**NEWS ARTICLE RECOMMENDATION**

**MITHUN RAAM M – 2018103562**

**NAREN T P – 2018103568**

**SANJAY CHINNI KARTHICK V – 2018103586**

**INTRODUCTION:**

A news article recommendation is an application used to provide similar news articles for users based on their history. This is achieved through lexical analysis; it recommends articles on various authors and categories. The project “ NEWS ARTICLE RECOMMENDTATION “ is developed in python programming language, which mainly focuses on Natural Language Processing (NLP) by recommending users the right articles at the right time. The steps involved are creating the dataset and giving it to the system and by using lexical analysis we tokenize into vector forms, and evaluate and compare each vector by Euclidean similarity and recommend news headline.

**RELATED WORKS:**

1. News recommendation based on collaborative semantic topic models and recommendation adjustment.

* **METHODOLOGY:**

Collaborative filtering such as Matrix Factorization is a typical way to extract the implicit factors for predicting the preference ratings of users.

1. Personalized News Recommendation Based on Collaborative Filtering.

* **METHODOLOGY:**

Three approaches for personalized news recommendation: collaborative filtering at the level of news items, content-based system recommending items with similar topics, and a hybrid technique.

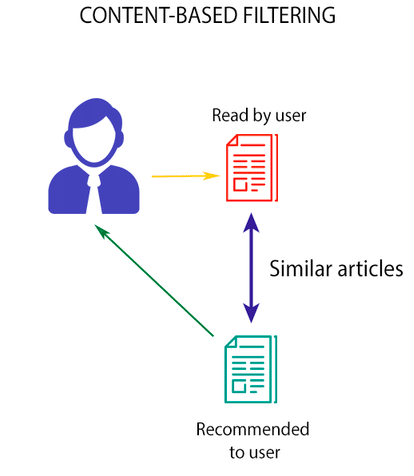
1. A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering.

* **METHODOLOGY:**

Measure the similarity between documents is tousle the terms within the documents to represent them as vectors and measure the similarity among them based on the angle or Euclidean distance between each pair.

