Topic Modeling using Text Visualization

Onkar Pednekar University of
North Carolina at Charlotte
opedneka@uncc.edu

Juhi Jadhav
University of North Carolina at
Charlotte jjadhav@uncc.edu

Trish Vilas
Bandhekar
University of North
Carolina at Charlotte
tbandhek@uncc.edu

Abstract:

Modeling of text data using text visualization involves visualizations to explore and understand the results of a topic modeling analysis of a large corpus of text. In this project we propose a method which comprises preprocessing the text data, visualization of the words and phrases using plots like word clouds, scatter plots, and network graphs and topic labeling based on the findings from the visualizations. We suggest utilizing interactive scatterplot in addition to these visualization techniques to improve the understanding of text data. Users can explore the most common and pertinent terms in the corpus and interactively change the settings to sharpen their analysis. To obtain insights and identify patterns and trends in the data, the visualizations and interactions can be evaluated and understood. It can be difficult to determine the topics of text documents using solely text visualization for a variety of reasons such as ambiguity of words and phrases, noisy data, insufficient background information, intertwined topics and subtopics and subjectiveness of topics in text data. Finding the most important topics in a big corpus of text and examining the frequency and distribution of particular words in the data are possible results of topic modeling using text visualization.

Introduction:

Modeling of text data using text visualization involves visualizations to explore and understand the results of a topic modeling analysis of a large corpus of text. A statistical method called topic modeling is used to find themes or topics in a group of documents. Yet, topic modeling can produce complex and challenging-to-interpret results, particularly when working with huge text corpora. Text visualization can help with this by helping users understand the relationships between topics and subtopics and to identify patterns that might not be easily obvious from the raw text data. Text visualization techniques such as Word Clouds can be used to display the most common term with the size and color of each word denoting the frequency and significance of the word, Word Trees can be used to depict the connections between topics and the keywords that go with them. The graph consists if words and edges which shows phrases containing the keywords visually and shows how words are related to the keywords in the document. As a supplementary tool to assist users in exploring and improving their interpretation of text documents, interactive scatter plots will be used as part of our text visualization method. Using interactive scatterplot, you can select a word and examine the connections between keywords in a corpus of text. The user will have an option to select a keyword and the word tree will be plotted based on that keyword. These characteristics can help with the identification of potential topics and subtopics as well as the improvement of topic categorization based on information discovered through interactive exploration. This method's inclusion in the methodology for text visualization can further improve the accuracy as well as effectiveness of topic modeling analysis and offer meaningful insights into the underlying themes and topics in a large corpora of text data.
Following tasks can be expected:

- Preprocessing of text data: Stop words, punctuation, and numerals must be removed from the text data in order to make it more useful. Tokenization and Lemmatization of the input text data.
- Text Visualization: The outcomes of topic modeling can be viewed using a variety of text visualization approaches, including word clouds, word tree, scatter plots or heatmaps. Additionally,

we would be integrating interactive scatterplots to enhance the analysis of text data.

- Analysis and Interpretation: In order to obtain insights and identify patterns and trends in the data, the visualizations can be analyzed and interpreted.
- Topic Labeling: For ease of interpretation, the identified topics need to be given labels that are both descriptive and meaningful.
- Evaluation: The topic modeling and text visualization techniques' success can be assessed by contrasting the outcomes with actual data or by employing the right performance metrics.

This project's overall goal is to examine and comprehend a sizable corpus of text data by locating underlying topics and visualizing them using a variety of text visualization approaches.
Due to the numerous difficulties that may be encountered, topic identification in text documents using text visualization can sometimes be challenging. The ambiguity of text data, where a word or phrase may have different interpretations depending on the context in which it is used, is a significant difficulty. The fact that different readers may interpret the same text in various ways might further aggravate this uncertainty. As a result, figuring out a document's actual topic using solely text visualization might be challenging. The potential presence of noise in text data presents another difficulty. This can contain material which is repeated or redundant, which can

make it challenging to separate crucial data from unimpor- tant details. Text visualization can be used to reduce certain amounts of this noise, however it's critical to keep in mind that it won't always be capable of distinguishing between vital data and the noise. Moreover, visualizations may occasionally be lacking in the prior knowledge or context which is essential to fully comprehend a document's subject. This might result in a misconception or a lack of comprehension of the material, which can be problematic if the document is being used as the basis for significant choices or actions. Text data might also include a range of topics and subtopics that are interrelated and challenging to sort out, in addition to these difficulties. It may be difficult to effectively identify the primary topics and set them apart from related but unrelated concepts due to their complexity. Lastly, topic identification in text data might be difficult due to its subjectivity. Because various readers may have different interpretations of a single text, it could be challenging to properly communicate the heterogeneity of the data using simple visualizations. While utilizing text visual- ization, it is indeed essential to take such subjectivity under consideration and to utilize additional analytical techniques in addition to visualizations to obtain a more comprehensive understanding of the data.

