANGUAGE PROCESSING COURSE PROJECT

Comparison of topic modeling techniques

Evgenia Slivko

Full list of author information is available at the end of the article
*Equal contributor

Abstract

Goal of the project: To analyze and compare different topic modeling techniques based on unlabeled Wikipedia dataset.

Main results of the project: BERTopic and Top2Vec topic models showed the highest values of the coherence measure which indicates high degree of semantic similarity between high scoring words in the topic. The coherence measure of LDA model showed slightly lower value of coherence measure.

Background

Topic models represent probabilistic models which refers to statistical algorithms for discovering the latent semantic structures in the text. Topic modeling is unsupervised machine learning technique on unlabeled data aiming to extract the hidden topics underlying a collection of documents. Topic modeling is usually associated with setting a huge number of parameters and hyperparameters that affect the quality of modeling. The evaluation and comparison of topic models is still a hard task since models may demonstrate different quality depending on the dataset. The goal of this project is a comparison of the several topic models based on unlabeled Wikipedia dataset that contains articles with a great variety of topics.

Data

Dataset

For the aims of comparison of topic modeling techniques the part (50000 articles in English) of Wikipedia dataset (available https://www.lateral.io/resources-blog/the-unknown-perils-of-mining-wikipedia) was used.

Data preprocessing

Data preprocessing includes the folowing steps:

- Convert texts to lowercase in order to avoid adding the same word several times,
- Removing of stop words that have little or no significance and are used many times in most of the texts.
- Removing of special characters which add extra noise to the text
- Lemmatization - converting words to the base form with means of WordNetLemmatizer from NLTK library.
- Tokenization - converting articles to list of tokens.

Methods

In this project several topic models, such as LDA, BERTopic, and Top2Vec were created, analysed and compared. The sections below provide basic information about the analysis approach and models used, as well as the criteria for models comparison - coherence measure.

LDA

Latent Dirichlet allocation (LDA) is a generative statistical model where each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words. LDA maps documents to a list of topics by assigning each word in the document to different topics. LDA ignores the order of occurrence of words. It treats documents just as a collection of words or a bag of words.

Modeling with BERTopic + dimensionality reduction + clustering

According to the description of the BERTopic provided by the author of the model (https://github.com/MaartenGr/BERTopic/):

”BERTopic is a topic modeling technique that leverages BERT embeddings and a class-based TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions”.

BERTopic uses sentence-transformers to create embeddings for the documents. In this project two sentence-transformer embedding models were used in order to build the BERTopic model and test its performance :

- sentence-transformers/distilbert-base-nli-mean-tokens maps sentences and paragraphs to a 768 dimensional dense vector space (according to the latest information this model is depreciated and is not recommended to use due to low quality of the embeddings produced),
- sentence-transformers/all-MiniLM-L6-v2 maps sentences and paragraphs to a 384 dimensional dense vector space.

Since embeddings are highly dimensional what clustering algorithms deal poorly with, dimensionality reduction is an important preparatory step before clustering documents. In this project UMAP technique was used for the purpose of dimensionality reduction. Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimension reduction technique that keeps a large part of the high-dimensional local structure in lower dimensionality.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm developed by Campello, Moulavi, and Sander. This algorithm works quite well with UMAP and is able to find clusters of varying densities. HDBSCAN shows good performance on the data with arbitrarily shaped clusters, clusters with different sizes and densities and noise in the data what is typical for document embeddings data.

Top2Vec + dimensionality reduction + clustering

Top2Vec is an algorithm for topic modeling and semantic search that identifies topics in the text and generates jointly embedded topic, document and word vectors.

By default Top2Vec uses Doc2Vec to generate the joint word and document embeddings, but it also allow to use pretrained embedding models for generating joint word and document embeddings. In this project Top2Vec with default embedding model was used. After creation of the Top2Vec model, the same approach was applied as for BERTopic model: dimensionality reduction of embeddings and documents clustering based on resulting embeddings.