**SYSTEM ARCHITECTURE:**

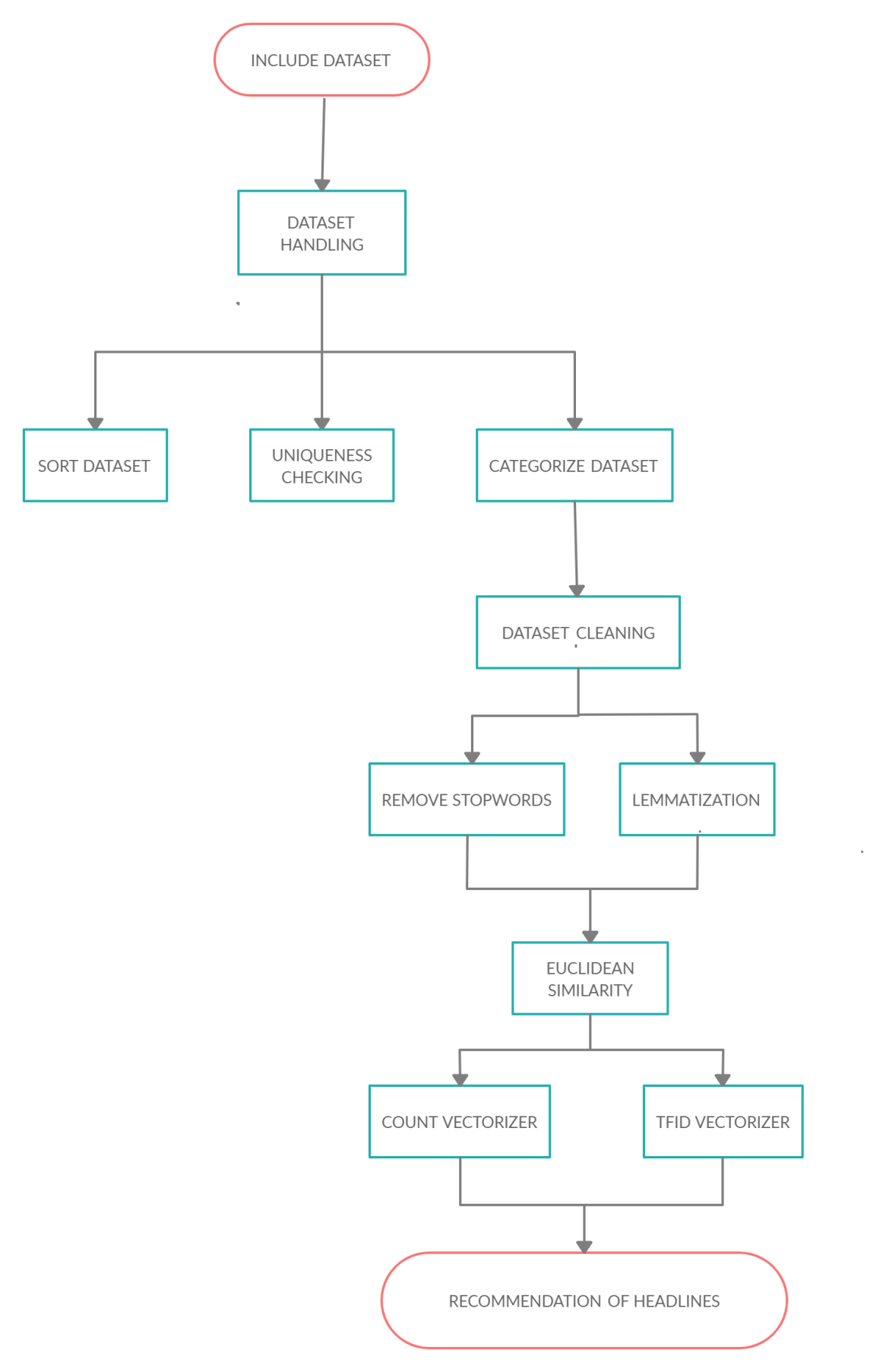
**APPROACH USED:**



**CONTENT BASED FILTERING:**

* This model doesn't need any data about other users, since the recommendations are specific to a user. This makes it easier to scale it to a large number of users.
* This model can capture the specific interests of a user, and can recommend niche items that very few other users are interested in.

**FLOW DIAGRAM:**



**MODULE 1:** **DATA HANDLING**

* **Include Dataset**
* The dataset which we have used here is the BCC News Dataset.
* **Phase 1: Sort dataset**
* Headlines having less than 5 words are sorted from the

dataset.

* The dataset has also been sorted using timestamp function.
* **Phase 2: Uniqueness Checking**
* Initially, the dataset is sorted in descending order for uniqueness checking.
* Final step, is removing duplicates from the dataset.
* **Phase 3: Categorize dataset**
* The number of unique categories are displayed.
* Then the number of headlines in each category are listed.

**MODULE 2: DATA CLEANING**

* Using strftime function we are converting time into desired and required format.
* The dataset is copied to another set for later use.
* **Phase 1: Removing Stopwords**
* First we import **“nltk”** library to download.
* It provides such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

* We loop through the headlines word by word by splitting and comparing the stopwords and removing them.
* **Phase 2: Lemmatization**
* We loop through the headlines word by word and using word\_tokenize we group the different forms of word and convert into rootword.
* At last, using strip function, the leading and trailing spaces are removed.

**Module 3: Euclidean Similarity**

* We have used Euclidean similarity to compare two lists of numbers (i.e. vectors), and compute a single number which evaluates their similarity. Most measures were developed in the context of comparing pairs of variables across cases. In other words, the objective is to determine to what extent two variables co-vary, which is to say, the same values for the same cases have.
* **CountVectorizer** - used to convert a collection of text documents to a vector of term/token counts. It also enables the pre-processing of text data prior to generating the vector representation.
* **TfidfVectorizer** - Transforms text to feature vectors that can be used as input to estimator. vocabulary\_ Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index.In each vector the numbers (weights) represent features tf-idf score.
* **Phase 1: CountVectorizer**

