Topic Modeling on Short Text Data

Paper : Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis
Authors : Rania Albalawi, Tet Hin Yeap and Morad Benyoucef
Year : 2020
Journal : Frontiers in Artificial Intelligence

Nidhi Gowri Srinath
801199302

Motivation

- NLP techniques enable computers to perceive textual and audio input in the same way that humans can.
- The majority of techniques for machine learning require structured and organized data. NLP enables us to work with unstructured data to uncover important insights. Typically, data is unstructured. A lot of crucial information is lost during the data refining process.
- Using topic modeling on textual data will aid in extracting hidden topics from enormous amounts of data.
- Topic modeling has various algorithms that can be employed.

Problem Statement

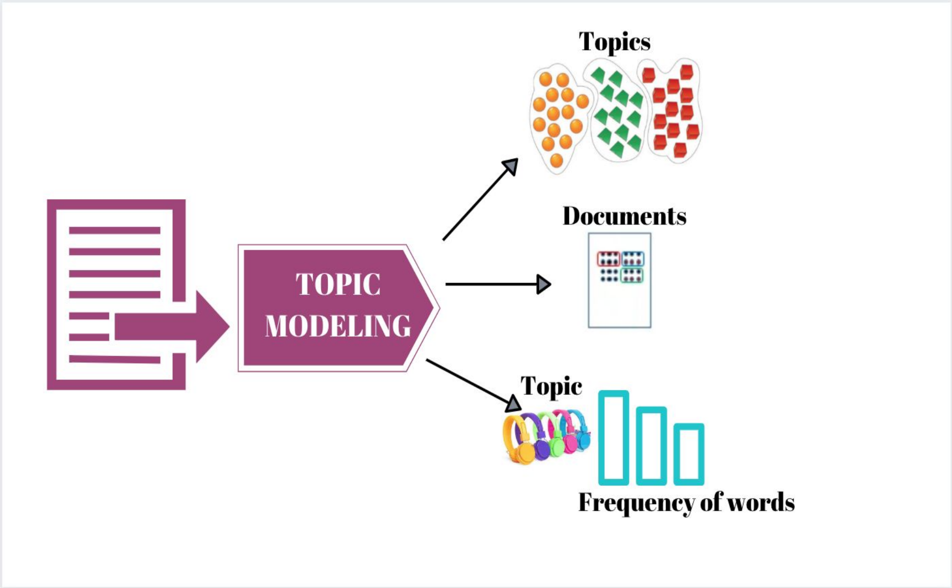
- Every online application that works with textual data necessitates the usage of categories to improve the way data is recommended to users. We need to discover categories in the data that will help us effectively cluster them.
- Topic modeling is an unsupervised machine learning technique that can scan a corpus of texts and identify word and phrase patterns.
- Identifying common “topics” can help segregate data into logical groups.

Paper Summary and Existing Approaches

“Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis”

- The authors intend to find application areas, methods and tools for topic modeling using machine learning and natural language processing techniques.
- They also compare five commonly used topic modeling techniques, namely; latent semantic analysis, latent Dirichlet allocation, non-negative matrix factorization, random projection, and principal component analysis.
- To evaluate these techniques, the authors consider statistical evaluation metrics such as f-score, precision, and topic coherence.
- As a result of the tests conducted by the authors, they found that Latent Dirichlet Allocation (LDA) had the most valuable results followed by Non-Negative Matrix Factorization ((NMF)

Topic Modeling



Topic modeling is a kind of statistical modeling used to identify abstract "topics" that appear in a collection of texts. Topic Models are extremely effective for document clustering, structuring big blocks of textual data, retrieving information from unstructured text, and feature selection.