A fundamental comprehension of the topics included in the text data could be obtained through topic identification using text visualization alone. However, because of the restrictions and difficulties associated with text visualization, it may not always provide a thorough or accurate representation of the un- derlying topics. The chosen visualization method, the quantity and quality of the text data, and the knowledge and expertise of the person evaluating the visualizations may all have an impact on the results. Text visualization can sometimes serve as a useful summary or overview of the material by drawing attention to the key themes that are present in the text data. The language employed or the frequency of particular phrases or words may also show trends or patterns that can provide light on the underlying topics. While there are undoubtedly limitations to utilizing text visualization to identify topics, there may also be advantages. For instance, patterns and connections between words and sentences that might not be immediately clear from reading the text alone might be found with the aid of visualizations. Moreover, visualizations can offer a quick and simple way for users to explore and navigate large amounts of textual data and quickly determine important themes and topics. Users can better understand the topics and ideas found in a big corpus of text by employing text visualization tools. These visualizations can be used to support decision-making processes in a range of sectors, such as social sciences, marketing, and business intelligence. They have the potential to help users find trends, patterns, and correlations within the text data.

RELATED WORK:

Text visualization has been an important area of research for extracting insights from a large volume of textual data, detecting trends, and highlighting main topics. Several studies have been done in this field with an emphasis on different visualization techniques such as word clouds, graph networks, and scatter plots. One of the studies by Shixia Liu et al [1] introduces a tool that combines text analytics with interactive visualization to summarize a set of documents and display the outcomes in a time-based visualization. Similarly, Chaney and Blei [2] suggest a technique for employing topic modeling to cate- gorize, summarize, and display large document collections. These studies aim to extract insights from a vast volume of text, detect trends and patterns, and highlight main topics.

Other studies such as S. Karpovich [3], Jiang, X and Zhang, J [4], and Saranya and Geetha [12] have focused on text visualization techniques for topic modeling analysis, with the purpose of aiding users in comprehending the organization of their text data. They employ a variety of visualization approaches such as word clouds, dendrograms, topic networks, word trees, and interactive elements like filtering, sorting, and clustering to allow users to experiment with various views and play with the data. Cui et al [5] proposes an algorithm for context-preserving visualization that allows users to understand the context of the words within the word cloud. In contrast, Endert, A., Fiaux, P., & North, C [6] presents a visual analytics approach for topic modeling called Semantic Interaction, which enables users to interact directly with the visualization, which in turn influences the underlying topic model.

El-Assady et al [7] introduces a novel method to combine topic modeling with a network graph visualization, repre- senting topic similarity as edges between nodes (topics). On the other hand, Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai [8] presents a novel approach to topic mod- eling, which incorporates network regularization to improve the effectiveness of traditional topic models. Other studies such as H. Ezzikouri, Y. Madani, M. Erritali, and M. Oukessou [9], F. Heimerl et al [10], A. Sarikaya and M. Gleicher [11], W. Zhu et al [14], and S. Sendhilkumar, M. Srivani, and G. S. Mahalakshmi [15] have explored different approaches for measuring semantic similarity between words, designing tools for text analytics based on word clouds, reviewing scatterplots, proposing algorithms for interactive word clouds, and determining the most important topics in a corpus of text.

In summary, the studies mentioned above have contributed to the development of text visualization techniques and ap- proaches that aid in comprehending and analyzing large vol- umes of textual data. These studies offer different visualization approaches such as word clouds, graph networks, and scatter plots, and have used these techniques to extract insights from textual data, detect trends and patterns, and highlight main topics. These studies offer a rich body of knowledge that can inform the development of future text visualization tools and approaches.

ALGORITHM:

