**CSE4077 - Recommender Systems**

***J Component – Project Report***

***Article Recommender System***

*By*

19MIA1001 Roshan Srinivaas

19MIA1003 Niranjan J

19MIA1093 Vignesh.N

M.Tech CSE with Specialization in Business Analytics

*Submitted to*

**Dr. A.Bhuvaneshwari,**

Assistant Professor Senior,

SCOPE, VIT, Chennai

**School of Computer Science and Engineering**



*November 2022*



*BONAFIDE CERTIFICATE*

Certified that this project report entitled “**Article Recommender System**” is a bonafide work of Roshan Srinivaas (19MIA1001), Niranjan J(19MIA1003) and Vignesh N (19MIA1093) who carried out the J-component under my supervision and guidance. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for award of any degree or diploma and the same is certified

**Dr.A.Bhuvaneswari,**

Assistant Professor Senior,

**SCOPE, VIT, Chennai**

**Abstract**

In the last decade we have observed a mass increase of information, in particular information that is shared through smartphones. Consequently, the amount of information that is available does not allow the average user to be aware of all his options. In this context, recommender systems use a number of techniques to help a user find the desired product. Hence, nowadays recommender systems play an important role. Recommender Systems aim to identify products that best fits user preferences. These techniques are advantageous to both users and vendors, as it enables the user to rapidly find what he/she needs and the vendors to promote their products and sales. As the industry became aware of the gains that could be accomplished by using these algorithms, also a very interesting problem for many researchers, recommender systems became a very active area. Having in mind that this is an ongoing problem the present project intends to observe the value of using a recommender algorithm to find users likes by observing his/her domain preferences in news article. In this project we will show how news topics can be used to recommend news articles. In this project, we use content-based filtering technique to determine the user ratings for an article.

# Acknowledgement

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr.Bhuvaneshwari** mam AssistantProfessor, School of Computer Science Engineering, for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We express our thanks to our HOD **Dr.Sivabalakrishnan** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the school for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

Roshan Srinivaas (19MIA1001)

Niranjan J (19MIA1003)

Vignesh.N (19MIA1093)



**School of Computing Science and Engineering**

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

FALL SEM 22-23

**Worklet details**

|  |  |  |
| --- | --- | --- |
| Programme | M.Tech Integrated Business Analytics | |
| Course Name / Code | Recommender Systems / CSE4077 | |
| Slot | E1 + TE1 | |
| Faculty Name | Dr.A.Bhuvaneshwari | |
| Component | J – Component | |
| J Component Title | Article Recommender System | |
| Team Members Name | Reg. No | Roshan Srinivaas | 19MIA1001 |
| Niranjan J  Vignesh. N | 19MIA1003  19MIA1093 |

**Team Members(s) Contributions – Tentatively planned for implementation:**

|  |  |
| --- | --- |
| *Worklet Tasks* | *Contributor’s Names* |
| Preprocessing | Niranjan, Vignesh, Roshan |
| Model building | Niranjan, Vignesh, Roshan |
| Visualization | Niranjan, Vignesh, Roshan |
| Technical Report writing | Niranjan, Vignesh, Roshan |
| Presentation preparation | Niranjan, Vignesh, Roshan |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Title** | **Page no.** |
| 1 | **Introduction** | 7 |
| 2 | **Literature survey** | 8 |
| 3 | **Data set and tools** | 10 |
| 4 | **Algorithms and Techniques used** | 11 |
| 5 | **Discussions on results** | 13 |
| 9 | **Conclusion** | 15 |
| 10 | **Implementation Screenshots (if any)** | 16 |
| 11 | **Github repository** | 23 |
| 12 | **References** | 24 |

**Introduction**

It is highly difficult and time-consuming to find the desired item of choice fast due to the easy availability of enormous things (services) on our favorite online platforms like e-commerce, e-articles, e-newspapers such as ‘Amazon Audible’. Content based recommendation using **News category** dataset, the goal is to recommend news articles which are similar to the already read article by using attributes like article headline, category, author and publishing date. The number of news releases has grown rapidly and for one individual, it is cumbersome to browse through all online news resources for relevant news articles. Search engines help users up-to some extent in searching through the vast information collections available online and the recommendation systems have emerged to address different challenges and provide users with the information which matches their needs either by their preferences or by content similarity. Each online news publisher tries to handle its news and use some mechanisms to recommend similar news to their readers. The recommendation of news articles is a hard task because of a highly dynamic environment, which leads several challenges, e.g. frequent changes in the set of news articles, set of users, rapid changes in user’s preferences, etc. Therefore, recommendation algorithms must be able to process continuous incoming news streams in real-time. The complex requirements of news recommendations based on some relevancy among news articles that best fullfill the user’s requirements lead to many rich research scenarios and make it more interesting.

**Literature Survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl No** | **Title** | **Author / Journal name / Year** | **Technique** | **Result** |
| 1 | News  Recommendation with Recurrent Neural Networks | Gabriel de souza p. moreira, Dietmar jannach, adilson marques da cunha,2019 | RNN, LSTM | Proposed hybrid approach leads to higher prediction accuracy. |
| 2 | News  Recommendation  Systems | Chong feng,  Muzammil Khan,  Journal of recommendation system,2020 | Collaborative  Filtering  Approach | News recommender systems (NRS) are developed to relieve the information overload problem and suggest news items that might be of interest for the news readers. |
| 3 | News  Recommendation system | Jing Qin  Journal of recommendation  model, 2020 | CF method, userbased collaborative filtering (UserCF) method, ItemCF,  CNN | CF methods are used to calculate the  similarity of users or news or MFbased modelling CF method have better accuracy than CB |
| 4 | Recommender system for news articles using supervised learning | Akshay kumar chaturvedi | Linear regression, Naïve bayes classifier, Logistic regression | Linear regression turned out to be the best one. Number of classes and performance of model are inversely proportional to each other. |
| 5 | Article | Nayana k | Content based | Collaborative |
|  | recommendation using python |  | and collaborative based filtering | filtering automatically  builds a recommendation system that evolves and decides which feature needs to be used. Content based filtering at  the start will help avoid problems  like cold start where users have no history. |
| 6 | News recommender system : a review of recent progress, challenges and opportunities | Shaina raza and chen ding | Deep neural networks | The general algorithms are  insufficient to provide news recommendations since they need to be modified, varied and extended to large extent. Deep learning-based algorithms have addressed those limitations.    Beyond accuracy other aspects like diversity, coverage, novelty, serendipity is important to provide better user experience in  news recommender system. |