Performance measures

Coherence score was used as a measure of quality of different topic modeling techniques. Coherence score - a measure of goodness of a topic reflecting the quality of human judgment. The higher the coherence score, the topic is more coherent. In this project 2 coherence measures were used:

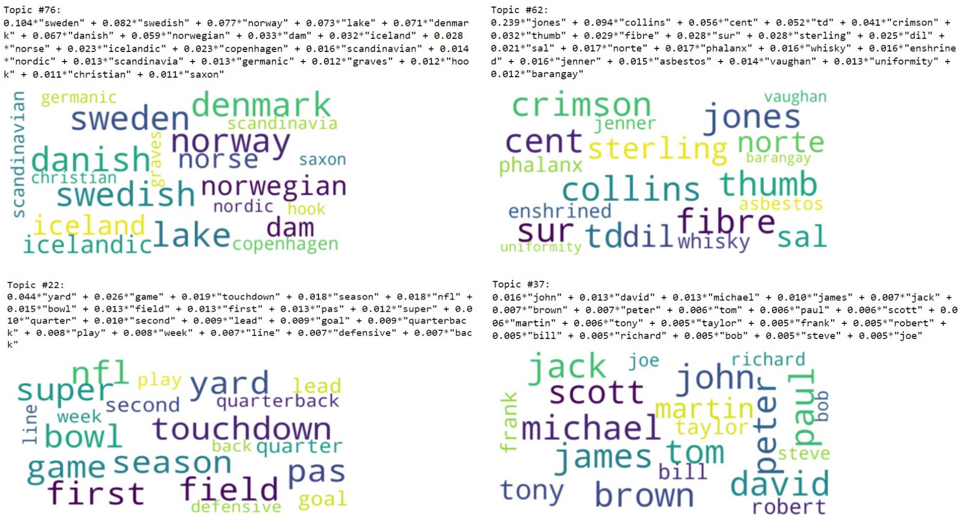
- c_npmi measure is based on a sliding window and the normalized pointwise mutual information of all word pairs of the given top words.
- c_v measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity

2 full iterations of the calculation were carried out using each of the coherence measures listed above. As expected, very similar results were obtained at each iteration.

Results

Results of LDA modeling with auto number of topics (100 topics)

LDA model with auto number of topics identified 100 topics in 50000 articles. Most of the topics identified by LDA models are pretty well interpretable. The wordclouds below show examples of the representation of the topics, generated by LDA model:



```
Topic #76:
0.104*"sweden" + 0.082*"swedish" + 0.077*"norway" + 0.073*"lake" + 0.071*"denmar
k" + 0.067*"danish" + 0.059*"norwegian" + 0.033*"dam" + 0.032*"iceland" + 0.028
*"norse" + 0.023*"icelandic" + 0.023*"copenhagen" + 0.016*"scandinavian" + 0.014
*"nordic" + 0.013*"scandinavia" + 0.013*"germanic" + 0.012*"graves" + 0.012*"hoo
k" + 0.011*"christian" + 0.011*"saxon"
```

```
Topic #62:
0.239*"jones" + 0.094*"collins" + 0.056*"cent" + 0.052*"td" + 0.041*"crimson" +
0.032*"thumb" + 0.029*"fibre" + 0.028*"sur" + 0.028*"sterling" + 0.025*"dil" +
0.021*"sal" + 0.017*"norte" + 0.017*"phalanx" + 0.016*"whisky" + 0.016*"enshrine
d" + 0.016*"jenner" + 0.015*"asbestos" + 0.014*"vaughan" + 0.013*"uniformity" +
0.012*"barangay"
```