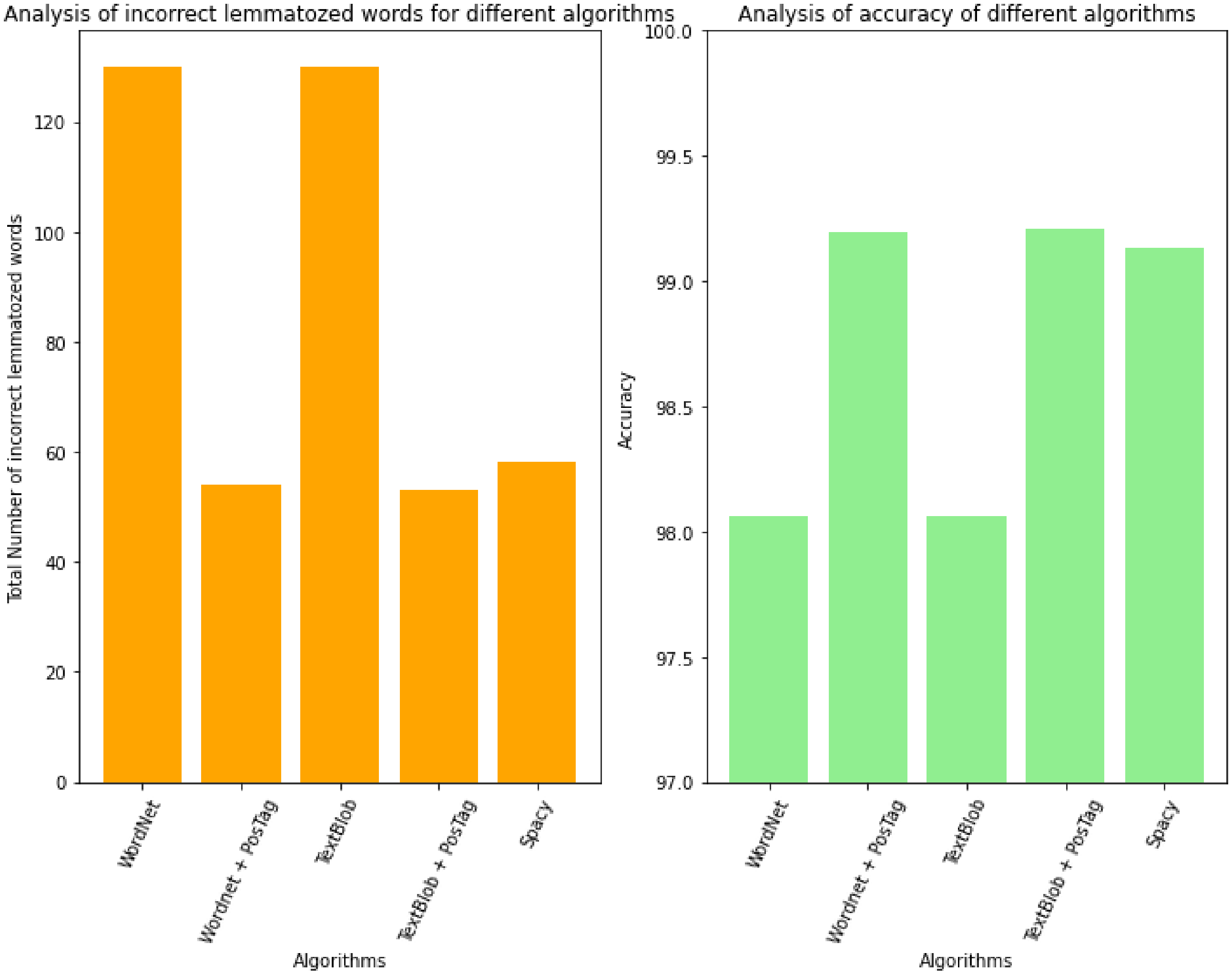
Our goal in this project is to investigate topic modeling for text data using a variety of visualization methods. To start, we'll tokenize the data, which involves splitting the text into separate words or "tokens" upon certain factors like whitespace. We can do this by iterating over the sentences and then appending the words separated by whitespace into an array. This is a crucial step because it allows us to determine the distribution and frequency of certain words or phrases in the text.

Next, we will eliminate words like "the," "and," "a," and other often used words that have no real significance from the text. To do so, we can use the NLTK library in Python. It has a built-in method stopwords.words(), which returns a list of English stop words. The stopwords.words() method can be used to filter out the stop words by dropping these words from the text data during preprocessing. By getting rid of such abstract terms, we can focus on the keywords that we believe are more crucial and more likely to convey the text's essential idea.

After the data has been preprocessed, we will perform lemmatization [9], which entails breaking down words into their parent or basic forms. First, we decided to compare the various algorithms that can be used for Lemmatization and compared their accuracy with a sample dataset. The algorithms used were:

- WordNet
- WordNet + PosTag
- TextBlob
- TextBlob + PosTag
- Spacy

The accuracy was as follow:



From the above figure, We can see that WordNet + PosTag & TextBlob + PosTag has the lowest amount of incorrect words and therefore the highest accuracy i.e. 99.2% We chose to go forward with WordNet + PosTag algorithm.

For this project we will use the WordNetLemmatizer() class from NLTK library where we'll pass each word along with its pos tag and get its base form. This process will enable us to combine several spellings of the same term and lessen the dimensionality of the data.

