Topic Modeling using Text Visualization

Onkar Pednekar
Juhi Jadhav
Trish Bandhekar

Introduction

- Text visualization is a technique used to explore and understand the results of topic modeling analysis of a large corpus of text data.
- It identifies themes or topics in a group of documents. However, the results of topic modeling can be complex and challenging to interpret.
- Hence text visualization techniques, such as Word Clouds and WordTree, can help users understand the relationships between topics and subtopics and identify patterns that may not be easily apparent from the raw text data.
- The overall goal of the project is to examine and understand a large corpus of text data by identifying underlying topics and visualizing them using various text visualization approaches.

Components implemented in this project includes:

- **Preprocessing of the text data:** Stop words and punctuation, and numerals are removed and Tokenization and Lemmatization has been performed.
- **Text Visualization:** Text visualization approaches such as word clouds, wordtree, scatter plots, boxplot and barplot are implemented to view the outcomes of topic modeling.
- **Analysis and Interpretation:** Analyzed and interpreted the visualizations to obtain insights and identify patterns and trends in the data.
- **Evaluation:** Evaluated the success of the topic modeling and text visualization techniques using appropriate performance metrics.

Algorithm

- Our goal is to investigate topic modeling for text data using a variety of visualization methods.
- Tokenization involves splitting the text into separate words or "tokens" and is a crucial step in determining the distribution and frequency of certain words or phrases in the text.
- Convert the tokens to lowercase and remove any punctuations which were characterised as tokens to remove irrelevant data.
- Eliminating stop words like "the," "and," and "a" from the text is done using the NLTK library in Python.
- Lemmatization involves breaking down words into their parent or basic forms, and we compared the accuracy of various algorithms such as WordNet, WorNet + PosTag, TextBlob, TextBlob + PosTag and Spacy for this task. To do so we used a subset of Rotten Tomato Dataset.
- After preprocessing, we examined trends and patterns in the text data using visualization approaches like word clouds, word trees and interactive scatter plot..

Visualization and Interaction

To explore the relationship between the words, we created four visualizations that represent different aspects of the data.

- BoxPlot and Bar Chart
- WordTree
- WordCloud
- Scatterplot

Evaluation

- To evaluate the visualizations and the accuracy of our algorithm, we decided to go forward with the Quantitative User Opinion (QUO).
- The dataset used for visualizations is a small text blob on smartphones.
- We sent a google form to 7 users and asked them to answer a few questions.
- Based on the survey, we found that all users could easily identify the overall topic of the text corpus. The users also said that the wordcloud was the visualization that helped them identify the overall topic.
- The users could also identify patterns such as smartphones, entertainment and media. The users could also group words together to find relations such as communication, Smartphone, media, feature.

Analysis of Results

- We asked users, what the topic of the paragraph would be ?
- All users selected the “Smartphone” option.
- Another question asked was which visualization assisted them the most in determining the topic?
- 95% of the respondents chose wordcloud
- Lastly, we asked the users, whether the visualizations were effective in conveying the message
- All users responded positively.

Conclusion and Future Work

- The reviewed studies highlight the effectiveness of visualization techniques, including word clouds, box plots, and word trees, in analyzing and understanding large volumes of textual data.
- User evaluation suggests that the visualizations were found to be effective and helpful in understanding word relationships and overall topics.
- Future research could explore more advanced text mining and natural language processing techniques, such as sentiment analysis or named entity recognition, to gain deeper insights from the text data.
- Incorporating advanced machine learning algorithms could improve the accuracy and effectiveness of topic modeling.
- Other visualization techniques, such as heatmaps or network graphs, could be explored to better convey word-topic relationships.
- Future research could also extend these techniques to analyze other types of data, such as audio or video recordings.