Image source : medium.com

Latent Dirichlet Allocation

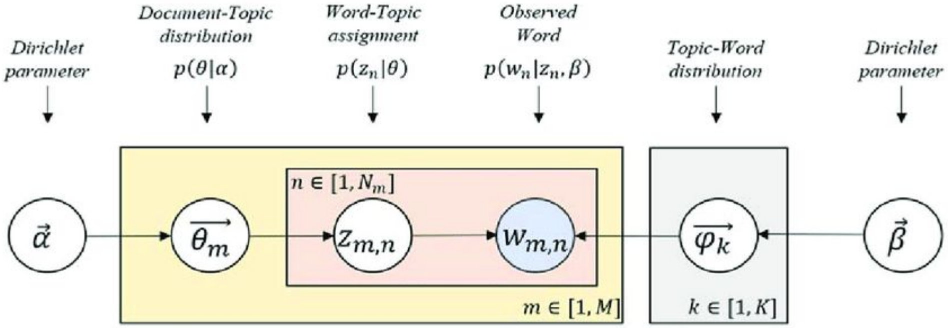
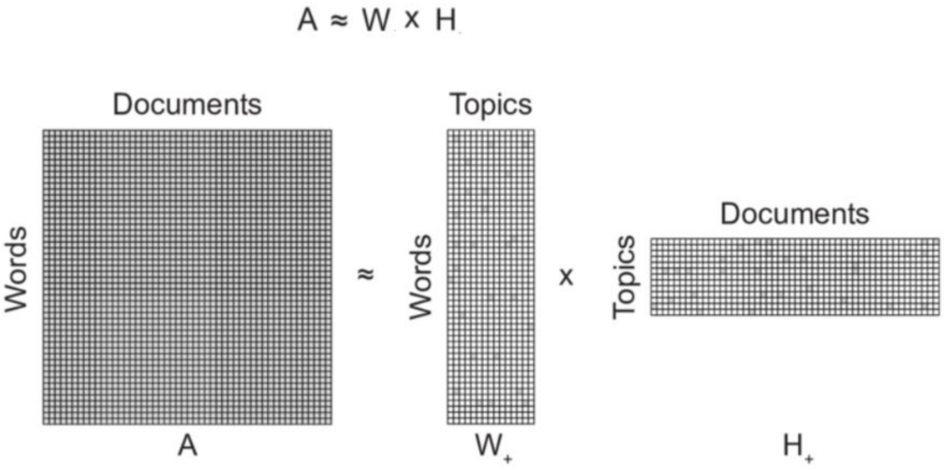


Image source : researchgate.net

- Latent Dirichlet Allocation (LDA) is a popular topic modeling technique to extract topics from a given corpus. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

  LDA assumes that documents are a mixture of topics and topics are a mixture of tokens or words.
- LDA's main purpose is to identify the best representation of the Document-Topic matrix and the Topic-Word matrix in order to determine the best Document-Topic distribution and Topic-Word distribution.

Non Negative Matrix Factorization



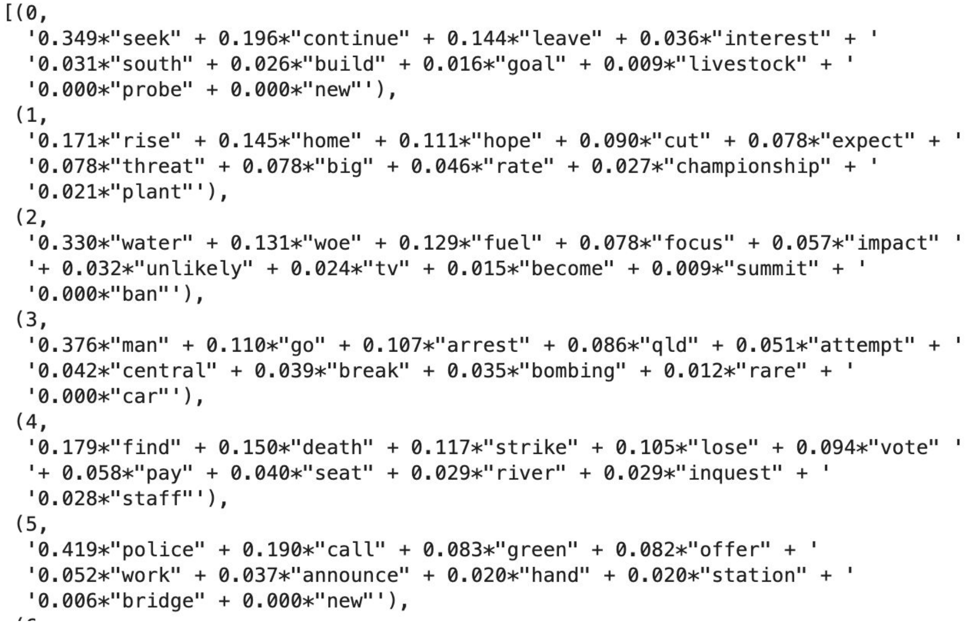
- Non-Negative Matrix Factorization (NMF) is an unsupervised approach, which means that the topics on which the model will be trained are not labeled.
- NMF works by decomposing (or factoring) high-dimensional vectors into a lower-dimensional form. Because these lower-dimensional vectors are non-negative, their coefficients are also non-negative.
- NMF is generally more scalable than LDA.