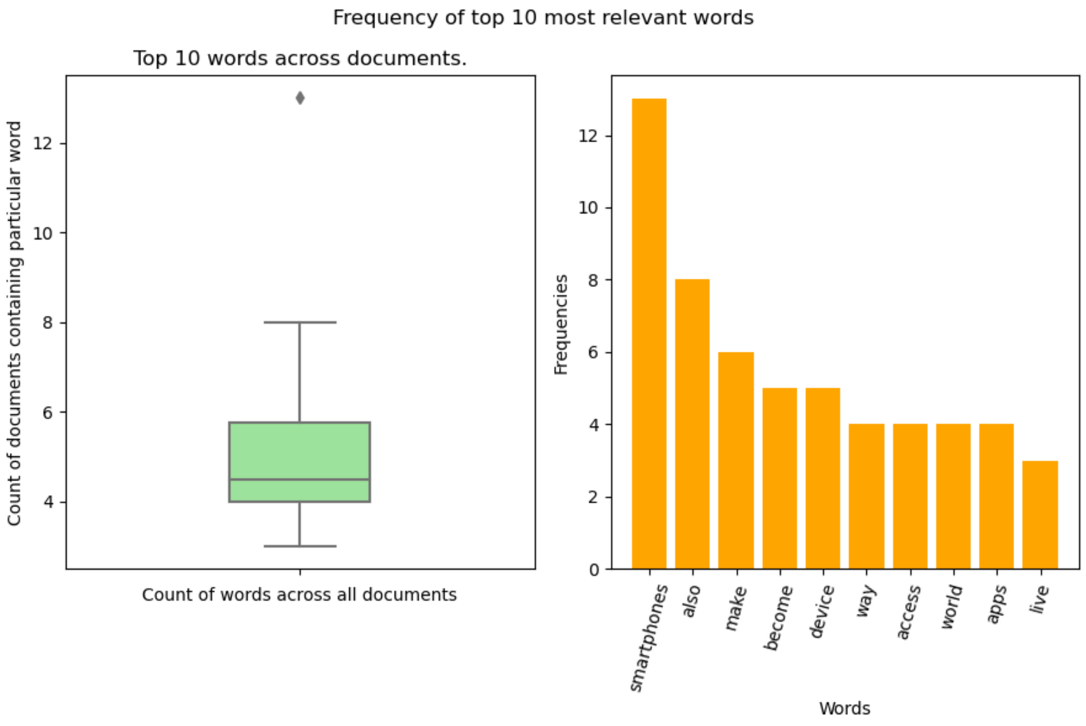
As soon as the preprocessed data is prepared, we will examine the trends and patterns in the text data using a variety of visualization approaches, including word clouds, word trees and scatter plots. We will be able to determine the most frequent terms or phrases in the text using these visualizations, as well as the connections between various words or topics.

Visualization and Interaction:

To explore the relationship between the words, we created four visualizations that represent different aspects of the data. The first visualization is a boxplot and a barchart. From this graph, we try to find the frequency of the words that are visualized.

Boxplots are used to display the distribution of a numerical variable. They show the median, quartiles, and any outliers in the data. The box itself represents the interquartile range (IQR), which is the range between the first and third quartiles. The line in the box represents the median. The ”whiskers” of the boxplot extend from the box to the minimum and maximum values that are not outliers. Boxplots are useful for comparing distributions across different categories and identifying potential outliers. They are particularly useful for identifying differences in the central tendency and spread of the data.

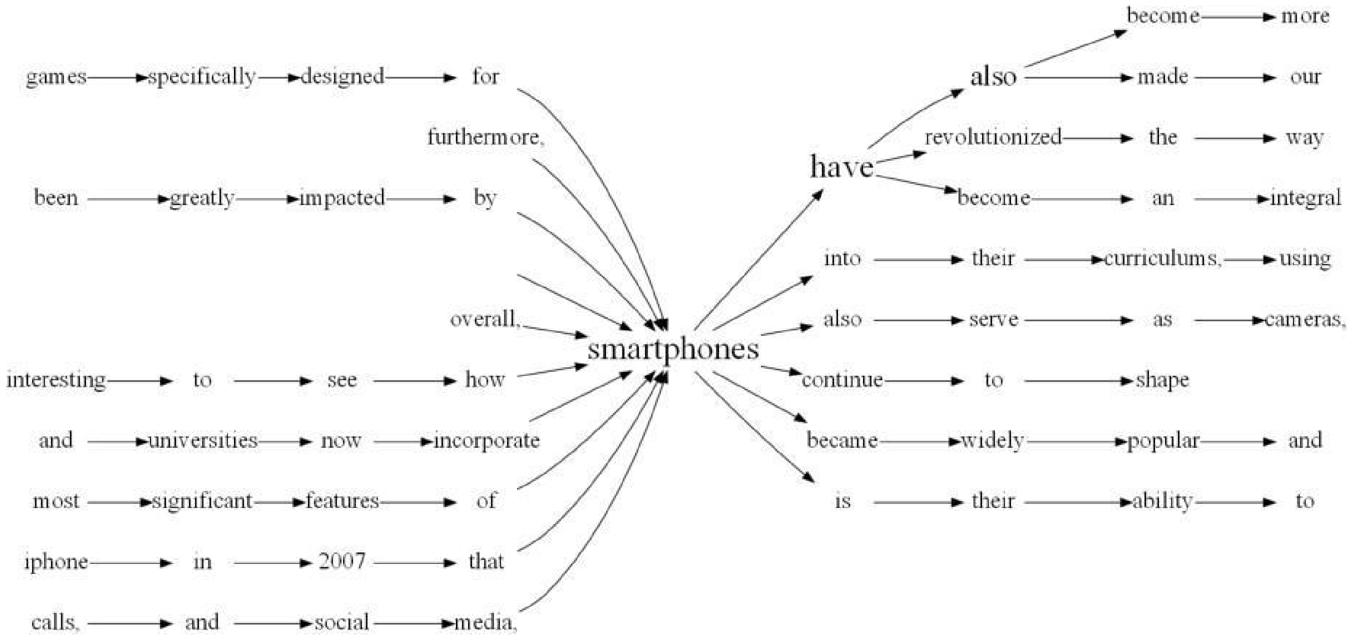
Barcharts represent the data using bars of equal width, where the height of each bar corresponds to the frequency or proportion of each category. Barcharts are useful for visualizing the relative frequency of different categories and iden- tifying any trends or patterns in the data. They are particularly useful for comparing categorical data across different groups or categories.