**Datasets And Tools Used**

The dataset contains around 200k news headlines from the year 2012 to 2018. The model trained on this dataset could be used to identify tags for untracked news articles or to identify the type of language used in different news articles.

Various categories in our dataset include :

Politics, Entertainment, Wellness, Travel, Style and beauty, Parenting, Healthy living and Queer voices, Food and drinks, Business, Comedy, Sports, Crime, Media, Science, Education etc. Within each category our dataset contains the Headlines of the news and the authors of that particular news. The link of that specified news is also provided with respect to that particular author.

There also includes a column for short\_description of each news and the particular date it was published. These specified categories and columns in our dataset help in segregating each news and they further enhance the results of our dataset. Our dataset is obtained from Huffpost.

The tool we are going to work on this project is ‘Python’, where use use libraries such as numpy, pandas, matplotlib, sklearn and nltk.

**Algorithms / Techniques description**

Collaborative filtering tackles the similarities between the users and items to perform recommendations. Meaning that the algorithm constantly finds the relationships between the users and in-turns does the recommendations. The algorithm learns the embeddings between the users without having to tune the features. The most common technique is by performing Matrix Factorization to find the embeddings or features that make up the interest of a particular user.

**Algorithm** used here is Content based filtering. The aim of this content-based recommender system is to estimate the utility of a set of objects belonging to a given domain, starting from the information available about users and objects. The basis for content-based recommendations is the similarity between users/items determined by their qualities. It makes use of extra data (meta data) about persons or objects, i.e., it depends on the kind of content that is already available. This meta data may include demographic information about the user, such as age, gender, occupation, location, and skill sets. Similar information may be found for products, including item name, details, category, registration date, etc. Therefore, the fundamental notion is to suggest items by identifying persons or items that are similar to the subject user or item based on those individual’s traits.

**Techniques description**: (following are the procedures for this technique)

1. Importing necessary libraries
2. [Loading Data](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "2.-Loading-Data)
3. [Data Preprocessing](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "3.-Data-Preprocessing)

* [Fetching only the articles from 2018](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "3.a-Fetching-only-the-articles-from-2018)
* [Removing all the short headline articles](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "3.b-Removing-all-the-short-headline-articles)
* [Checking and removing all the duplicates](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "3.c-Checking-and-removing-all-the-duplicates)
* [Checking for missing values](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "3.d-Checking-for-missing-values)

4) [Basic Data Exploration](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "4.-Basic-Data-Exploration)

* [Basic statistics - Number of articles, authors, categories](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "4.a-Basic-statistics---Number-of-articles,authors,categories)
* [Distribution of articles category-wise](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "4.b-Distribution-of-articles-category-wise)
* [Number of articles per month](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "4.c-Number-of-articles-per-month)
* [PDF for length of headlines](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "4.d-PDF-for-length-of-headlines)

5) [Text Preprocessing](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "5.-Text-Preprocessing)

* [Stopwords removal](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "5.a-Stopwords-removal)
* [Lemmatization](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "5.b-Lemmatization)

6) [Headline based similarity on new articles](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "6.-Headline-based-similarity-on-new-articles)

* [Using Bag of Words method](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "6.a-Using-Bag-of-Words method)
* [Using TF-IDF method](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "6.b-Using-TF-IDF-method)
* Using Word2Vec embedding
* Weighted similarity based on headline and categories
* Weighted similarity based on headline, category and author

[Weighted similarity based on headline, category, author and publishing day](https://www.kaggleusercontent.com/kf/45395921/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..IFdmmWV2Ksvj4JLhYWEdjg.MmN3nABSgrGHjghXANLNYbgim4elwjNd2LjtfGyz2wCRZ6CkvX_j07LLHvD1QjiKbLFl3fG2dgu9VnWikcI4uu43IjelRGQAOzCqKTNrtgLTM5vC0gusojgsXqfRs491eL1rXeesXwZL8F2O-gYavud0nLDgZR53Q6ggJQz79Oj_zRgKZTE-Aazkgd5DX-mZHik2IOZcuT8_gkdalvNm1OytVjOpGdBaRF855SVy5Na6-LFwR2sTXWchVwBSlZEzqHnpPM_TLxWIw7JQg9dKcOKTLzIWbIhR9fsSTZeFqZLt8C9Ata-KIxnkFNU4wNRFn992EEL1--YL4IytRbQnSOsPqAJHdotuvoUQqCmuuICRDgxFWtyYqr6KV4jUh8b74lbz0HV397M_mBmwoZf7hRJpHskJ3ZCVwscYq43FCLw2n4i6fmtfq-2aa_xovloZnVur7dfwImTRQrRfKUbUrGDLZMxA8G3B7TYLfdr2svRtpLHH__fuhWKEaS5zG7NzJ0mrtS6BCSyH5Ogbc56PZ1Xwikw9g1erMpMqnpmDavu-huAX1tH66SW2Hq95iFTC1I3GDnE1A_1ZD2KFtcH1Rq7tfjsmynxRePRodGCHTcjVCxnQ259LPleRn-5W0etCf3s5QeJEj1t2nFlsEpbWXML6B-BqYTdA2WFt9LNHULr-qf1obDlQG2wT0gwSeDS9GxF74b4zR5K5YSBbskJ-9w.vy5P4hdS-vVr8nhS4-tpsg/__resultx__.html?sharingControls=true" \l "6.f-Weighted-similarity-based-on-headline,-category,-author-and-publishing-day)

**Discussion on results**

Generally, we assess similarity based on distance. If the distance is minimum then high similarity and if it is maximum then low similarity.