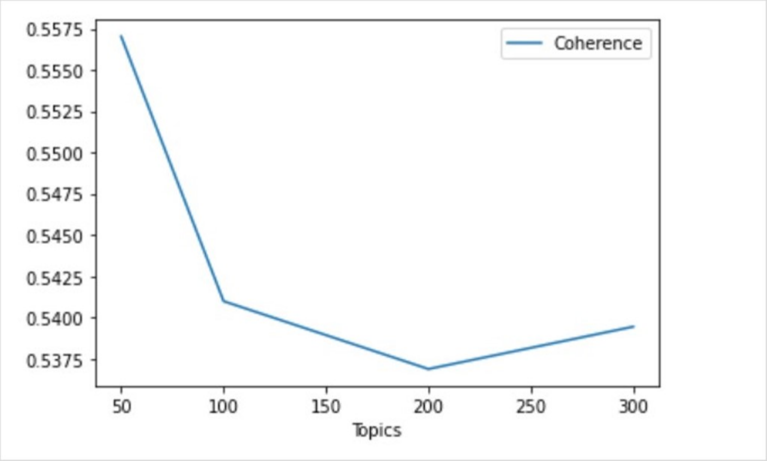
```
Topic #22:
0.044*"yard" + 0.026*"game" + 0.019*"touchdown" + 0.018*"season" + 0.018*"nfl" +
0.015*"bowl" + 0.013*"field" + 0.013*"first" + 0.013*"pas" + 0.012*"super" + 0.0
10*"quarter" + 0.010*"second" + 0.009*"lead" + 0.009*"goal" + 0.009*"quarterbac
k" + 0.008*"play" + 0.008*"week" + 0.007*"line" + 0.007*"defensive" + 0.007*"bac
k"
```

```
Topic #37:
0.016*"john" + 0.013*"david" + 0.013*"michael" + 0.010*"james" + 0.007*"jack" +
0.007*"brown" + 0.007*"peter" + 0.006*"tom" + 0.006*"paul" + 0.006*"scott" + 0.0
06*"martin" + 0.006*"tony" + 0.005*"taylor" + 0.005*"frank" + 0.005*"robert" +
0.005*"bill" + 0.005*"richard" + 0.005*"bob" + 0.005*"steve" + 0.005*"joe"
```

Wordcloud representation of the LDA model (number of topics = 100)

Results of modeling and hyperparameter tuning with LDA

In order to understand whether it is possible to improve the results of the LDA model, several LDA models with different number of topics were built and for each of them the coherence measure 'C_v' was calculated and used as the performance score.

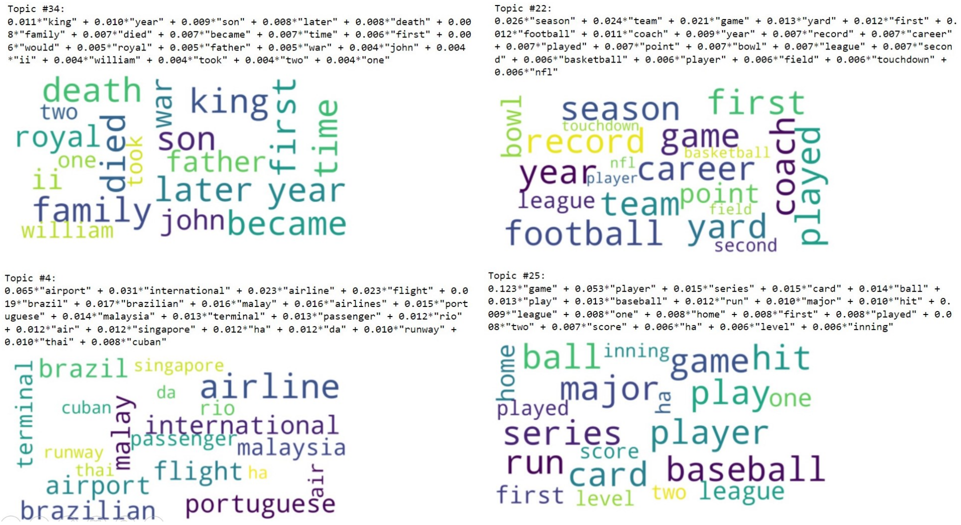


Search for the best number of topics of the LDA model based on coherence measure

The comparison of coherence measure values shows that LDA model with 50 topics has the highest coherence value and thus it should provide the most consistent and well interpretable topics.

