Topic Modeling Algorithms

Govardhan Digumurthi
CWID: A20286842

INTRODUCTION: Repeated clusters of statistically important tokens or phrases in a corpus. Models are being used in social science to discover variables or characteristics as well. Topic model doesn't tell you how many topics might be present in the corpus but we may tell the model to look for the mentioned number of topics. A topic model produces two results. Following a list of terms with a high likelihood of being related to each subject, each document is assigned two themes. Results may appear as a list of the most popular terms related to each topic. the same term used differently in each topic; the same word used differently in several topics

ALGORITHMS:

It is the process of automatically finding the hidden topics in a text data using unsupervised methods such as

1. LDA(Latent Dirichlet Allocation),
2. LDA_top model,
3. LSA(Latent Semantic Allocation),
4. NMF(Negative Matrix Factorization),
5. HDP(Hierarchical Dirichlet process)
6. LSI(Latent semantic Indexing)

Here, we are finding the topics from the corpus and corpus being a group of documents containing news articles.

Before applying algorithms we need to clean and preprocess the data.The text will be transformed into a representation of a bag of words, and the linguistic form for every word will be condensed using lemmatization and stemming methods into a single root.
LDA depicts documents as a combination of subjects, with each topic represented by a combination of words.

DATASET:

I have news article dataset for this project which is dataset “newsarticles.csv”

- Import all the required libraries. Initially Data is cleaned by creating a function clean_text, make text lowercase, remove text in square brackets, remove punctuation, remove read errors, and remove words containing numbers.
- A function “noun” is created to extract and lemmatize text. The text is tokenized, and nouns are extracted.
- A mask is created to isolate words that are nouns. Stores function to split string of words into a list of words or tokens. Then each word is lemmatized. A list is created with all the nouns which are lemmatized previously.
- Also, single character words are removed and returns list of nouns.
- All the nouns from articles are stored in a data frame. Remove articles which have 2 or less than 2 nouns in it because we need at least more than two nouns to distinguish the topics with a better coherence value.
- A corpus is created to store everything as topic modeling can be applied on a corpus.

Dataset source is from kaggle.com.

Topic Modeling Techniques Implemention:

NMF: Non-Negative Matrix Factorization is a popular technique, its main goal is to reduce the dimension of input corpora. It uses parameters to compare less weighted words to the words which have less coherence. The input non-negative matrix is TF-IDF normalized. This method uses basic linear algebra for topic modelling.

As the data is cleaned and processed. We use the nmfmodel variable to store the model of NMF with the required parameters passed.

NMF is also be implemented by scikit learn(TF_IDF vectorizer) without using “genism”. So that we can compare the NMF(genism) manually Vs NMF(scikitlearn) results.

LSI: Latent Semantic Indexing which inputs a matrix containing a bag of words. Similar to NMF, LSI also reduces the dimension. The concept of rare or infrequent words adds the weight model. But in the training record its occurrence is increased within the same document whereas if it occurs on corpus the occurrences are decreased. The matrix is

high-dimensional and consists of low frequency words. So using SVD we can reduce the dimension

Similar to NMF implementation, we use lsi model variable to store the model of LSI with the required parameters passed.

LDA: Latent Dirichlet Allocation. Unlike LSI and NMF, LDA purely deals with Topic modeling problems.

Model Evaluation: A function evaluate_bar_graph(coherences, indices), coherences: list of coherence values, indices: Indices to be used to mark bars. Length of this and coherences should be equal. Created to evaluate the above techniques implemented.

Futher Implementation or limitations:

1. **With the above three algorithms cannot determine which might be the exactly best fit algorithm. So we may implement the LDA_top model and HDP(Hierarchical Dirichlet process) to evaluate all models for a higher coherence score which may be the best model.**
2. **One of the problems with LDA is that if we train it on many topics, the topics get "lost" among the numbers. We can try to dig out the best topics from the best LDA model we can produce. That function can be used to control the quality of the LDA model we produce.**
3. **Hence, we can try creating two other models using LDA Technique as LDA best and “LDA top” which is implemented using a threshold of coherence value.**