To calculate the distance, we need to represent the headline as a d-dimensional vector. Then we can find out the similarity based on the distance between vectors.

A Bag of Words(BoW) method represents the occurrence of words within a document. Here, each headline can be considered as a document and set of all headlines form a corpus.

Using BoW approach, each document is represented by a d-dimensional vector, where d is total number of unique words in the corpus. The set of such unique words forms the Vocabulary.

TF-IDF method is a weighted measure which gives more importance to less frequent words in a corpus. It assigns a weight to each term(word) in a document based on Term frequency(TF) and inverse document frequency(IDF).

TF(i,j) = (# times word i appears in document j) / (# words in document j)

IDF(i,D) = log\_e(#documents in the corpus D) / (#documents containing word i)

weight(i,j) = TF(i,j) x IDF(i,D)

So, if a word occurs a greater number of times in a document but a smaller number of times in all other documents then its TF-IDF value will be high.

Word2vec is a technique for natural language processing published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence.

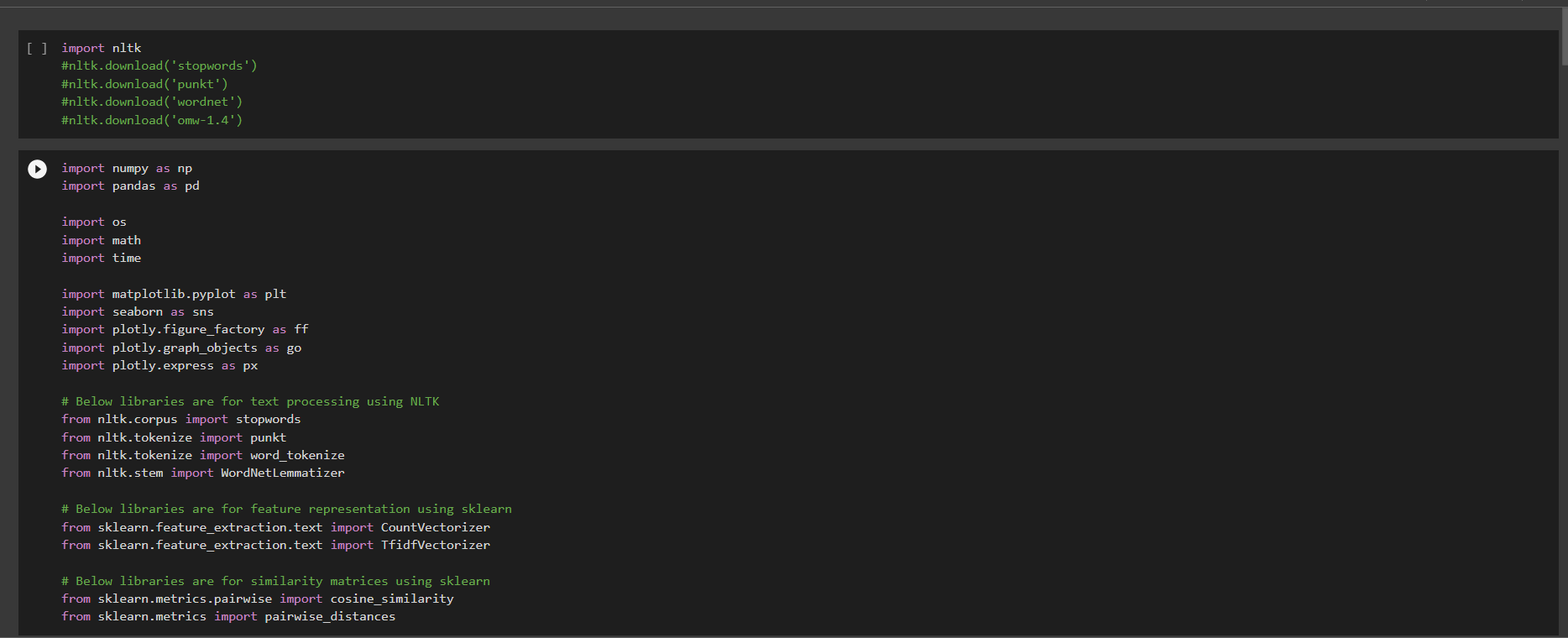
**Results shows that Word2Vec performs better than BoW, TF-IDF and weighted similarity methods.**