Results of LDA model with optimal number of topics (50 topics)

LDA model with number of topics equal to 50 identified very well interpretable topics (See topic representation as wordclouds below):



```
Topic #34:
0.011*"king" + 0.010*"year" + 0.009*"son" + 0.008*"later" + 0.008*"death" + 0.00
8*"family" + 0.007*"died" + 0.007*"became" + 0.007*"time" + 0.006*"first" + 0.00
6*"would" + 0.005*"royal" + 0.005*"father" + 0.005*"war" + 0.004*"john" + 0.004
*"ii" + 0.004*"william" + 0.004*"took" + 0.004*"two" + 0.004*"one"
```

```
Topic #22:
0.026*"season" + 0.024*"team" + 0.021*"game" + 0.013*"yard" + 0.012*"first" + 0.
012*"football" + 0.011*"coach" + 0.009*"year" + 0.007*"record" + 0.007*"career"
+ 0.007*"played" + 0.007*"point" + 0.007*"bowl" + 0.007*"league" + 0.007*"secon
d" + 0.006*"basketball" + 0.006*"player" + 0.006*"field" + 0.006*"touchdown" +
0.006*"nfl"
```

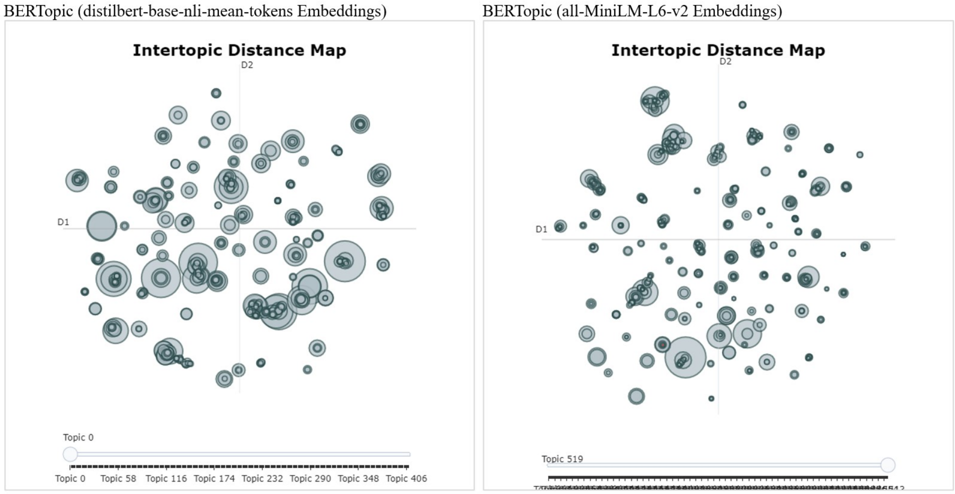
```
Topic #4:
0.065*"airport" + 0.031*"international" + 0.023*"airline" + 0.023*"flight" + 0.0
19*"brazil" + 0.017*"brazilian" + 0.016*"malay" + 0.016*"airlines" + 0.015*"port
uguese" + 0.014*"malaysia" + 0.013*"terminal" + 0.013*"passenger" + 0.012*"rio"
+ 0.012*"air" + 0.012*"singapore" + 0.012*"ha" + 0.012*"da" + 0.010*"runway" +
0.010*"thai" + 0.008*"cuban"
```

```
Topic #25:
0.123*"game" + 0.053*"player" + 0.015*"series" + 0.015*"card" + 0.014*"ball" +
0.013*"play" + 0.013*"baseball" + 0.012*"run" + 0.010*"major" + 0.010*"hit" + 0.
009*"league" + 0.008*"one" + 0.008*"home" + 0.008*"first" + 0.008*"played" + 0.0
08*"two" + 0.007*"score" + 0.006*"ha" + 0.006*"level" + 0.006*"inning"
```

Wordcloud representation of the LDA model (number of topics = 50)

Results of modeling with BERTopic

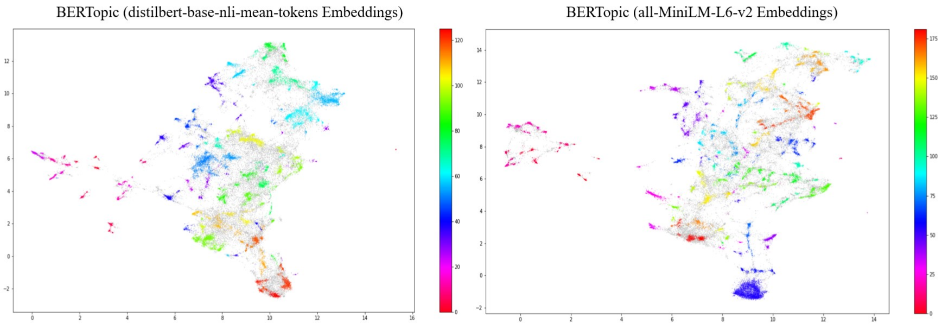
BERTopic models with two different embeddings were analyzed. Both models extracted high number of the topics (406 and 519). On the intertopic distance map presented for both BERTopic models we can observe scattered clusters of the close or similar topics that could be combined. Topics extracted by BERTopic with all-MiniLM-L6-v2 embeddings are more granulated in comparison to BERTopic model with distilbert-base-nli-mean-tokens.



