Visualization and Analysis of Text Using NLP

Atharva A. Deshpande *from Department of Computer Engineering of TSSM's, BSCOER*
Saurabh J. Joshi *from Department of Computer Engineering of TSSM's, BSCOER*
Ritesh B. Kothawade *from Department of Computer Engineering of TSSM's, BSCOER*
Prajakta P. Patil *from Department of Computer Engineering of TSSM's, BSCOER*
Prof. Mr. Shubham Bhadre *from Department of Computer Engineering of TSSM's, BSCOER*

Abstract
Various tools for text mining and visualization are readily available in the market, empowering creative individuals to uncover valuable insights related to emerging technology. With the ever-increasing volume of data generated each year, there is a growing need to synthesize and extract knowledge from an expanding pool of literature. Time holds greater importance than money in the corporate world, making it a challenge to grasp the nature, purpose, and significance of the data. This research paper presents a case study that employs computational techniques like Natural Language Processing (NLP) for text analytics and data visualization. Often, this data exists in an unstructured text format, falling into the realm of big data, which requires analysis to extract meaningful information. The project utilizes Python libraries to process a vast amount of text and offers graphical visualizations along with text analysis operations such as Word Cloud, Mendenhall Curve, Tokenization, Graph, Processed Text, and Named Entity Recognition (NER). The study demonstrates the effectiveness of NLP-based text analytics in comprehending the data. The paper aims to optimize time and effort by providing a comprehensive assessment of the data and enabling data-driven decision-making. This approach allows users to simultaneously understand the data and improve grammar, which can be beneficial for meetings, analysts, teachers, and employees alike.

Keywords – Text mining, text analytics, visualization, NLP.

Introduction
Every organization relies on its data, as it forms the foundation for decision-making and enables competitive advantage. However, monitoring and accessing this data can be challenging, as it is often stored in backend databases or spreadsheets that are difficult to navigate. Natural Language Processing (NLP) offers a solution to this problem. By utilizing computer algorithms to analyze text, NLP allows us to extract valuable information that would be otherwise difficult to comprehend.

NLP, which is the process of understanding and interpreting human language, enables us to derive meaningful insights from vast amounts of data, thereby enhancing business decision-making. Text analysis is the technique employed to derive meaning from textual data. Through NLP, we can gain a deeper understanding of our audience's needs and preferences, providing them with the knowledge required to make informed decisions. Text analysis proves to be useful in utilizing text data within a business context.

Text visualization plays a vital role in presenting data in a visually appealing manner, aiding readers in better understanding and utilizing the information for decision-making. By using visual representations and images, text visualization enhances comprehension and highlights key details. We simplify the process of making informed business decisions by providing data insights and contextualizing important trends. Leveraging our NLP-based product, we maximize productivity and save time. Upon specifying your requirements, our product swiftly and comprehensively analyzes your data, uncovering crucial components that enable you to make intelligent choices. Supported by an intelligent platform, this technology assists in comprehending complex data and making better business decisions.

Motivation
This project aims to investigate advance NLP research. It is challenging to quickly abstract the data from large and unstructured data, so MNCs benefit from this project's quick understanding of the data. The reader typically reads the information rapidly when there is a time constraint and an important meeting is about to begin, and if it is not prepared swiftly, they are unable to match the points and understand the relationships between words. Text analysis and visualization can make it simple, rapid, and scalable for the user to understand qualitative data. Word cloud, which uses charts, will transform plaintext into captivating narrative. The Text Analyzer is the ideal tool for individuals since it not only summaries the data but also highlights the Nouns, Verbs, and Pronouns in the text, enhancing both readability and learnability.

Problem Definition
Comprehending and examining written information poses a major hurdle. Each day, a substantial volume of data is produced, and multinational corporations (MNCs) face the pressure of working within strict time constraints, often with spontaneous meetings. With limited time, employees cannot fully delve into all the data, leading to an indirect influence on decision-making. Similar circumstances arise in educational institutions, where students struggle to grasp the teacher's data and notes due to the inherent limitations of the human brain in quickly comprehending abstract concepts compared to visually depicted graphs and images.

An additional challenge arises from the human brain's limited capacity to retain information without summarization over a prolonged duration. This problem is particularly notable in countries like India, where English is not widely spoken, leading individuals to forget fundamental grammar elements like nouns, verbs, pronouns, and more. To tackle these issues, we have developed a project that facilitates rapid comprehension of data while simultaneously prioritizing learnability and readability for users.

Project Scope

1. Companies no longer need to have a pre-requisite detailed theory of data in order to respond quickly to decisions based on data.
2. Employees can use this to highlight key data segments, present data as a wordcloud and other visual formats during stakeholder meetings, and assist in data comprehension.
3. By using a visual format and minimal effort, this approach can be used in colleges and universities to teach students challenging concepts.
4. While reading, students can easily comprehend the content and refine their English language skills.