Image source: towardsdatascience.com

Implementation Steps

- Importing the necessary packages
- Importing the dataset and initialize it to the required columns
- Data Preprocessing
- Creating Bigrams and Trigrams
- Removing Stopwords
- Performing Lemmatization
- Creating a dictionary and a corpus
- Initializing the LDA model
- Visualizing the topics and evaluating the model
- Initializing the NMF model
- Visualizing the topics and evaluating the model

Methods Used - LDA



```
[(0,
  '0.349*"seek" + 0.196*"continue" + 0.144*"leave" + 0.036*"interest" + '
  '0.031*"south" + 0.026*"build" + 0.016*"goal" + 0.009*"livestock" + '
  '0.000*"probe" + 0.000*"new"'),
 (1,
  '0.171*"rise" + 0.145*"home" + 0.111*"hope" + 0.090*"cut" + 0.078*"expect" + '
  '0.078*"threat" + 0.078*"big" + 0.046*"rate" + 0.027*"championship" + '
  '0.021*"plant"'),
 (2,
  '0.330*"water" + 0.131*"woe" + 0.129*"fuel" + 0.078*"focus" + 0.057*"impact" '
  '+ 0.032*"unlikely" + 0.024*"tv" + 0.015*"become" + 0.009*"summit" + '
  '0.000*"ban"'),
 (3,
  '0.376*"man" + 0.110*"go" + 0.107*"arrest" + 0.086*"qld" + 0.051*"attempt" + '
  '0.042*"central" + 0.039*"break" + 0.035*"bombing" + 0.012*"rare" + '
  '0.000*"car"'),
 (4,
  '0.179*"find" + 0.150*"death" + 0.117*"strike" + 0.105*"lose" + 0.094*"vote" '
  '+ 0.058*"pay" + 0.040*"seat" + 0.029*"river" + 0.029*"inquest" + '
  '0.028*"staff"'),
 (5,
  '0.419*"police" + 0.190*"call" + 0.083*"green" + 0.082*"offer" + '
  '0.052*"work" + 0.037*"announce" + 0.020*"hand" + 0.020*"station" + '
  '0.006*"bridge" + 0.000*"new"'),
```

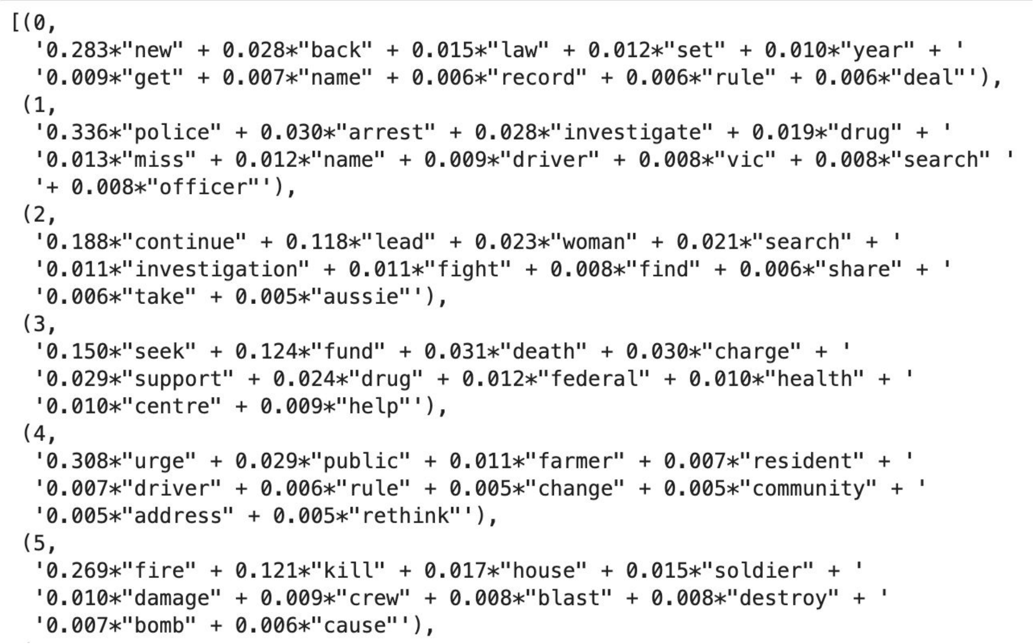
5 sample topics to understand how LDA assigns weights to each keyword in every topic.