Visualization of topics extracted by BERTopic models

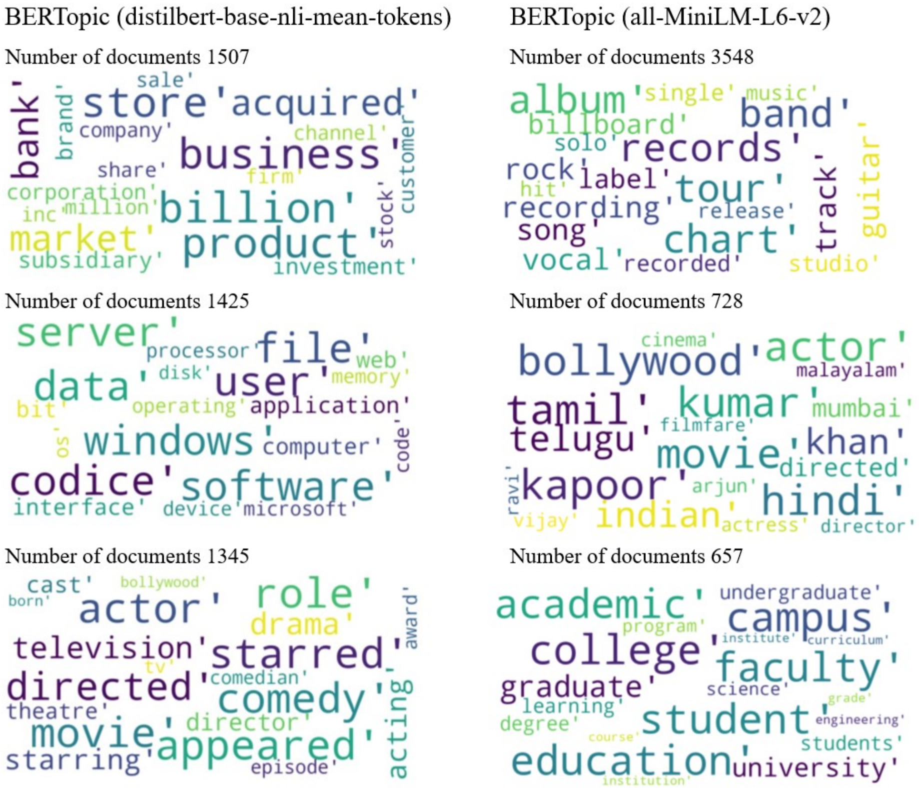
For both models dimensionality reduction and clustering was performed. After dimensionality reduction and clustering huge amount of articles (> 30%) we classified as outliers in both models, which means that they were aggregated to a mutual fragmented class (grey dots on the map): 25443 articles for distilbert-base-nli-mean-tokens and 24857 articles for BERTopic+ all-MiniLM-L6-v2.

The resulting number of topics of the BERTopic models with distilbert and all-MiniLM-L6-v2 embeddings: 127 and 183 topics correspondingly.



Representation of the BERTopic model topics as clusters after dimensionality reduction on 2d map

The screenshot below shows the topics with the highest number of articles identified by BERTopic + distilbert-base-nli-mean-tokens and BERTopicwith + all-MiniLM-L6-v2. The models have defined different topics: top 3 topics by the 1st model are marketing and business, software, movies and TV and top 3 topics by the 2nd model are music, bollywood and education.