From the above fig., we can see that the word ”Smartphone” has the highest number of occurrences followed by ”also”, ”make” and ”become”.

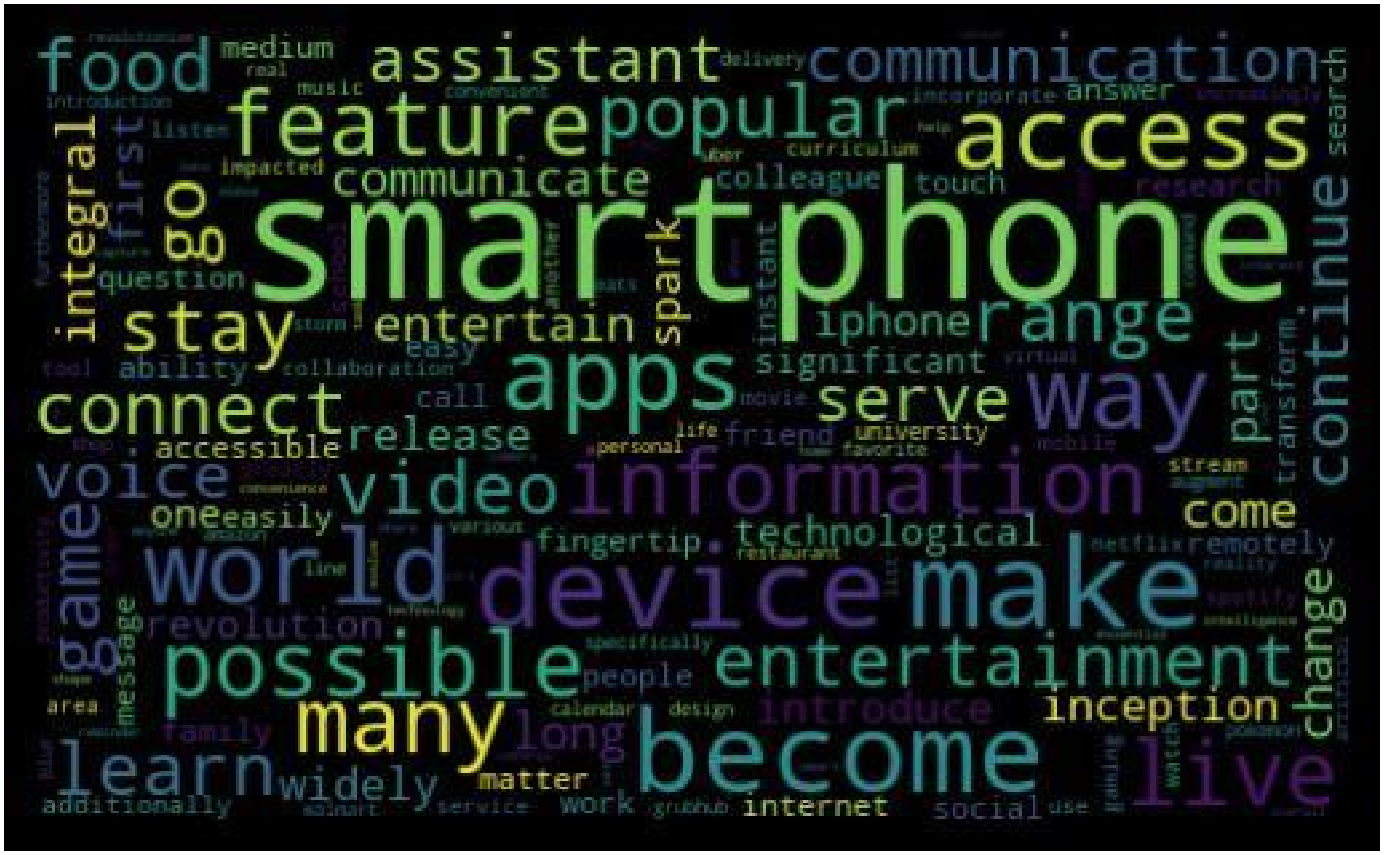
The second visualization we created was WordTree. The wordtree visualization tries to shows how words and phrases are related to each other in a corpus of text data. It is a powerful tool for exploring and analyzing large amounts of text data, and it is particularly useful for identifying patterns, trends, and relationships between words and phrases.

