CSCE 5290: Natural Language Processing Project Proposal

Project Title:

Topic Modeling and Sentiment Analysis for movie review

Team Members:

Gayatri Annapurna Vadlani

Harsha Preetham Pukkalla

Harshitha Ambilpur

Rakshith Dyavari Shetty

GITHUB Link:

https://github.com/Gayatri345/CSCCE5290_ProjectNLP.git

Topic Modeling and Sentiment Analysis for movie review

Motivation:

The main goal of this project is to provide user the information about a movie whether it is good or bad and what people are talking about it. This is possible by getting the data by web scraping, cleaning the extracted data and then perform topic modeling to discover abstract topics that occur in a collection of documents so as to determine what people are thinking of a particular movie. We then perform sentimental analysis to determine if the particular document is positive or negative. When a user enters a movie name, it performs web scraping from the IMDb website and extracts the review data related to that movie name. It then performs data cleaning followed by topic modeling and sentimental analysis and fetches the user with the information if the movie is good or bad and also what people are thinking about movie.

Significance:

Movie reviews can be used as a data source for mining public opinion on a movie. In websites like Imdb we will have number of reviews provided for each movie. By analyzing these reviews and using text mining techniques such as topic modeling and sentiment analysis we can know about a movie. We apply topic modeling to infer the different topics of discussion(knowing the genre of a movie), and sentiment analysis is applied to determine overall feelings.

Objectives:

The data for this project will be web scraped from the IMDb website. This is done using the Beautiful Soup python package. Beautiful Soup is a python package used for parsing HTML and XML documents.

Topic modeling a method to perform unsupervised classification of collected documents and those similar to clustering on numeric data, which finds natural group of items even when we’re not sure what we’re looking for. In text mining, sources often have collections of various documents, such as blog posts/news articles, that we would like to divide into different groups to understand them separately.

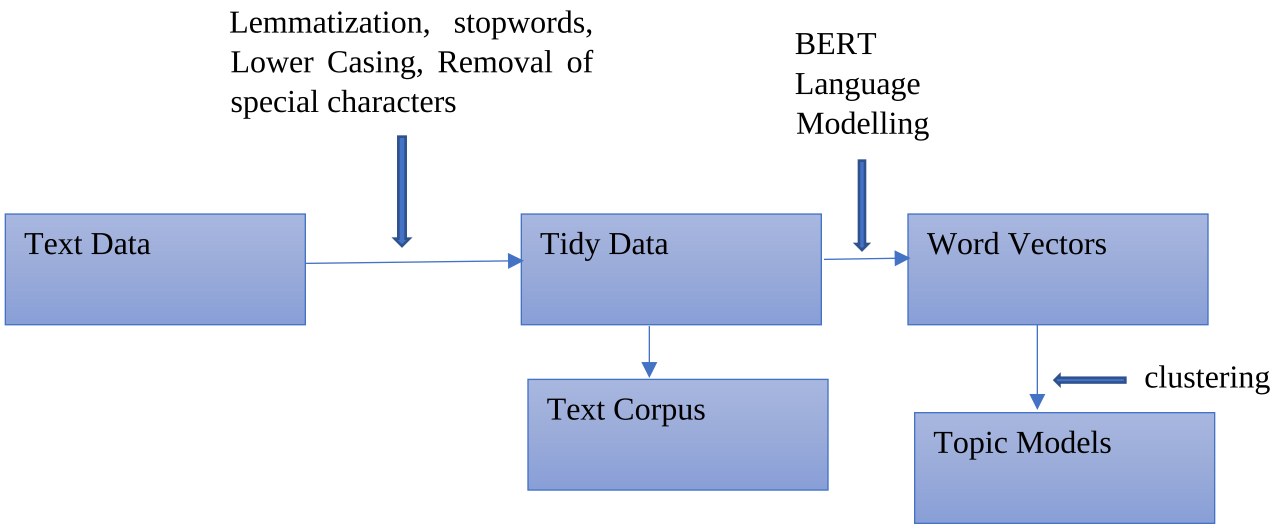

Fig.1. Flowchart for text analytics

We achieve topic modeling using BERT (Bidirectional Encoder Representations from Transformers) which main purpose is to extract embeddings based on the context of the word. We use clustering and TF-IDF to create topics and words occurring in topic.

Now, we perform sentiment analysis. It is a natural language processing technique used to determine whether data is positive, negative or neutral. This will give the information whether the movie is good or bad. For this we use unsupervised algorithm.

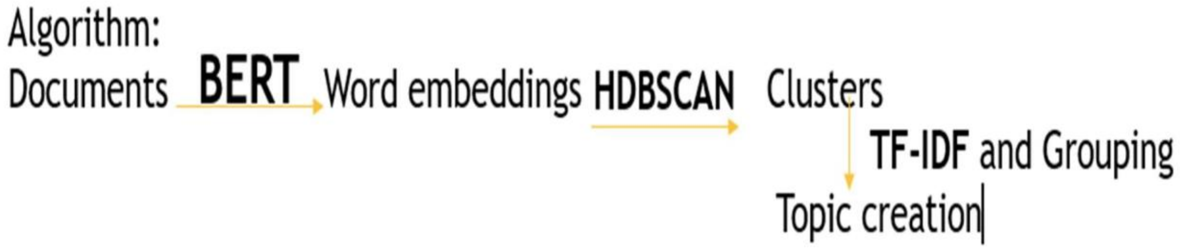

Fig.2 Flowchart of BERT

Features:

1. Web scraping the data from IMDb website.
2. Cleaning data free from stop-words, lowercasing, removing special characters, lemmatization.
3. Using BERT which is language modelling technique to generate word vectors.

4. Perform clustering so as to identify important topics in the set of word vectors created from the documents.
5. Perform unsupervised sentiment analysis with the word vectors created from BERT.

References:

1. https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6
2. https://www.pluralsight.com/guides/extracting-data-html-beautifulsoup
3. https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483