BERTopic (distilbert-base-nli-mean-tokens)

BERTopic (all-MiniLM-L6-v2)

Number of documents 1507

Number of documents 3548

Number of documents 1425

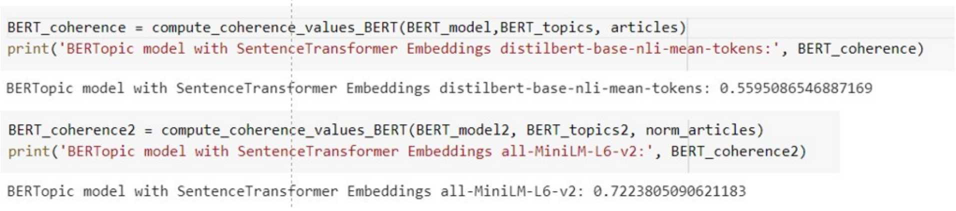
Number of documents 728

Number of documents 1345

Number of documents 657

Wordcloud representation of the BERT model topics with the highest number of articles

The coherence measure 'c_v' was calculated for both models. BERTopicwith + all-MiniLM-L6-v2 model showed a higher value:

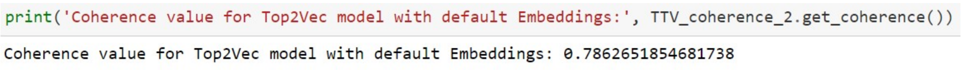


```
BERT_coherence = compute_coherence_values_BERT(BERT_model,BERT_topics, articles)
print('BERTopic model with SentenceTransformer Embeddings distilbert-base-nli-mean-tokens:', BERT_coherence)

BERTopic model with SentenceTransformer Embeddings distilbert-base-nli-mean-tokens: 0.5595086546887169

BERT_coherence2 = compute_coherence_values_BERT(BERT_model2, BERT_topics2, norm_articles)
print('BERTopic model with SentenceTransformer Embeddings all-MiniLM-L6-v2:', BERT_coherence2)

BERTopic model with SentenceTransformer Embeddings all-MiniLM-L6-v2: 0.7223805090621183
```

Coherence measure values for BERTopic models

Results of the modeling with Top2Vec

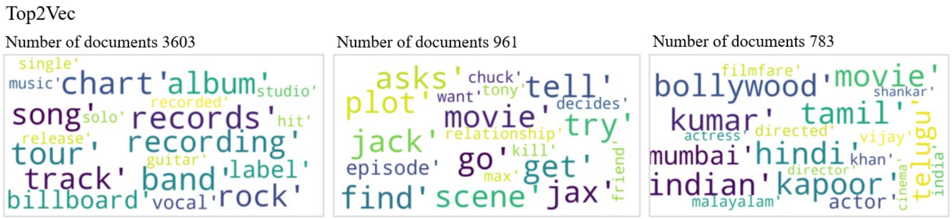
In order to analyze results of Top2Vec model i used an approach similar to BERTopic models analysis. Top2Vec model identidied 323 topics in total. After the initial formation of topics with the Top2Vec model, I used UMAP to reduce the dimensionality of embeddings, as well as hdbscan for clustering. As a result, 17772 articles were classified as outliers which is the best result among BERTopic and Top2Vec models. The coherence measure 'c_v'for Top2Vec model also showed the highest value among all the models.



```
print('Coherence value for Top2Vec model with default Embeddings:', TTV_coherence_2.get_coherence())

Coherence value for Top2Vec model with default Embeddings: 0.7862651854681738
```

Coherence measure value for Top2Vec model

Wordcloud representation for 3 topics with the highest number of articles demonstrates that 2 out of 3 topics by Top2Vec similar to topics with the largest number of articles generated by BERT+all-MiniLM-L6-v2 model.



Top2Vec

Number of documents 3603

Number of documents 961

Number of documents 783

Wordcloud representation for 3 topics with the highest number of articles (Top2Vec model)

Comparison of results

The comparison of the models showed that the topic models are very sensitive to the changes of parameters and hyperparameters.

Depending on the purpose of modeling, it is possible to create a model with more granular topics or more generalized ones.

On the example of 3 largest and 3 smallest topics (by the number of documents in the topic) we can see that all topics are well interpretable. ”Large” topics represent broad subjects (software and programming, business, TV, movies and actors, music, animals), while ”small” topics represent specialized questions (encoding and fonts, taxes, healthcare). Topics with the largest number of articles according to BERTopic with ”all-MiniLM”-L6-v2 embeddings and Top2Vec are quite similar (Music, Movies and TV, Bolywood and Music, Bollywood and education).