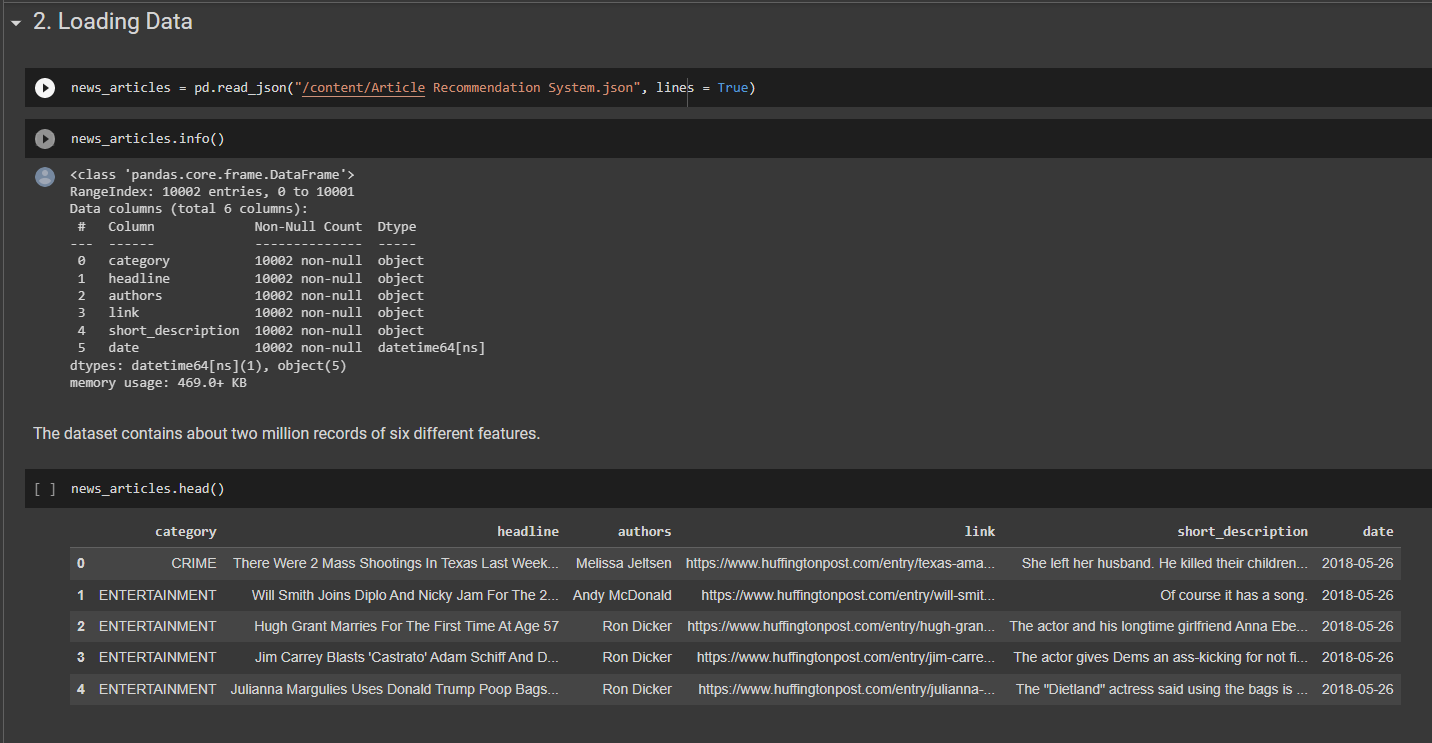
**Conclusion**

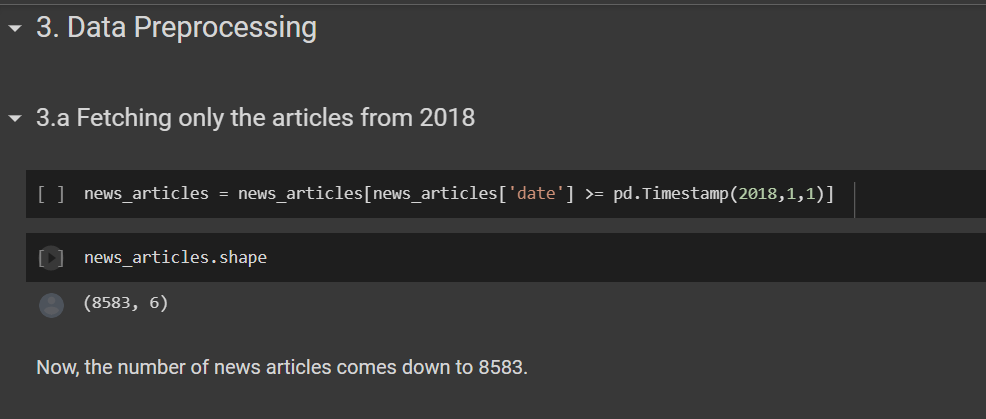
Article Recommendation was done with the help of content-based filtering. Compared to BoW method, here TF-IDF method recommends the articles with headline containing words like "employer", "fire", "flip" in top 5 recommendations and these words occur less frequently in the corpus. Also we used Word2Vec and weighted similarity methods for recommendation.

