**Topic Modeling on News Articles**

**Manoj Kumar**

**1.Data Description:**

The dataset contains a set of news articles for each major segment consisting of business, entertainment, politics, sports and technology. we need to create an aggregate dataset of all the news articles and perform topic modeling on this dataset. Verify whether these topics correspond to the different tags available.

**2.Problem Description:**

In this project task is to identify major themes/topics across a collection of news articles. using probabulistic model such as Latent Dirichlet Allocation (LDA) to find out the patterns and themes in between topics

**3. Methods involved:**

we are using a two methods here

Topic Modeling on News Articles Using NLTK\_Gensim

* **Topic Modeling on News Articles Using NLTK\_Gensim**

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.

**steps involved in nltk:**

* Loading the data
* Clean the Data
* Handling the Stop-words
* Lemmatization
* TF-IDF Vectorization
* Topic Modelling using Gensim's LDA

**4. Lemmatization:**

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meanings to one word.

Text preprocessing includes both Stemming as well as Lemmatization. Many times people find these two terms confusing. Some treat these two as the same. Actually, lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

Applications of lemmatization are

* Used in comprehensive retrieval systems like search engines.
* Used in compact indexing

**5.TF-IDF Vectorization:**

Bag of words (BoW) converts the text into a feature vector by counting the occurrence of words in a document. It is not considering the importance of words. Term frequency — Inverse document frequency (TFIDF) is based on the Bag of Words (BoW) model, which contains insights about the less relevant and more relevant words in a document. The importance of a word in the text is of great significance in information retrieval.

Example — If you search something on the search engine, with the help of TFIDF values, search engines can give us the most relevant documents related to our search.

**6. Bag of words:**

Bag of words is a Natural Language Processing technique of text modelling. In technical terms, we can say that it is a method of feature extraction with text data. This approach is a simple and flexible way of extracting features from documents.

A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a “bag” of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

One of the biggest problems with text is that it is messy and unstructured, and machine learning algorithms prefer structured, well defined fixed-length inputs and by using the Bag-of-Words technique we can convert variable-length texts into a fixed-length vector.

Also, at a much granular level, the machine learning models work with numerical data rather than textual data. So to be more specific, by using the bag-of-words (BoW) technique, we convert a text into its equivalent vector of numbers.

**7. Topic Modelling using Gensim's LDA:**

Topic Modeling is a technique to extract the hidden topics from large volumes of text. Latent Dirichlet Allocation(LDA) is a popular algorithm for topic modeling with excellent implementations in the Python’s Gensim package. The challenge, however, is how to extract good quality of topics that are clear, segregated and meaningful. This depends heavily on the quality of text preprocessing and the strategy of finding the optimal number of topics.

Can LDA be used for topic Modelling?

Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

The Work Flow for executing LDA in Python

After importing the required libraries, we will compile all the documents into one list to have the corpus.

We will perform the following text preprocessing steps (can use either spacy or NLTK libraries for preprocessing):

Convert the text into lowercase

Split text into words

Remove the stop loss words

Remove the Punctuation, any symbols, and special characters

Normalize the word (I’ll be using Lemmatization for normalization)

**8. Conclusion:**

* Loading the data
* Clean the Data
* Transforming text into something an algorithm can digest it a complicated process. We cannot feed the data as it is, some preprocessing needs to be done. In this task we will be doing some preprocessing to convert our data in a form that we can feed our model with.
* Handling the Stop-words
* Text may contain stop words like ‘the’, ‘is’, ‘are’. Stop words can be filtered from the text to be processed. There is no universal list of stop words in nlp research, however the nltk module contains a list of stop words. We will remove these stopwords in this task.
* Lemmatization
* TF-IDF Vectorization
* Apart from Count vectorizer an alternative to calculate word frequencies , and by far the most popular method is called TF-IDF. This is an acronym than stands for “Term Frequency – Inverse Document” Frequency which are the components of the resulting scores assigned to each word.
* Bag of words
* Topic Modelling using Gensim's LDA
* One of the drawbacks of LSA is that though it is really fast, its effectiveness in finding good topics is not great. One assumption that LSA makes is that the topics are orthogonal to each other, while Latent Dirichlet Allocation (LDA) relaxes this assumption. Moreover, LDA generalizes the way the documents are generated and this modelling assumption leads to better topics.
* Effectiveness of LDA Bag of words and LDA TF-IDF model in which LDA TF-IDF gave best probability of 57% and LDA using Bag of words giving 48% probability
* Effective modeling of topic for unseen document with 42% probability which helps news companies to connect underlaying patterns between topics and can identify similarities with in the topics