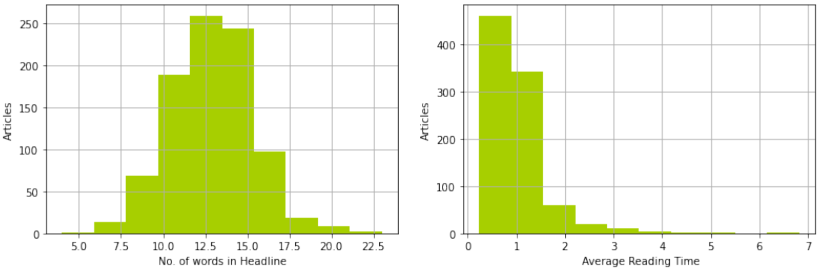
Strategy Plan For Jhakass Newswala

Chhavi(MS18136), Amlan(MS18197),
Smriti (MS18142), Rimjhim(MS18133), Dheer(MS18140)

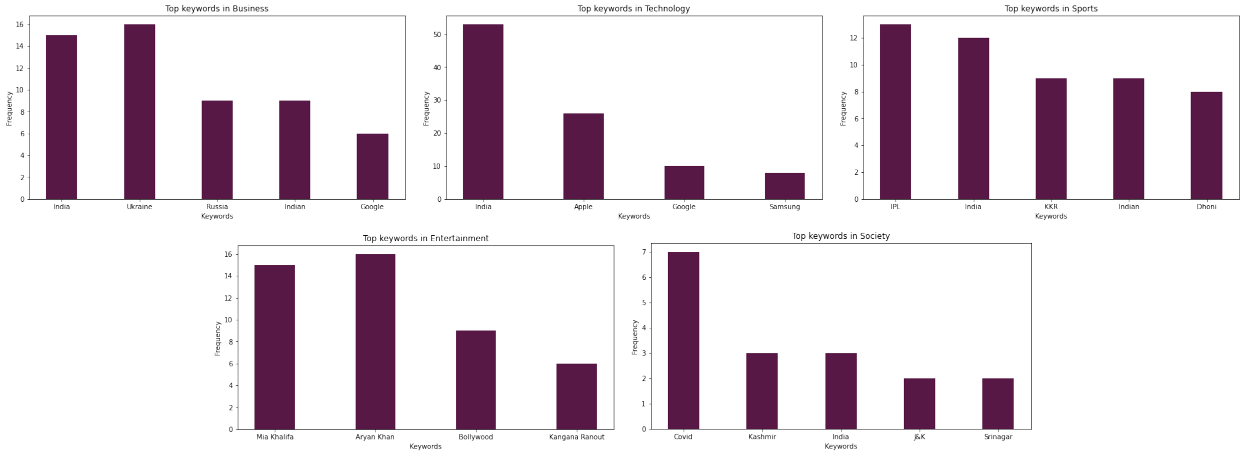
1 Research and Exploratory Data Analysis

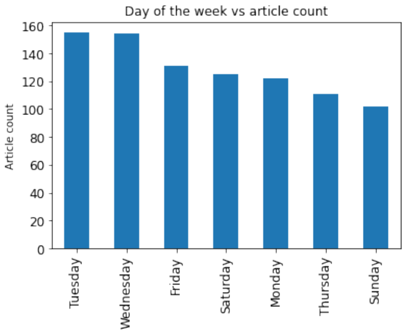
To do Exploratory Data Analysis, we performed web scraping on the website of International Business Times to get articles from different topics. The code is attached along with the report. We did the following analysis on the scrapped data.

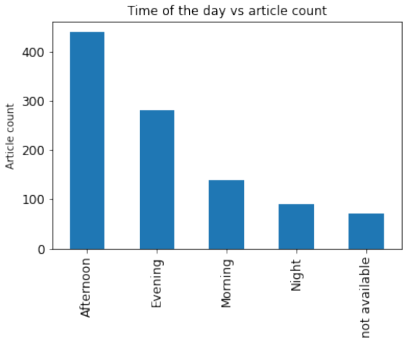
We got the 1,020 articles and the following number of articles for the given categories : technology(240), entertainment(240),society(180),business(180),sports(180). Each had 60 articles for every subcategory.



1) We tried to plot the Number of Articles vs Number of words in Headline for the data. We observe that the graph is a Gaussian distribution. From this, we can conclude that there is a mean length of headline for a randomly picked article in a pool of articles that this might also to relevant to user preference of picking an article to read.
2) Since we have a word count per article and considering each person reads at the speed of 250 words/minute we plotted the expected reading time of articles. This can be used further to make appropriate article length recommendations to user.
3) This is a plot of recurring keywords in each category across all articles. This can be used to further make the trending category shown to 1st time users.
4) We plotted Number of articles written on a certain time of the day and day of the week. This information can be used to hypothesise which day or time of the day might experience most number of readers due to large inflow of articles.







2 Strategy

2.1 Pre-processing and vectorizing

We are planning to form a gigantic corpus of web scraped news articles. We will track the dates, category, headline, article content and links. We also want to incorporate likes and no. of comments on each article if possible.

Next, we will use spacy library to execute standard filtering of article text as taught in class. Raw data obtained will go through text cleaning involving stemming, lemmatization, stop words removal and lowercasing. After this, the documents will be vectorized through a pipeline of processes involving presence & frequency of words, TF-IDF, LDA, hashing and Word2Vec as taught in class depending on time availiability.

We want create both hashing and embedding models, benchmark, and then choose the data processing approach.

2.2 Hybrid approach for content recommendation

We plan to use both - content based filtering and collaborative filtering for this project. Content based recommendation is based on the weighted average of articles consumed (explicit feedback) while also incorporating implicit feedback such as the time spent on each article as compared to the expected time. The model will recommend articles based on the similarity coefficient between user profile and news repository, becoming more personalized as the user consumes the recommendations leading to enhancement of increasing retention time. For first time users, we plan to serve top rated and trending stories from various categories and track their engagement time with each article. This data will be used for further recommendations. For collaborative filtering, we plan to find similarity between the users and compare the weighted average of similar user profiles with the news dataset. The articles will be recommended on the basis of cosine similarity between the weighted user profile and the news repository.

3 Plan for personalisation

3.1 Reduce bias in data collection

We plan to display most trending news stories across categories to the user. If the user skips a particular story, the entire category will be scrapped from the feed to avoid displaying trending similar-stories which disinterest the user. We can use registered IP of the user to display local trending news. This reduces a category of stories relevant to the user and provides more personalised experience to increase engagement. Instead of showing top stories from every subcategory, we show trending news from every subcategory, recommending more articles on the same category based on the engagement.

3.2 Learn as much as possible about the users on their first visit

The user is asked to choose from a list of topics after logging in to understand user subject preferences. Final user preferences are decided on the basis of time of engagement with a particular article. If a user prefers surfing over in depth reading, varied topics in subcategories will be recommended, depending on the expected reading time from our EDA. Contrasting news is displayed to user on first visit and their reaction is collated with final like dislike decision of the user to eliminate categories that are not liked.

3.3 Maximise coverage of the news corpus

We propose to include a ”Discover” section recommends articles out of their users' interests. Based on the engagement from this section, new type of articles will be included in the field. In user's usual feed, we plan to put a few articles per every 10 articles. Depending on the response, the frequency of these articles will increase and viewed stories will be added to the usual feel, adding to both personalisation and increase of news corpus.

This report has been made with equal contributions by every member.