One of the main advantages of WordTree visualizations is that they allow users to see how words and phrases are used in context. This can help users to understand the meaning of individual words and phrases, as well as the relationships between them.



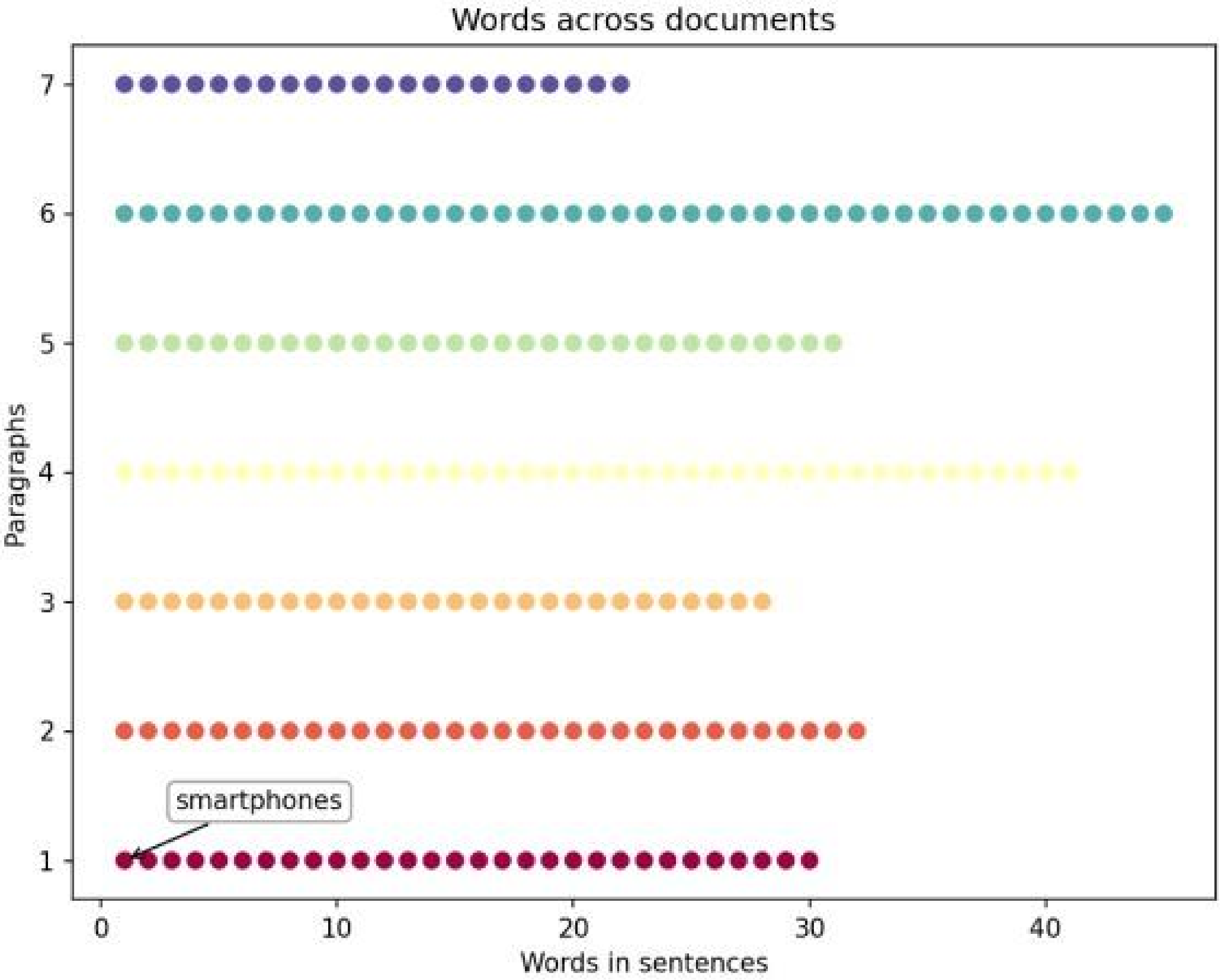
From the above fig., we can see that smartphones is at the center of the wordtree. The other words show how the words are related to each other. It also helps to identify underlying patterns and trends. For example, we can see how the word ”game” and ”smartphones” are related in the paragraph. It shows that games are being specifically being designed for smartphones. Another example is, how smartphones have gained popularity in recent times.

The third visualization we created was WordCloud. This visualization was particularly useful for providing a visual representation of text data, which can be difficult to process and understand when presented in traditional formats. By displaying the most frequently occurring words in a corpus, word clouds offer a way to quickly identify common themes and patterns within a dataset. This can be especially helpful in situations where large amounts of text data need to be analyzed and summarized quickly, such as in market research, social media analysis, and sentiment analysis. Wordclouds provide a quick overview and are eye-catchy and easy to understand.



From the above fig., we can see that the word ”Smartphone” is the largest in terms of font size. The fontsize and placement of the word signify the importance of the word from the corpus. Other important words that can be found are feature, access, possible. The placement also shows how closely each word is related to each other.