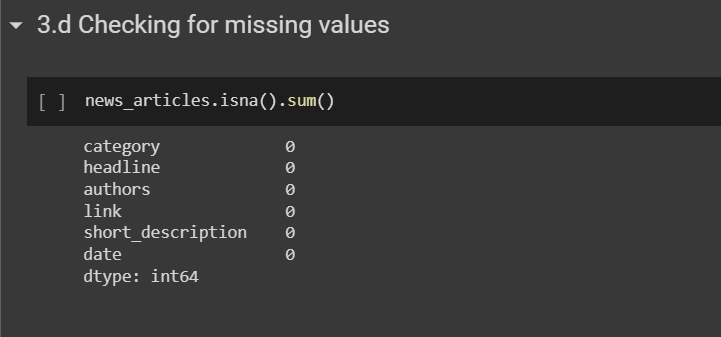
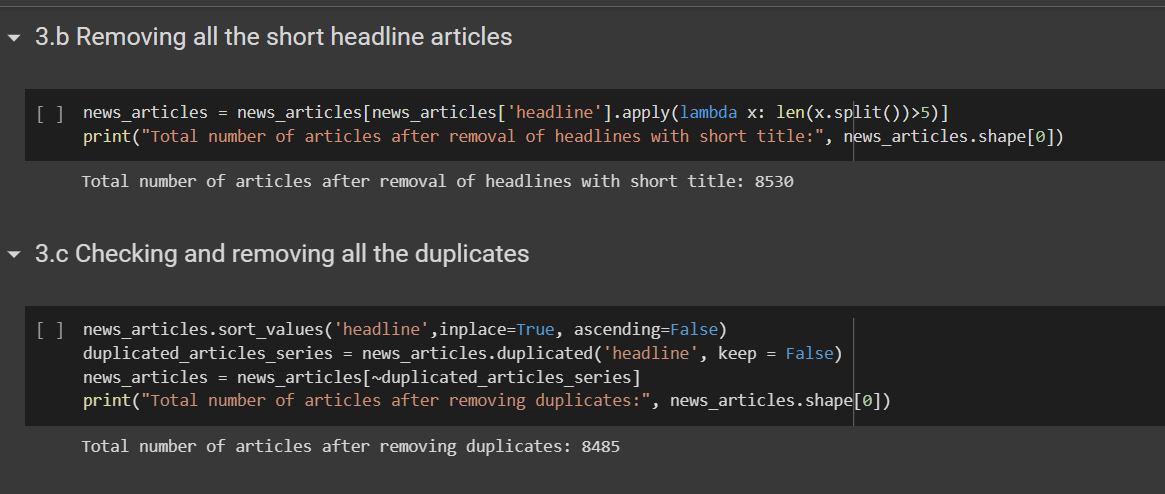
**Implementation**

