News Recommendation System

1. Research --

Timeline followed -

a) 14 Sep - Group Discussion and jotting down the main points for the dataset.
b) 15 Sep - Dataset formation- Preprocessing.
c) 16 Sep - EDA on the data collected
d) 17 Sep - Final Documentation

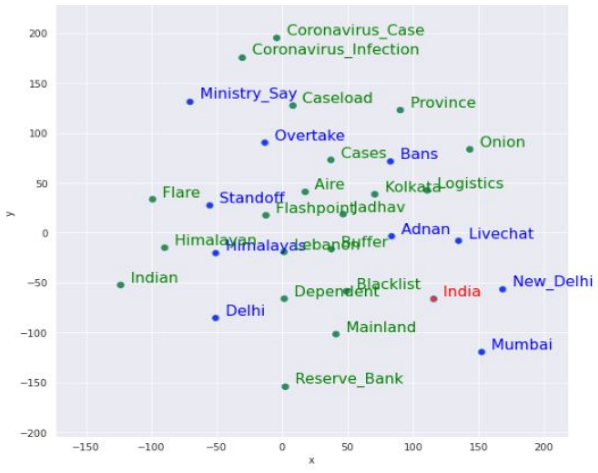
Articles read -
ttps://www.datacamp.com/community/tutorials/recommender-systems-python,
https://blog.insightdatascience.com/news4u-recommend-stories-based-on-collaborative-reader-behavior-9b049b6724c4,
https://www.offerzen.com/blog/how-to-build-a-content-based-recommender-system-for-your-product,
https://github.com/vivekpabani/News-Recommendation-System

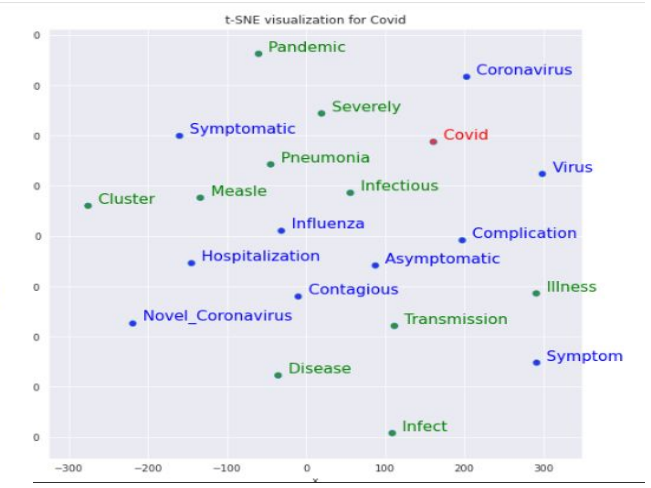
Data sources - We created our own dataset. As there was no dataset available which was up to date. We used https://newsapi.org/ , to extract links of various articles and used **"newspaper3k"** for extracting **title,dates,author,summary,keyworks etc.**
We are also trying https://pypi.org/project/news-fetch/ to extract data.

2. Exploratory Data Analysis --

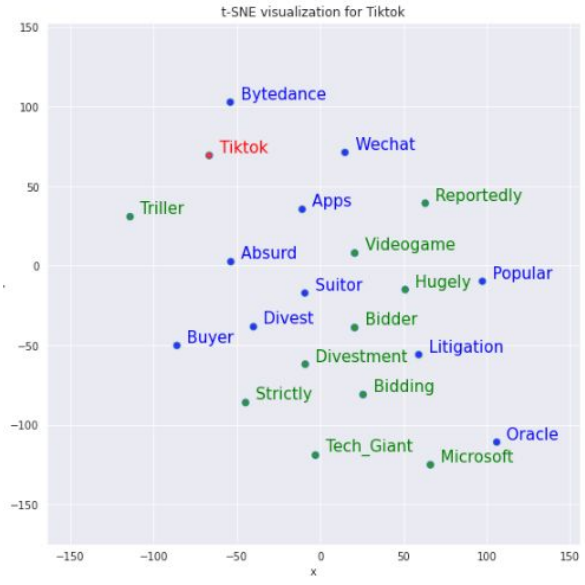
What do you look for -- correlated data, similar phrases, and
Conclusions drawn -- rich corpus , meaningful, bigrams, trigrams, similar tags .