LDA

LDA model takes an alpha and a beta value

The LDA model is initialized and made to identify 20 topics from the data. The model's coherence score and perplexity is calculated.

Methods Used - NMF



```
[(0,
  '0.283*"new" + 0.028*"back" + 0.015*"law" + 0.012*"set" + 0.010*"year" + '
  '0.009*"get" + 0.007*"name" + 0.006*"record" + 0.006*"rule" + 0.006*"deal"'),
 (1,
  '0.336*"police" + 0.030*"arrest" + 0.028*"investigate" + 0.019*"drug" + '
  '0.013*"miss" + 0.012*"name" + 0.009*"driver" + 0.008*"vic" + 0.008*"search" '
  '+ 0.008*"officer"'),
 (2,
  '0.188*"continue" + 0.118*"lead" + 0.023*"woman" + 0.021*"search" + '
  '0.011*"investigation" + 0.011*"fight" + 0.008*"find" + 0.006*"share" + '
  '0.006*"take" + 0.005*"aussie"'),
 (3,
  '0.150*"seek" + 0.124*"fund" + 0.031*"death" + 0.030*"charge" + '
  '0.029*"support" + 0.024*"drug" + 0.012*"federal" + 0.010*"health" + '
  '0.010*"centre" + 0.009*"help"'),
 (4,
  '0.308*"urge" + 0.029*"public" + 0.011*"farmer" + 0.007*"resident" + '
  '0.007*"driver" + 0.006*"rule" + 0.005*"change" + 0.005*"community" + '
  '0.005*"address" + 0.005*"rethink"'),
 (5,
  '0.269*"fire" + 0.121*"kill" + 0.017*"house" + 0.015*"soldier" + '
  '0.010*"damage" + 0.009*"crew" + 0.008*"blast" + 0.008*"destroy" + '
  '0.007*"bomb" + 0.006*"cause"'),
```

5 sample topics to understand how NMF assigns weights to each keyword in every topic.

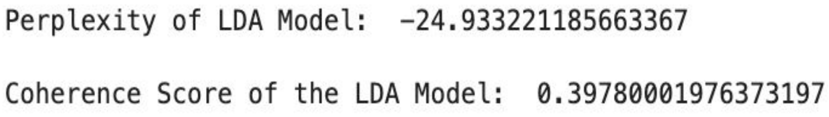
NMF

The NMF model is initialized and made to identify 20 topics from the data. The model’s coherence score is calculated

Coherence score indicates the readability of the model’s output data

Model Evaluations

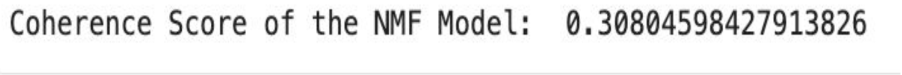
LDA



```
Perplexity of LDA Model:  -24.933221185663367

Coherence Score of the LDA Model:  0.39780001976373197
```

NMF

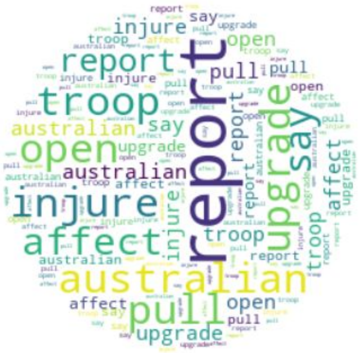


```
Coherence Score of the NMF Model:  0.30804598427913826
```

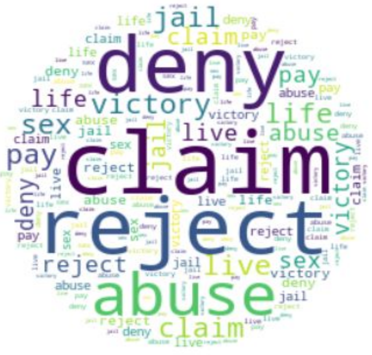
- Based on the results obtained from the model, we can see that the LDA model performs better.
- Perplexity is a measure of how well the probability model predicts a sample
- Coherence score scores a single topic by measuring the degree of semantic similarity between high scoring words in a topic