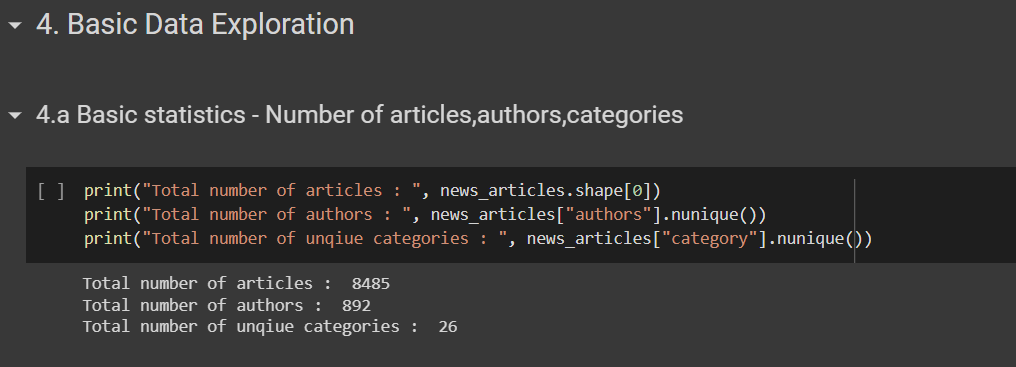
Importing necessary libraries

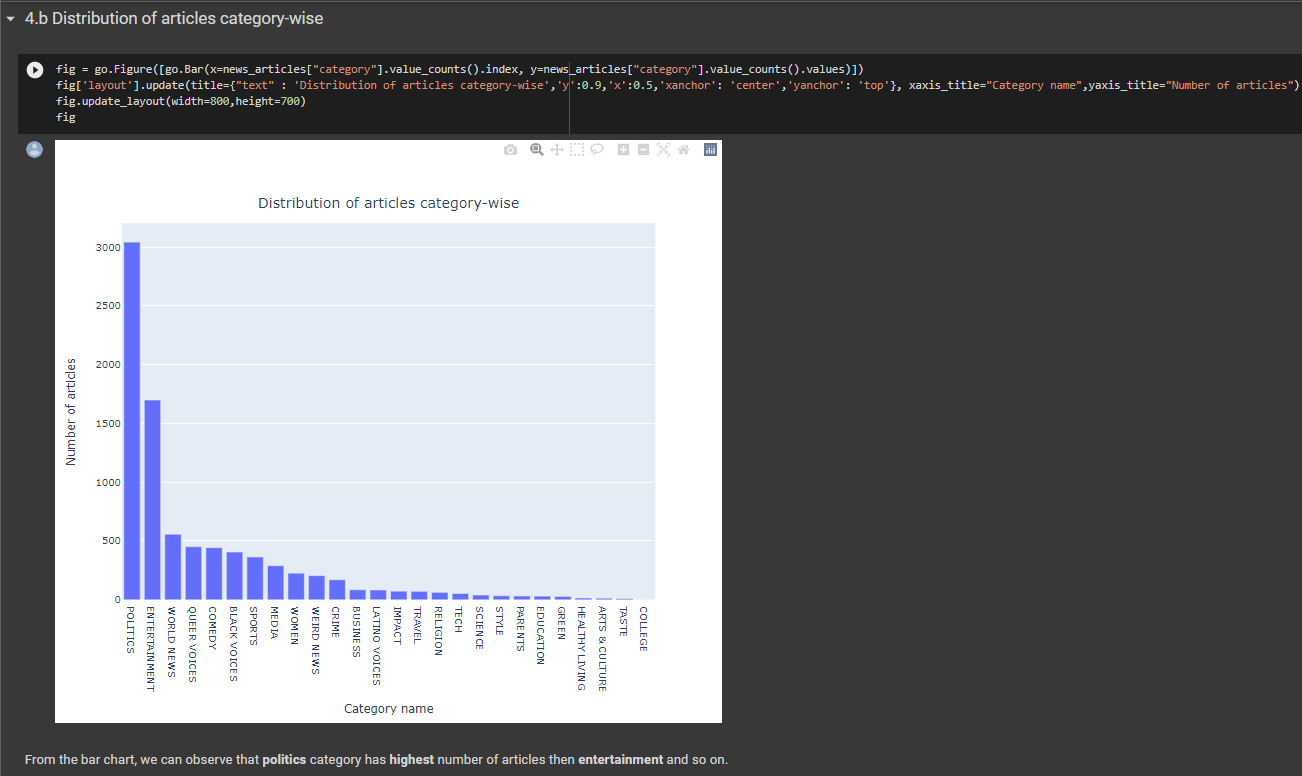