The fourth visualization we created was an interactive scatterplot. Interactive scatterplots allow users to explore and discover patterns, trends, and relationships within their data. By interacting with the plot, users can zoom in, pan, and hover over individual data points to obtain detailed information, enabling them to identify interesting clusters, outliers, or correlations that may not be apparent in a static plot.



From the above fig, we can see that smartphones has an occurrence of 30 in the first paragraph. On hovering over this visualization, we get to see the word that the data point highlights. When a user clicks on the data point, they can see the wordtree visualization based on that word.

EVALUATION:

• Data and Material

The dataset used for visualizations is a small text blob on smartphones.

• Procedure

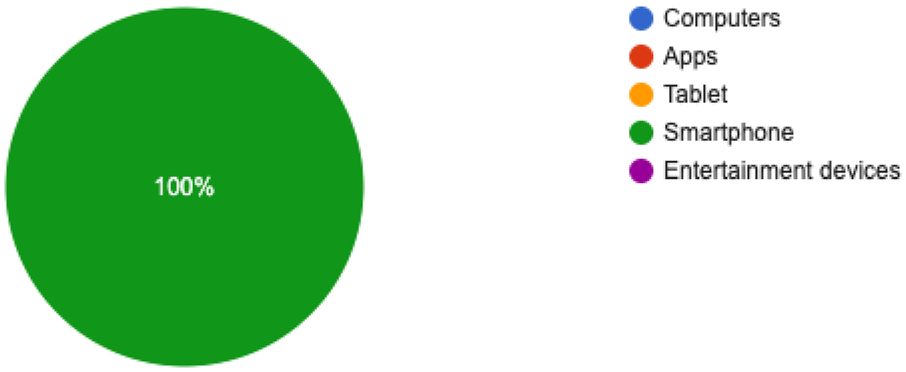
We sent a google form to 7 users and asked them to answer a few questions. To evaluate the visualizations and the accuracy of our algorithm, we decided to go forward with the Quantitative User Opinion (QUO).

• Questionnaire

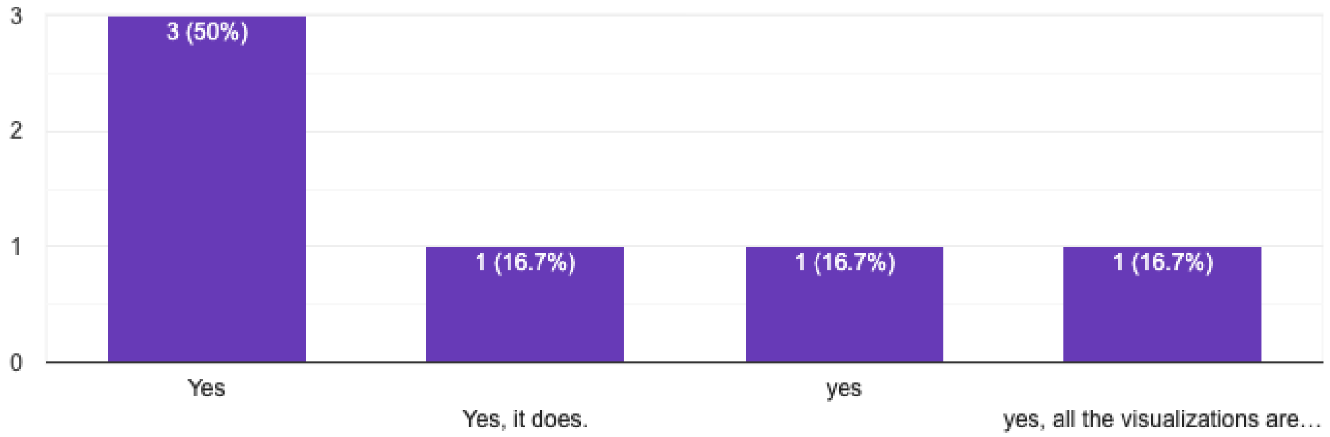
We asked users, what the topic of the paragraph would be ? All users selected the "Smartphone" option. Another question asked was which visualization assisted them the most in determining the topic? 95% of the respondents chose wordcloud. Lastly, we asked the users, whether the visualizations were effective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

• Analysis

The following graph shows the split of answers from the survey for the question what topic of the paragraph would be ?



The second question we asked was on the effectiveness of the visualizations. The users responded in the following manner:



From the survey, we can see 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, media, connect, Smartphone, media, feature. Overall, the visualizations effectively conveyed the overall structure or organization of the data.

CONCLUSION AND FUTURE WORK

In conclusion, the studies and research reviewed highlight the effectiveness of visualization techniques, particularly word clouds, box plots, and wordtree, in analyzing and compre- hending large volumes of textual data. These visualization tools enable users to identify trends, patterns, and relationships within the data, and provide a more engaging and accessible way of conveying complex information.