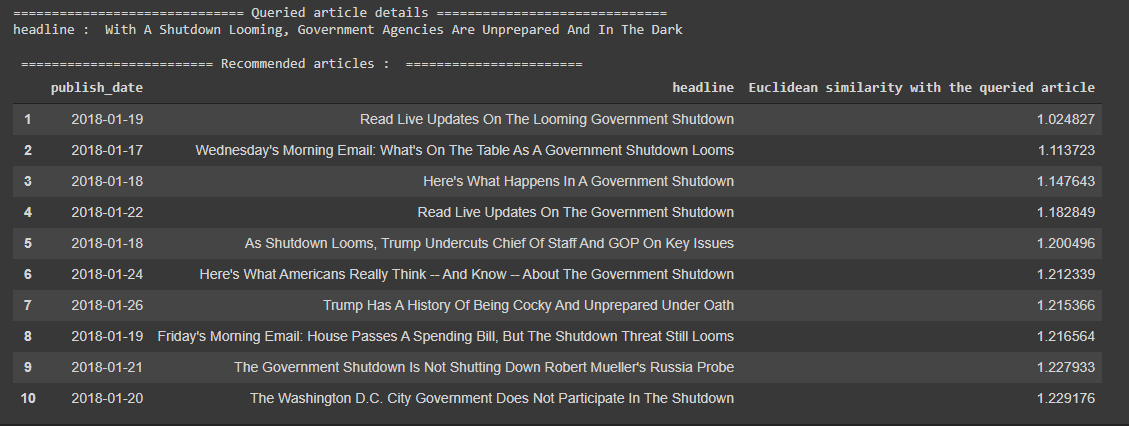
* Count Vector is used to transform the given text into a vector form on basis of the frequency of each word occurs in the entire text.
* It creates a matrix in which each unique word is represented by a column of the matrix and the each text sample from

the data is the row in a matrix.

* **Phase 2 : TfidVectorizer**
* TfidfVectorizer is used to convert a collection of raw documents to a matrix of TF-IDF features.
* The TfidfTransformer transforms a count matrix to a normalized tf or tf-idf representation.
* Although both the CountVectorizer and TfidfTransformer (with use \_idf=False) produce term frequencies, TfidfTransformer is for normalizing the count.

**RESULT AND DISCUSSION:**

For a given headline as input, similar headlines are recommended.



First, we have included the dataset and sorted and cleaned it. Then, a headline from the dataset is given (input is given as: headline number, number of recommendations for that headline). From the given input, the headline is fetched and it is vectorized using countvectorizer and tfidvectorizer (a matrix is created in which each row is a headline and each column represents a word in the dataset and each cell represents the count of each word in the dataset). Eventually, based on euclidean similarity (difference between the vectors) the similar headlines for the given input are recommended. The headline which has a minimum difference is given as first recommendation and so on (number of recommendations to be displayed is given as the input initially).

**FINAL SUMMARY:**

A person generally reads a news article, the application will analyse the headline of that article. Before creating any feature from the raw text, we must perform a cleaning process to ensure no distortions are introduced to the model. We have followed these steps:

* **Special character cleaning**
* **Up case/down case**
* **Punctuation signs**
* **Possessive pronouns**
* **Lemmatization**
* **Stop words**

The idea for the above steps is to tokenize the sentence, and proceed further. During the vectorization process, it will compare the given articles with the data set provided to the system. Eventually, the articles are recommended based on user’s history.

**CONCLUSION:**

With roughly a thousand news articles published by journalists at BBC News every day, it is a challenge to bring the right articles to the right readers at the right time. So, to provide right articles to the readers, we have recommended similar articles to them (readers) based on their interests and reading history.

**REFERENCES:**

1. Y. Liao, J. Lu and D. Liu, "News Recommendation Based on Collaborative Semantic Topic Models and Recommendation Adjustment," 2019 International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, 2019, pp. 1-6, doi: 10.1109/ICMLC48188.2019.8949259.
2. F. Garcin, K. Zhou, B. Faltings and V. Schickel, "Personalized News Recommendation Based on Collaborative Filtering," 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, 2012, pp. 437-441, doi: 10.1109/WI-IAT.2012.95.
3. A. Heidarian and M. J. Dinneen, "A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering," 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, 2016, pp. 142-151, doi: 10.1109/BigDataService.2016.14.
4. C. Feng, M. Khan, A. U. Rahman and A. Ahmad, "News Recommendation Systems - Accomplishments, Challenges & Future Directions," in IEEE Access, vol. 8, pp. 16702-16725, 2020, doi: 10.1109/ACCESS.2020.2967792.
5. Yuqi Wang and Wenqian Shang, "Personalized news recommendation based on consumers' click behavior," 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, 2015, pp. 634-638, doi: 10.1109/FSKD.2015.7382016.