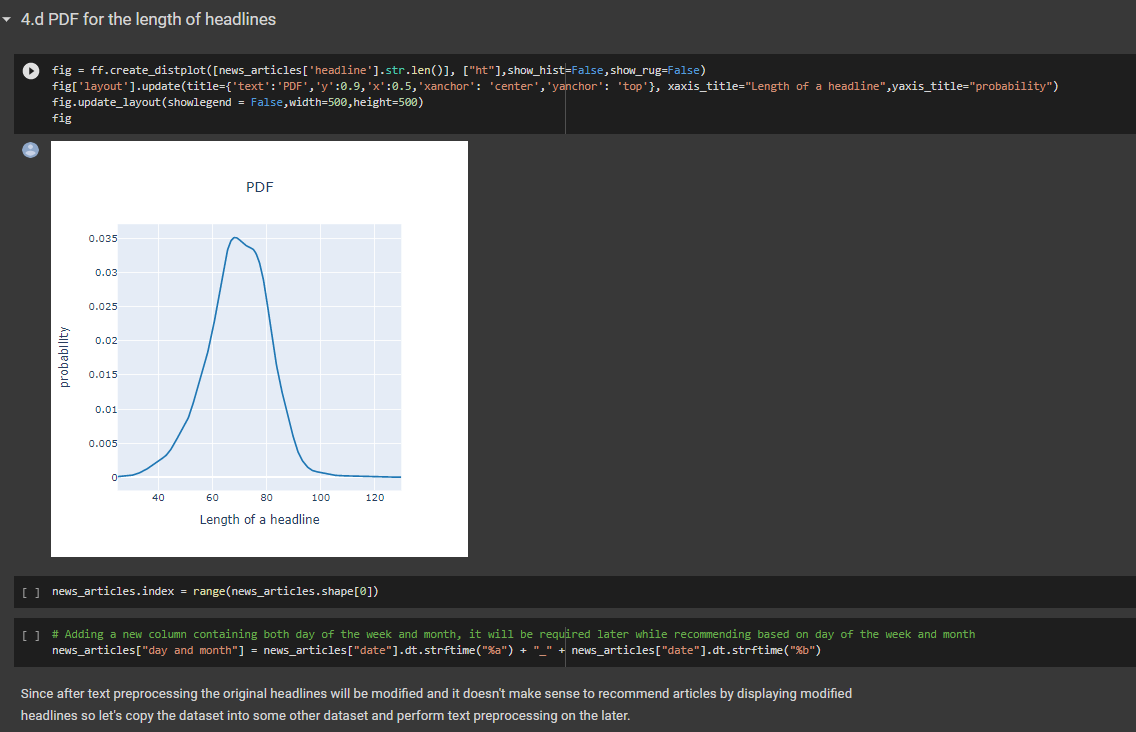


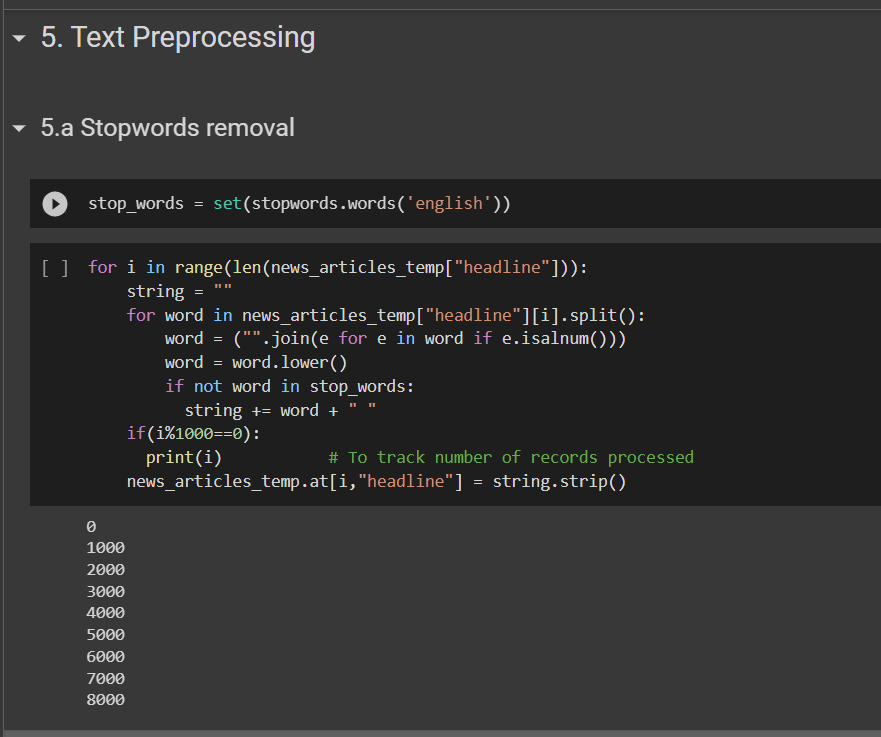


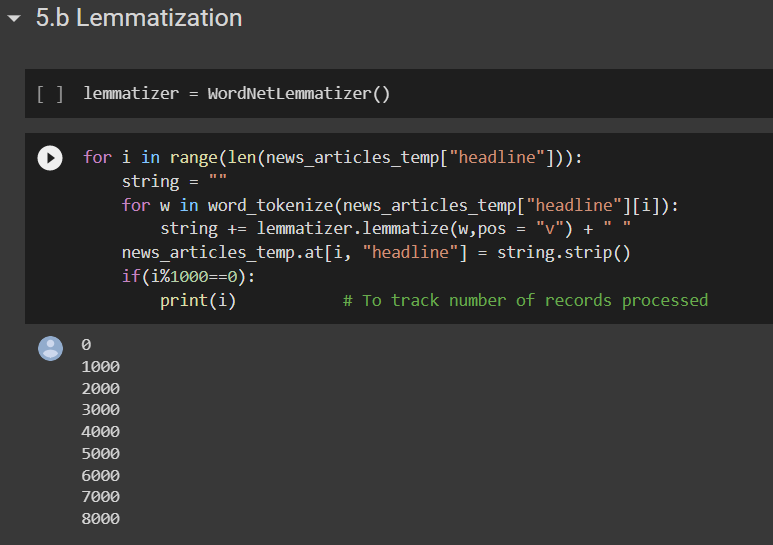