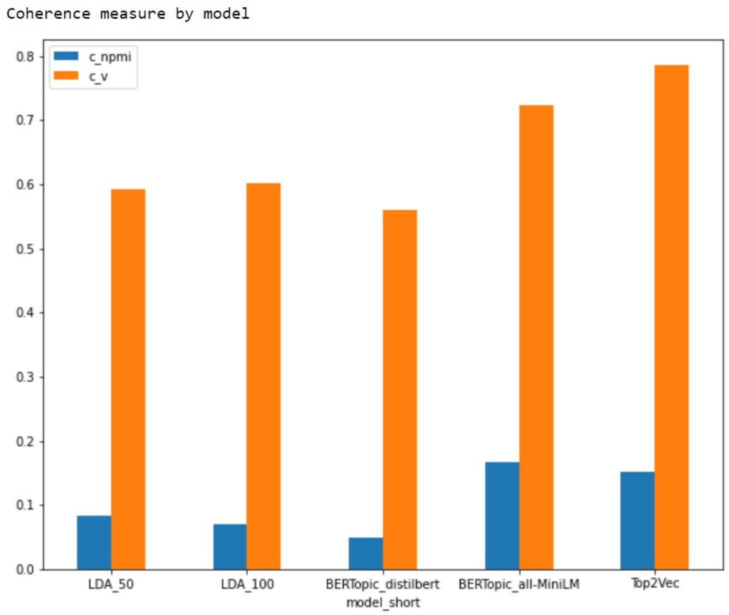
Comparison of models based on coherent measures

In this project coherence measures ’c_npmi’ and ’c_v’ were calculated for all topic models.

The model Top2Vec and BERTopic with ”all-MiniLM-L6-v2” embeddings showed the highest values of the coherence measures.

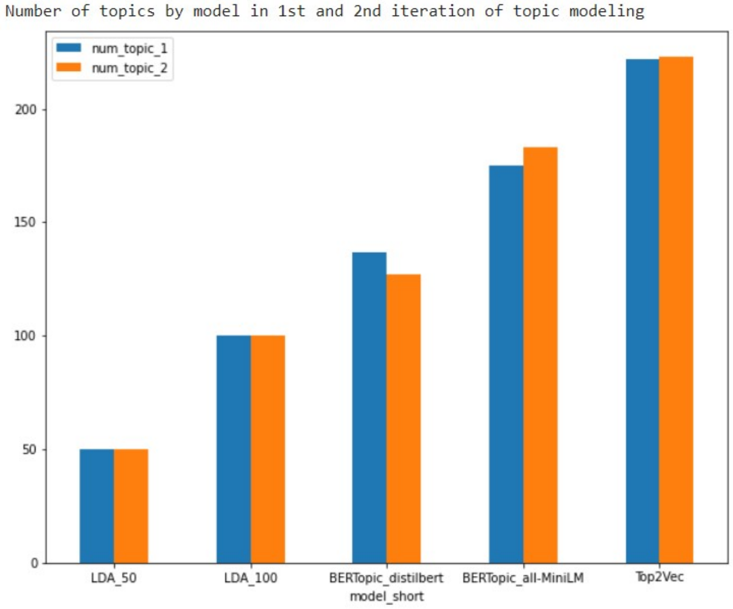
LDA with number of topics = 50 had the highest coherence value ’c_npmi’ among LDA models with 50, 70, 100, 200 and 300 topics. At the same time, LDA with number of topics = 100 had the highest coherence value ’c_v’. Slightly different coherence values calculated for LDAMulticore models for the same number of topics (For details see hyperparameter tuning section) were observed.

Nevertheless, coherence values of LDA models are lower than the coherence values of BERTopic and Top2Vec.



Comparison of the coherence measure value

During the model evaluation two iterations of modeling were performed with the same parameters. Resulting models showed slightly different number of topics in two iterations which does not affect the overall result.



Comparison of the number of topics

Possible development of the project

For a dataset as vast and varied as wikipedia, it is worth to consider the possibility of hierarchical topic modeling (hierarchical latent Dirichlet allocation (hLDA), hierarchical additive regularization of topic models (ARTM), etc.).