# News Recommendation System

# Team: Likith Kumar, Anirudh Jilla, Shashank Sai, Vishwak Subba Reddy

Reading the news online has exploded as the web provider’s access to millions of news sources from around the world. These volumes of articles can be overwhelming to readers. Therefore, building a news recommendation system to help users find articles that are relevant and most interesting to them.

# Strategy and Implementation:

The strategy adopted in order to build the news recommender system-



**Web Scrapping & Data Accumulation:**

We will be using Beautiful Soup & Selenium packages for scraping the data from the news websites.

Our news corpus will contain following attributes:

1. News id
2. News Title
3. News Content
4. Date of publication
5. Link

# Types of Filtering Techniques:

There exist 3 different types of filtering techniques that can be used in our news recommendation system:

* Content based filtering
* Collaborative filtering
* Hybrid filtering

**Collaborative filtering:** Collaborative filtering for recommender systems is a method that is based on the past interactions recorded between users and their referred articles to produce new recommendations. The main idea is that the past user-article interactions are sufficient to detect similar users and/or similar articles and make predictions based on these estimated features.

However, the recommendation system using collaborative filtering recommends news articles that are not chronologically catalogued, i.e. the news articles being referred could also be a year old and hence their content might be irrelevant to the contemporary times. This would affect the knowledge of users and create a faulty image of the system in the user’s perception. Moreover, this technique only considers user’s rating and past-user interactions as the only parameters to recommend a particular article to others. This would mean that for recent articles the system would face a severe problem as it would not be able to recommend due to low user interactions and ratings, this is referred to as a cold start problem.

**Content based filtering:** Content based filtering uses additional information from users and articles over plain user interactions in their recommendation mechanism. The basic idea of a content based method is to build a model based on the available features, that explain the observed user-article interactions. It would consider features such as age, gender, job and any other personal information along with the article, published date and time, author and sources. However, content based filtering suffers a lower cold start risk factor compared to the collaborative technique as recent articles can be recommended to the users based on their characteristics and hence every article is recommended to minimum users at the least.

# Approach adopted:

We have selected on 2 different approaches to achieve this project:

* **Approach-1**

Using tf-idf and cosine similarity to recommend our user Top-K articles. Assuming the user has at least read one article.

* **Approach -2 (Doc2Vec + LSH)**

Hash items such that similar items go into the same bucket with high probability. Restrict similarity search to the bucket associated with the query item.

From the above given approaches we agreed upon Approach -2 rather 1. The primary reason for our decision is that the Word2Vec embedding in Doc2Vec allow the capture of the context of words in relation to other words of the document. Also it allows to calculate the weighted sum of vectorized representation of words in a document to create a document 2 vector representation.

Moreover the primary reason to reject approach 1 is due to its higher time complexity , i.e. O(n) which is not good for large corpus, whereas LSH has a O(logn) time complexity.

# Inference:

We use User based Recommendation system rather than Item based recommendation system as:

The initial stages do not accommodate the User Similarity index due to lower data in terms of users and the user-article interactions. Hence the cosine and adjusted cosine similarities will not be efficient enough to predict which users are similar due to sparsity in data. Also the item based approach is equipped with the means to calculate the similarity score unlike the previous case.

1) Initially we have thought of using both Doc2Vec algorithm for vectorization and LSH algorithm for classification of the articles this would have allowed us to easily convert and catalogue the articles based on their vectors and eventually the recommendation system can access similar vectors and suggest articles to the user. Also, the implementation of Doc2Vec + LSH would have given us a time complexity of O(log n) but our inability to do so has now left us with a time complexity of O(n)[here n represents the number of articles in the corpus because we calculate the distance between the user vector and vectors of all the articles existing in the corpus.

2) Currently we have used content-based filtering in order to recommend articles to the user in spite of being suggested to use Hybrid filtering that would have allowed for a better article recommendation process. This is because our attempts at trying to use Hybrid filtering had been unsuccessful as the model would most likely give different sets of recommendations than the preferences of the user hence we had to eventually drop the idea due to time constraints.

3) If we had more time, we would have tried to reduce the time complexity by using the Doc2Vec and LSH with Hybrid Filtering rather than the ones we have used currently.

4) The theory that had been referenced during the development of this project had led us to presume that the use of Hybrid Filtering was more profitable in terms of recommending better articles to the user and could be easily consolidated into the programming part. But during the development of the model we had encountered many errors, failed test cases and had realized that the implementation was not as direct as the theory had led us to presume, we had realized that we require deeper insights into the Filtering technique in order to compile it into the recommender system.

5) The primary idea was to use the Hybrid Filtering Technique in the recommender system, we were able to program the Content Based Filtering into the model but in case of Collaborative Filtering as it was previously mentioned we had encountered many errors, failed test cases and eventually were not able to successfully implement it into the model.