CS432/532: Final Project Report

Project Title: Analyzing the Evolution: Trends, Sentiments, and Styles in NY Times Reporting

Team Member(s): Janvi Bhalala, Sumeet Patil

I. PROBLEM

Examining the news landscape within the New York Times involves a comprehensive analysis of evolving trends and content over the years. This exploration extends beyond factual reporting to understanding the emotional responses of readers. By delving into the temporal evolution of news stories, we aim to uncover patterns, shifts, and sentiments that shape the narrative. The project's focus lies in discerning not only the factual progression of events but also the nuanced emotional undertones conveyed by the readership. Through this analysis, we seek to capture the dynamic nature of news coverage, offering insights into both content evolution and the emotional resonance with the audience over time.

We are utilizing a publicly available data set sourced from Kaggle for Analyzing News in New York Times Report. Our analysis involves three intricate queries on the dataset.

A. Analyzing the yearly frequency of keywords reveals the predominant trends and popular categories, offering insights into evolving themes and interests over time.

B. Examining reader comments provides a window into understanding the emotions expressed by the audience. Analyzing these comments offers valuable insights into the sentiments conveyed by readers.

C. The evolution of writing styles and journalistic preferences over time reflects dynamic changes within sport sections, showcasing nuanced shifts in storytelling approaches.

II. SOFTWARE DESIGN AND IMPLEMENTATION

A. *Software Design and NoSQL-Database and Tools Used*

1) MongoDB

2) Python

3) Modules Used

a) Pandas -We employed Pandas, a sophisticated Python toolkit with advanced data structures and extensive analytical functionalities, designed to streamline and simplify the process of data analysis.

b) Py-Mongo -For handling MongoDB, a document-oriented database, we leveraged PyMongo, a Python distribution providing an advanced interface for managing and manipulating data within MongoDB databases. It supports various database operations, including find, insert, update, delete, and aggregation.

4) Jupyter Notebook, Visual Studio Code and MongoDB Compass

5) JavaScript, CSS and HTML is used to make Interactive User Interfaces.

B. Analysis

1) Analysis 1:

We can identify the yearly trends and most popular categories by examining the frequency of keywords from 2000 to 2023. This approach reveals the most popular topics and themes in stories, giving readers a sense of the country's mood. The dynamic storyline of events and societal shifts in the USA over each period is reflected in the popularity of particular topics in The New York Times. By dissecting the language used in news stories, one can discover the topics that caught the public's interest as well as historical context, illuminating key events and cultural trends that influenced speech across time.

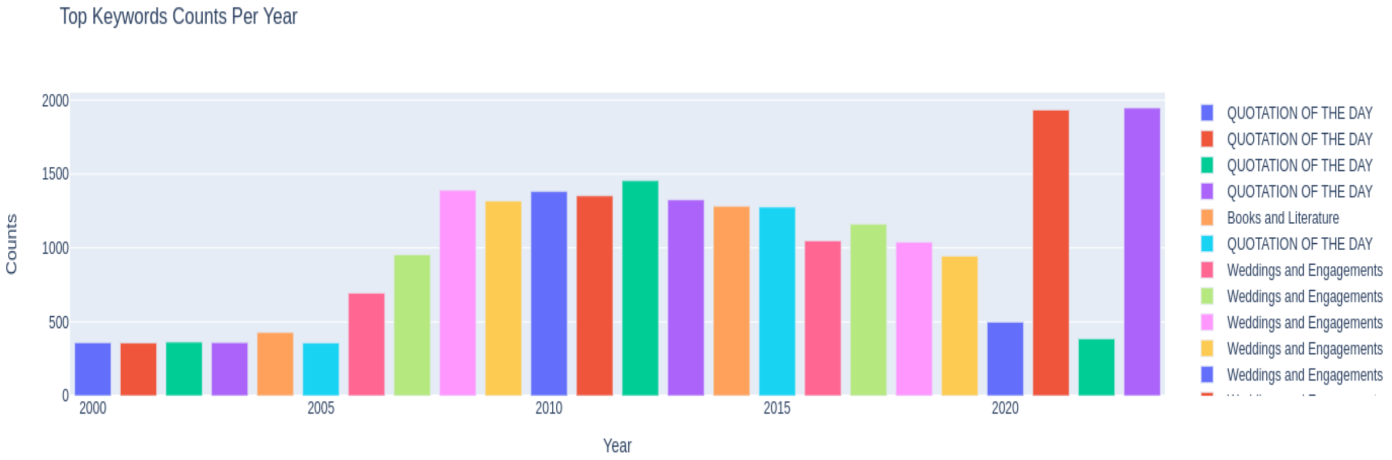


Fig. 1. Analysis of keywords .