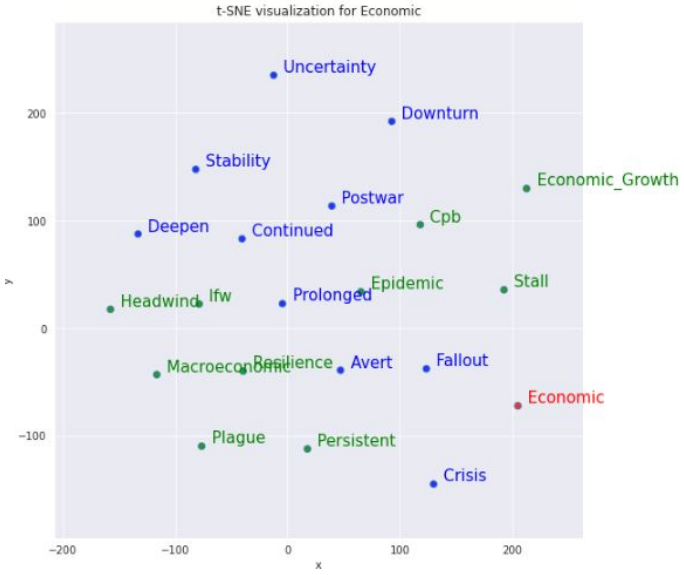
To make the visualizations more relevant, we will look at the relationships between a query word (in **red**), its most similar words in the model (in **blue**), and other words from the vocabulary (in **green**).



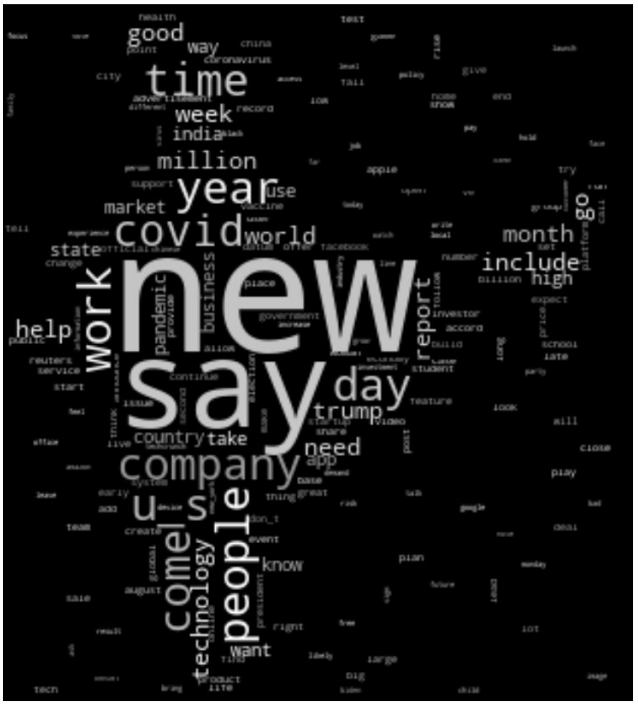


These are 2 examples to show how rich the corpus is that if we type just one word like "India" or "Covid" it is able to show similar/ correlated data. This is required as these words will be used as keywords to define a document and later used for whichever recommender system architecture we choose for the current problem statement.





The above data shows that the dataset has also covered variety of news articles.



This is a Word Cloud that shows 100 most similar words to India. And the probabilities of the selection is --

```
w2v_model.wv.most_similar(positive=["india"])
```

```
[('new_delhi', 0.7648840546607971),
 ('bans', 0.7335999608039856),
 ('delhi', 0.7161003351211548),
 ('adnan', 0.7041406035423279),
 ('overtake', 0.6986569166183472),
 ('himalayas', 0.6860007643699646),
 ('mumbai', 0.6801972389221191),
 ('standoff', 0.6796436905860901),
 ('livechat', 0.6768172383308411),
 ('ministry_say', 0.6754418611526489)]
```

3. Build Strategy -- The following steps were and are going to be followed while building this project.

1. Research - Detailed study of recommender Systems and input data required.
2. Pseudo-Code -,dataset formation, Pre-processing, EDA, finding meaningful conclusions. And finalizing the dataset/corpus to be used.
3. Research on what will be the model/ architecture will be deployed(in-Progress) .