Based on the evaluation, it can be inferred that the users found the visualizations to be effective and helpful in un- derstanding the relationships between words and identifying the overall topic of the paragraph. This highlights the im- portance of incorporating visualizations in text analysis and visualization, as it enhances comprehension and provides a more engaging and interactive experience for users.

For future work, we can see that it does have information related to frequency but that alone is not sufficient to convey the overall structure. We can also consider different algorithms to show how similar or dissimilar the word associations are.

One possible direction for future research is to explore the use of more advanced text mining and natural language processing techniques, such as sentiment analysis or named entity recognition, to gain further insights from the text data. Additionally, the inclusion of more advanced machine learning algorithms could improve the accuracy and effectiveness of the topic modeling.

Another potential area for future work is to investigate the use of other types of visualization techniques, such as heatmaps or network graphs, to better convey the relationships between words and topics. Additionally, it may be beneficial to conduct a comparative study of different visualization techniques to determine the most effective approach for a particular use case or domain.

Lastly, while the current study focused on analyzing written text data, future research could explore the use of similar techniques and visualizations to analyze other types of data, such as audio or video recordings.

REFERENCES:

1. Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. 2012. “TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis.” ACM Trans. Intell. Syst. Technol. 3, 2, Article 25 (February 2012), 28 pages. https://doi.org/10.1145/2089094.2089101
2. Chaney, A., & Blei, D. (2021). Visualizing Topic Models. Proceedings of the International AAAI Conference on Web and Social Media, 6(1), 419-422. https://doi.org/10.1609/icwsm.v6i1.14321
3. S. Karpovich, A. Smirnov, N. Teslya and A. Grigorev, ”Topic model visualization with IPython,” 2017 20th Conference of Open Innovations Association (FRUCT), St. Petersburg, Russia, 2017, pp. 131-137, doi: 10.23919/FRUCT.2017.8071303.
4. Jiang, X., Zhang, J. A text visualization method for cross- domain research topic mining. J Vis 19, 561–576 (2016). https://doi.org/10.1007/s12650-015-0323-9
5. Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X., & Qu, H. (2010). Context- Preserving, Dynamic Word Cloud Visualization. IEEE Computer Graph- ics and Applications, 30(6), 42-53.
6. Endert, A., Fiaux, P., & North, C. (2012). Semantic Interaction for Visual Text Analytics. Proceedings of the SIGCHI Confer- ence on Human Factors in Computing Systems (CHI ’12), 473-482.https://dl.acm.org/doi/10.1145/2207676.2207741
7. El-Assady, M., Hautli-Janisz, A., Gold, V., Butt, M., Holzinger, K., & Keim, D. (2016). Interactive Visual Analysis of Transcribed Multi-Party Discourse. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1, 370-379. Link:https://www.researchgate.net/publication/318738992 Interactive Visual
8. Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In Proceedings of the 17th international conference on World Wide Web (WWW ’08). As- sociation for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/1367497.1367512
9. F. Li, L. Liao, L. Zhang, X. Zhu, B. Zhang and Z. Wang, ”An Efficient Approach for Measuring Semantic Similarity Combining WordNet and Wikipedia,” in IEEE Access, vol. 8, pp. 184318-184338, 2020, doi: 10.1109/ACCESS.2020.3025611. https://ieeexplore.ieee.org/document/9201502
10. F. Heimerl, S. Lohmann, S. Lange and T. Ertl, ”Word Cloud Explorer: Text Analytics Based on Word Clouds,” 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 2014, pp. 1833- 1842, doi: 10.1109/HICSS.2014.231.
11. A. Sarikaya and M. Gleicher, ”Scatterplots: Tasks, Data, and Designs,” in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 402-412, Jan. 2018, doi: 10.1109/TVCG.2017.2744184. https://ieeexplore.ieee.org/document/8017602/references#references
12. M. Wattenberg and F. Vie´gas, ”The Word Tree, an Interactive Visual Concordance” in IEEE Transactions on Visualization & Computer Graphics, vol. 14, no. 06, pp. 1221-1228, 2008. doi: 10.1109/TVCG.2008.172
13. M. S. Saranya and P. Geetha, ”Word Cloud Generation on Clothing Reviews using Topic Model,” 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0177-0180,

doi:10.1109/ICCSP48568.2020.9182111.https://ieeexplore.ieee.org/document/9182111

14. W. Zhu, J. Zang and H. Tobita, ”Wordy: Interactive Word Cloud to Summarize and Browse Online Videos to Enhance eLearning,” 2020 IEEE/SICE International Symposium on System Integration (SII), Honolulu, HI, USA, 2020, pp. 879-884,doi:10.1109/SII46433.2020.9026306.https://ieeexplore.ieee.org/document/9026306
15. S. Sendhilkumar, M. Srivani and G. S. Mahalakshmi, ”Generation of Word Clouds Using Document Topic Models,” 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, India, 2017, pp. 306-308, doi: 10.1109/ICRTCCM.2017.60.https://ieeexplore.ieee.org/document/8057554