2) Analysis 2

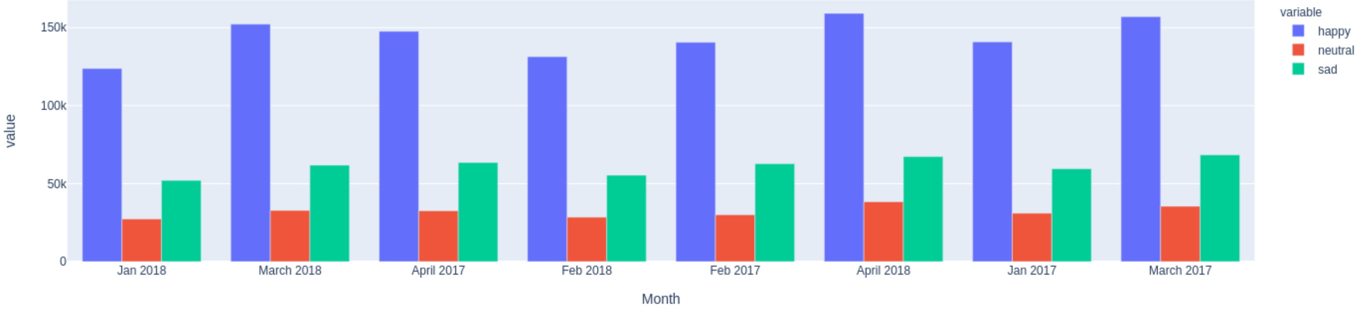
The analysis delved into the sentiments expressed in the comments on NY Times articles. Using Natural Language Processing (NLP) techniques, sentiments were categorized into "happy," "sad," or "neutral." The sentiment analysis aimed to understand the emotional tone of reader comments over time.

Data from the 'comment months' collection in the 'nytimes' MongoDB database was retrieved, encompassing comments on NY Times articles. The dataset included information such as comment text ('commentBody') and the corresponding date ('Date'). The TextBlob library in Python was employed to perform sentiment analysis on each comment. Sentiments were categorized into three groups: "happy," "sad," or "neutral" based on the calculated sentiment polarity.

A month-wise aggregation of sentiments was conducted, tracking the frequency of each sentiment category. The results were stored in the 'monthcomments' collection in the 'nytimes' MongoDB database. The analysis generated a detailed breakdown of sentiments for each month, providing a nuanced understanding of the emotional responses associated with NY Times articles. The aggregated results were stored in a structured format, including the month, count of "happy" comments, count of "sad" comments, and count of "neutral" comments.

3)

Fig. 2. Analysis of User’s Comments



3) Analysis 3

The analysis focused on unraveling the dynamic trends within the Sports section of the NY Times, examining the evolution of top keywords over the years. The process involved complex queries, data preprocessing, and insightful visualizations.The data was retrieved from the MongoDB collection 'article' within the 'nytimes' database, specifically targeting articles categorized under the 'Sports' desk.The publication dates were standardized to datetime format, and a new 'year' column was created for temporal analysis.

A safe literal evaluation was applied to handle potential errors in the 'keywords' column, allowing for subsequent expansion. The 'keywords' were expanded to discern their individual values, and a new column 'keyword' was generated based on dictionary values. The analysis involved grouping the data by 'year' and 'keyword' to calculate the frequency of each keyword. The top keywords per year were identified, and the results were sorted in descending order to highlight the most prominent terms. The findings were stored in the 'top_keywords_per_year' collection within the 'nytimes' database. A comprehensive visualization was generated, illustrating the trends in the top keywords by year within the Sports Desk.

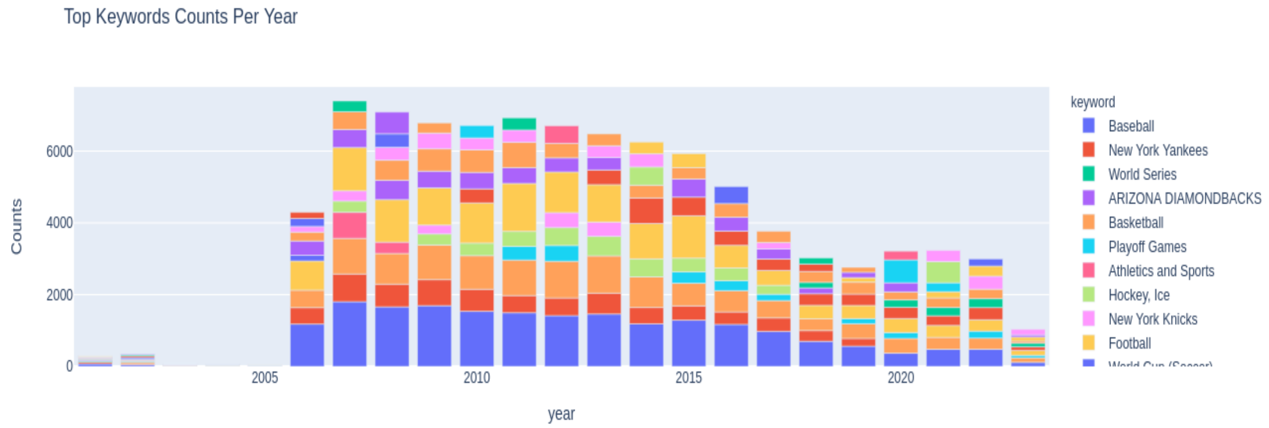


Fig 3. Analysis of evolution of articles in sports section

References

[1] https://www.kaggle.com/datasets/aryansingh0909/nyt-articles-21m-2000-present

[2] https://www.kaggle.com/datasets/aashita/nyt-comments

[3] https://www.mongodb.com/docs/