News Recommender for JhakaasNewsVala: Strategy and Implementation Plan

Puneet Singh Bhooi, Karanjot Singh Saggu

(Team 🚀🚀🚀)

Sabudh Foundation

1 Exploratory Data Analysis

Following is the analysis of the news articles' readability and insights that can be drawn from them. Flesch Ease Test score and Flesch-Kincaid Grade Level score try to measure the readability of a piece of text using Average Sentence Length (in the number of words), ASL, and Average number of Syllables per Word, ASW. The number of syllables in a word is an approximation and not an exact count.

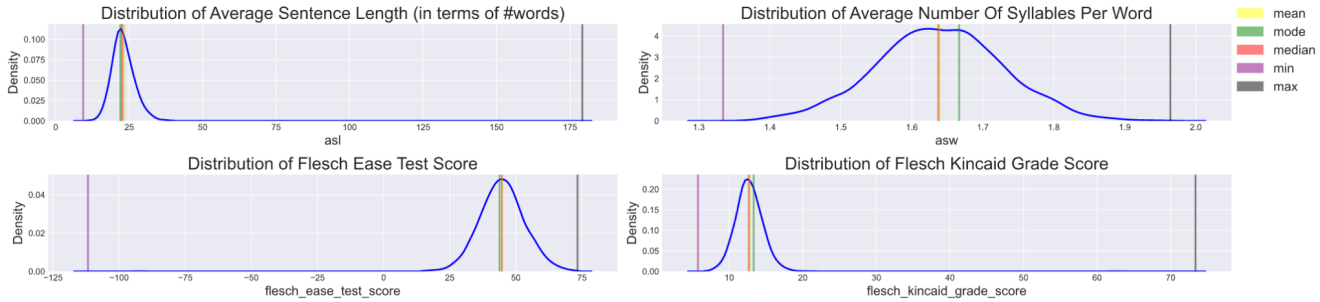

Figure 1. ASL has a skew of 11.87. ASW has a skew of 0.02; the mean and median of this variable almost coincide. Flesch Ease Test score has a skew of -3.04, Flesch-Kincaid Grade Level score has a skew of 9.11.

News article with the least Flesch Ease Test score (most readable) says the following, "A man who built a raft to save a swan's nest of eggs was "determined" to help her[...]". And the article with the most Flesch-Kincaid Grade Level score (least readable) talks about the Bafta TV Craft Awards nominations and mostly consists of celebrity names. This evaluation of readability, along with the length of the article, can be used in generating the time spent by the users on reading a news article.

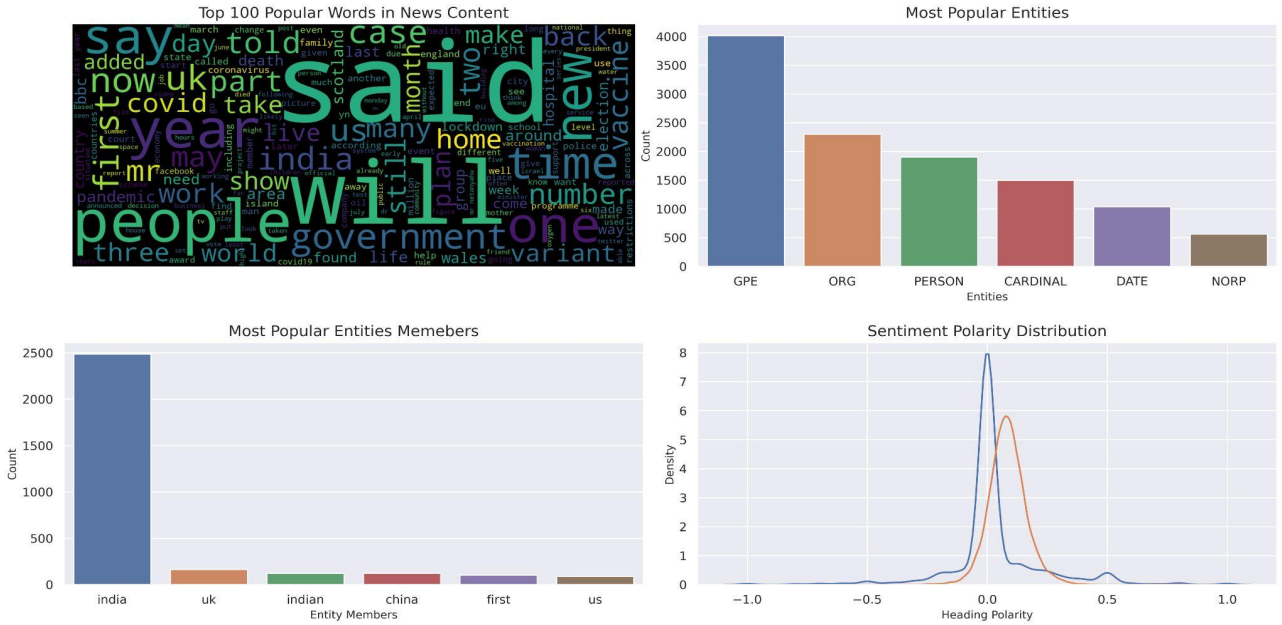

Figure 2. Word Cloud; Entity tag distribution; Entity member distribution; Polarity distribution

The above plot represents the semantic analysis of the dataset. The top left subplot presents the traditional word cloud of news content. The top right plot gives an overview of the most popular categories of entities present, similar to the bottom left plot. The button right plot shows the polarity of the text headings as a kernel distribution function.

2 Strategy and Implementation Plan

2.1 Data Scraping

The news articles are scraped from the following news websites using Selenium for navigating through the website and Beautiful Soup and Requests for scraping the data:

1. BBC News (https://www.bbc.com/news)
2. Times of India (https://timesofindia.indiatimes.com)
3. Yahoo News (https://in.news.yahoo.com)

2.2 Generate Clickstream Data

The clickstream data will include features like userID, visitID, itemID, and interaction metrics which capture the engagement of the user with the said item, e.g.: whether the user clicks on the recommendation or not. The engagement can be captured using proxy metrics like time spent by the user on the article, depth of scrolling through the webpage, etc. The motive of the recommender system(s) will be to maximize the frequency with which the user opens the app to consume stories and increase the clickthrough rate. Once user-item interaction data has been generated, similar users based on their common news preferences, and behaviour can be found.

2.3 Content-Based & Collaborative Filtering Recommender System

The Content-Based Filtering Recommender System is the first type of Recommender System that we can build with only a few user preferences. The article text, title (headline), and, additionally, tags, images, and other articles' metadata can be used to find similarity between the articles that can be recommended to the user. This approach, though fast, will soon overfit the user preferences in the train data. To take a step forward as the clickstream gets populated with different users, we can use clickstream data to capture the interaction between different users and items hence Collaborative Filtering comes into action. Algorithms like SVD, NeuMF, and other matrix factorization algorithms will be used to capture negative and positive interaction of an item with a user based on the other users.

In the next section, we present a technique that will provide better results that not only works on a subset of item metadata but all user-item interaction data as well.

2.4 Hybrid Recommender System

Reflecting on the limitations of both Content-Based and Collaborative Filtering Recommender Systems we wish to tackle the problem by using a combination of both techniques to get the best of both worlds. The Hybrid Recommender System will combine recommendations from both techniques using a deciding function that will dynamically rank recommendations on the go.

2.5 Metrics

To evaluate the Recommender Systems, metrics like ARHR (Average Reciprocal Hit Rank), Precision@k, Recall@k will be logged. Since the real estate available to display the news articles is limited to 10 articles only, k can be chosen as 10.