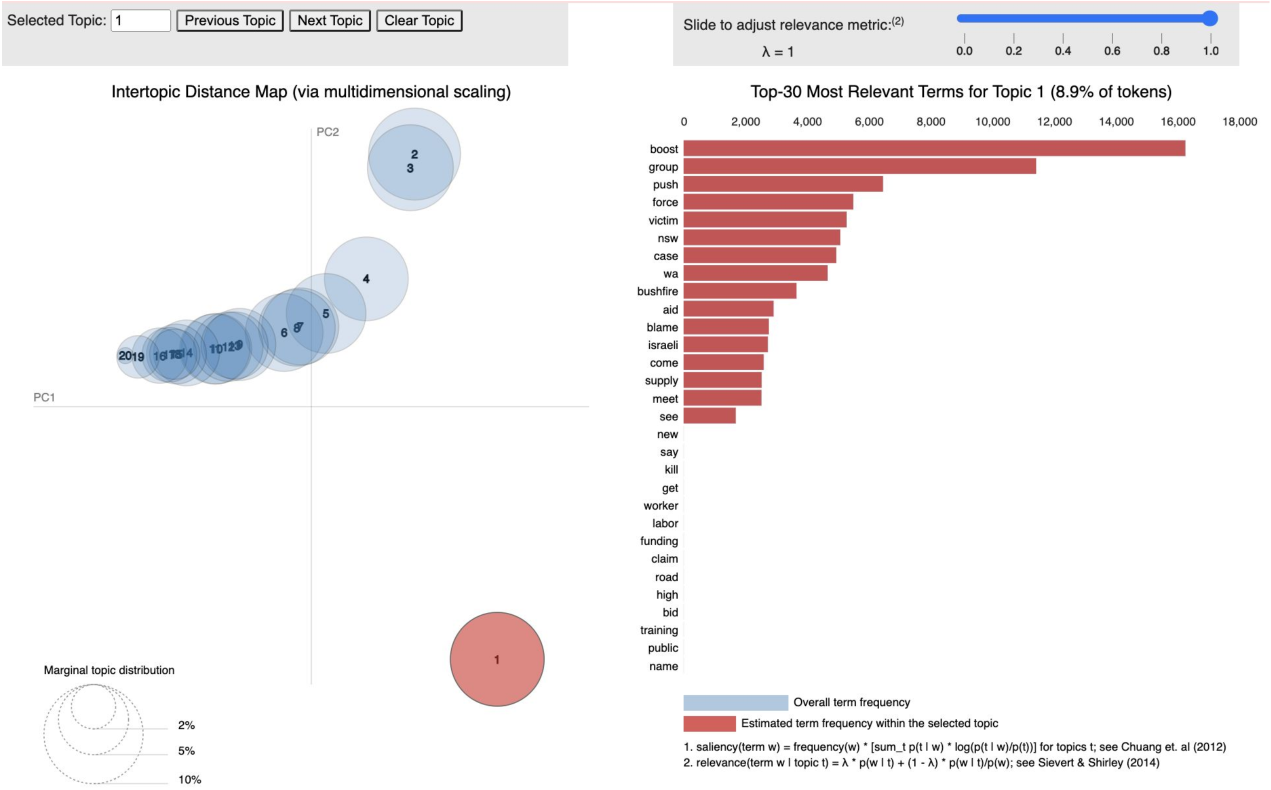
Visualizations

LDA Topic Wordcloud



NMF Topic Word Cloud





1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Future Work

- The model may be updated to include a looping mechanism for hyperparameter tweaking to determine the best values for alpha, beta, kappa, chunk sizes, passes, maximum iterations, and so on.
- The model could include a method for selecting the optimal number of topics that yield the best results in the shortest period of time. As of now, the number of topics has been manually entered to determine which combination of parameters produces the best results and coherence scores.

References

- Paper 1 - Albalawi, Rania, et al., “Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis”, Frontier in Artificial Intelligence, 2020, DOI : 10.3389/frai.2020.00042 (link)
- Paper 2 - Jelodar, Hamed, et al., “Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey”, arXiv, 2018, arXiv:1711.04305 (link)
- Topic Modeling and Latent Dirichlet Allocation (LDA) in Python (link)
- Topic Modeling using Gensim-LDA in Python (link)
- Topic Modeling Articles with NMF (link)