However, despite the existence of various natural language processing (NLP) tools in the market, none of them offer comprehensive features or exhibit a high level of quality. Multiple websites and mobile applications are available to support these features, but they often lack integration, causing inconvenience and time consumption for users who need to access different platforms for different purposes. Moreover, unwanted advertisements pose a significant obstacle as developers rely on them for monetization. Additionally, many systems restrict input options, prohibiting users from uploading files and limiting them to copying and pasting data within a restricted word limit, which can be frustrating. While numerous systems provide different features, there is a lack of a unified platform that encompasses all these capabilities. Some mobile applications allow text scanning through the camera, but the results are occasionally unreliable, providing only partial accuracy.

Related Work

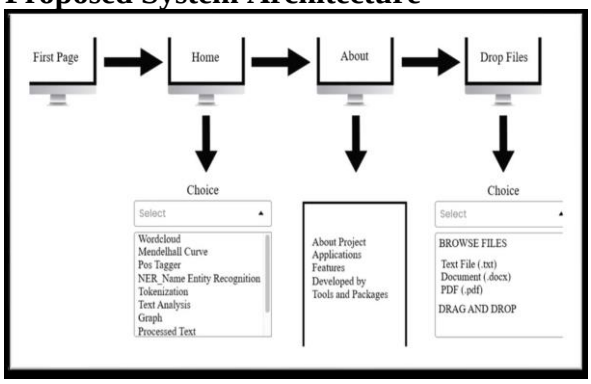
Significant progress has been made in the field of natural language processing (NLP), enabling people to make data-driven decisions more efficiently with the aid of advanced tools and implementations. Although some organizations are undertaking projects in the vast and occasionally complex realm of NLP, none of them have been adequately developed and designed to meet user needs. To address this issue, we are developing an impressive website that incorporates state-of-the-art NLP techniques. This project is being built using Python and leveraging Sublime and the Streamlit framework. We aim to provide users with a comprehensive platform that integrates essential functions such as text analysis, visualization, name entity identification, and part-of-speech tagging, all in one place. The motivation behind this project stems from the diverse features and outputs of existing NLP tools, which often lead to confusion and inefficiency when users have to rely on multiple programs for different tasks, resulting in a waste of time and resources.

Proposed System

In the input phase, users are provided with the capability to directly upload files up to 200 megabytes in size, including both text and PDF formats. This feature is essential as it allows users to obtain comprehensive insights from their data in a single step, eliminating the need for tedious module-based breakdowns or copy-pasting. Many existing systems impose word limits, which can lead to inaccurate results and time wastage. By enabling direct file uploads, this project addresses those limitations and enhances the user experience.

Moving on to the second phase, users are required to apply various operations to the data. These operations include generating word clouds, analyzing Mendelhall curves, utilizing Part of Speech taggers (POS), performing Name Entity Recognition (NER), tokenizing the text, conducting text analysis, generating graphs, and summarizing the text. These functionalities provide users with a wide range of options to explore, manipulate, and visualize their data effectively.

Proposed System Architecture



Future Work

The primary objective of this paper was to emphasize the significance of text analysis and visualization. Having recognized the system's pivotal role in the modern world, there is potential for further integration of this system with enhanced algorithms on large-scale datasets, specifically in enterprise settings, to facilitate swift data-driven decision-making. Looking ahead, there are plans to expand the system's capabilities by incorporating regional languages such as Hindi and Sanskrit. This expansion would enable researchers to extract valuable insights from materials written in these languages, thereby contributing to addressing contemporary challenges and issues.

Conclusion

The findings of the study strongly affirm the considerable benefits of employing natural language processing (NLP) techniques, particularly for individuals who engage in frequent reading and evaluation tasks. Notably, these techniques enable users to swiftly and succinctly communicate the results to administrators. The project effectively fulfills its objective of extracting pertinent information, generating visual representations, and presenting the data in a manner that yields valuable insights.

References

1. Sarthak J Shetty , Vijay Ramesh, " An open-source Python package for scientific
text analysis" , https://doi.org/10.1002/ece3.8098
2. Paritosh D. Katre, " NLP Based Text Analytics and Visualization of Political
Speeches" , ISSN: 2277-3878, Volume-8 Issue-3, 10.35940/ijrte.C6503.098319
3. Akshaya Udgave, Prasanna Kulkarni " Text Mining and Text Analytics of
Research Articles" , (ISSN 1567-214x), PJAEE, 17 (6)
4. John Risch, Shawn J bohn, Anne kao, Steve Poteet " Text Visualization",DOI
10.1007/978-1-4899-7502-7_837-1
5. Diksha Khurana, Aditya Koli, Kiran khatter & Sukhdev Singh " Natural language
processing: state of the art, current trends and challenges", 10.1007/s11042-022-
13428-4
6. YunYun Yang, Lucy Akers, Thomas Klose, Cynthia Barcelon, Yang "Text mining
and Visualization Tools [World Patent Information 30(2008) 280-293] BristolMyers Squibb, P.O. Box 4000, Princeton NJ 08543-4000 USA.