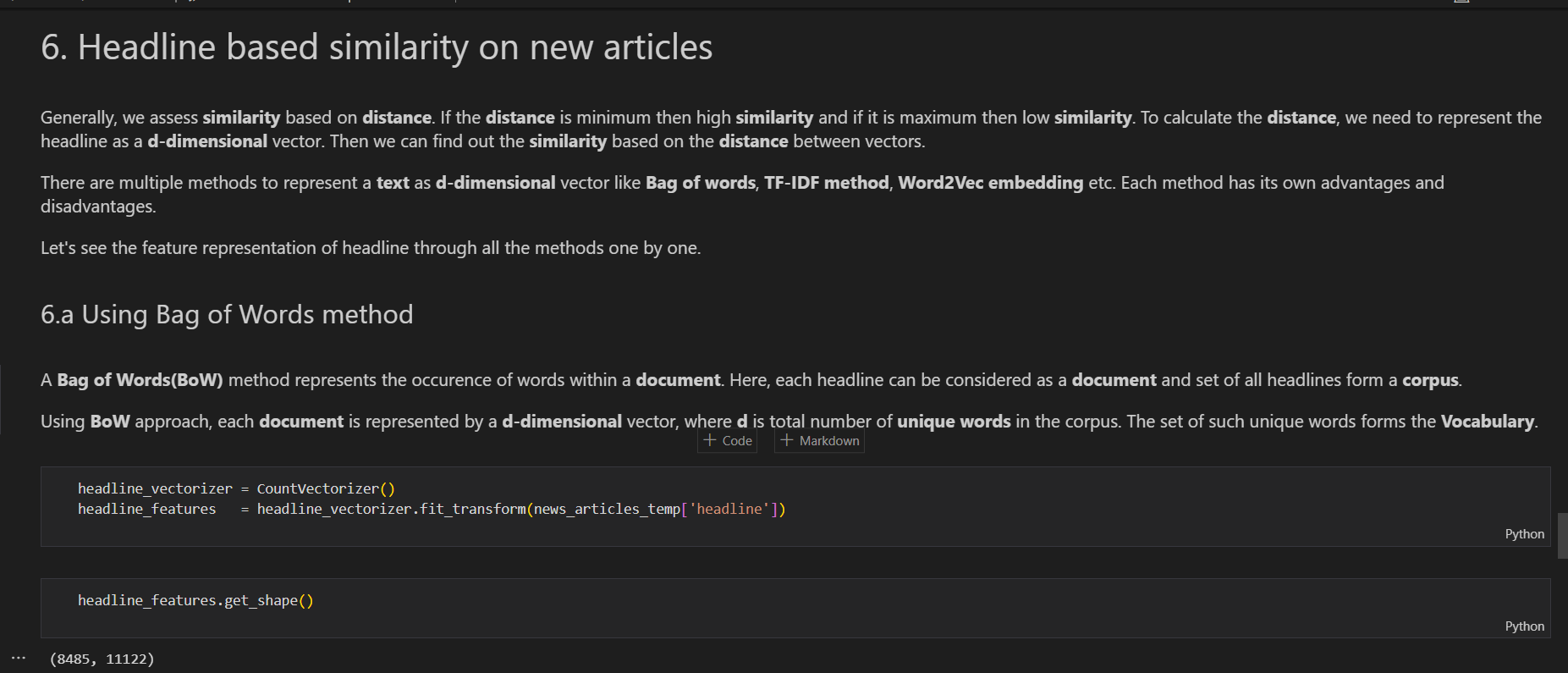


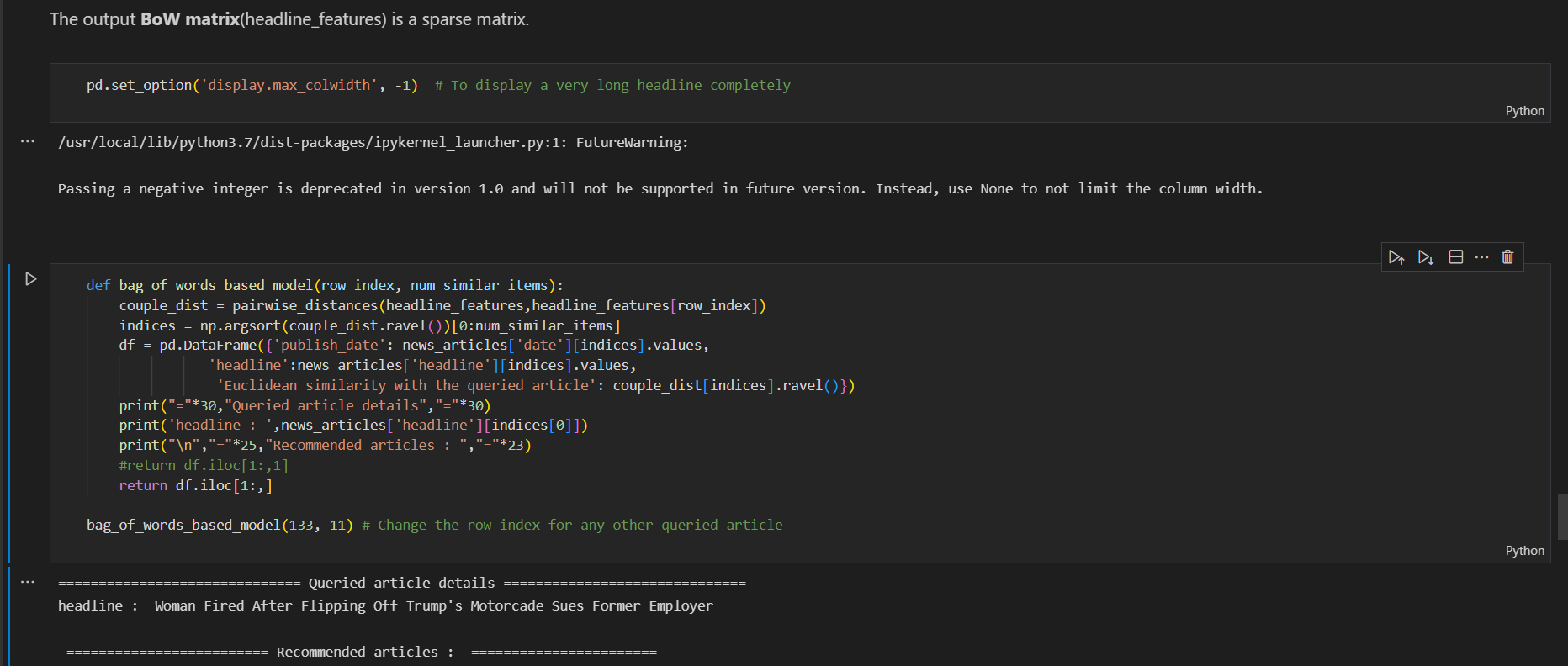


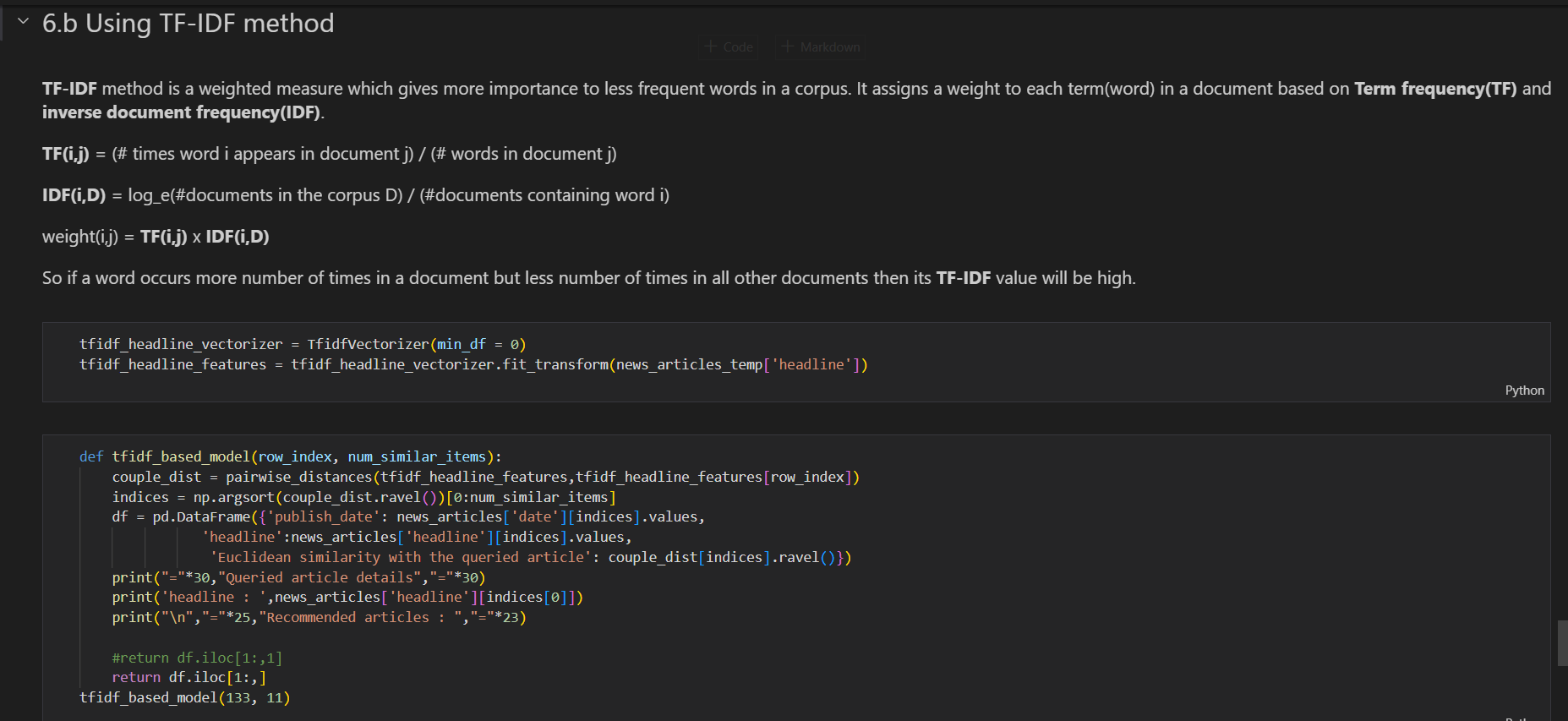


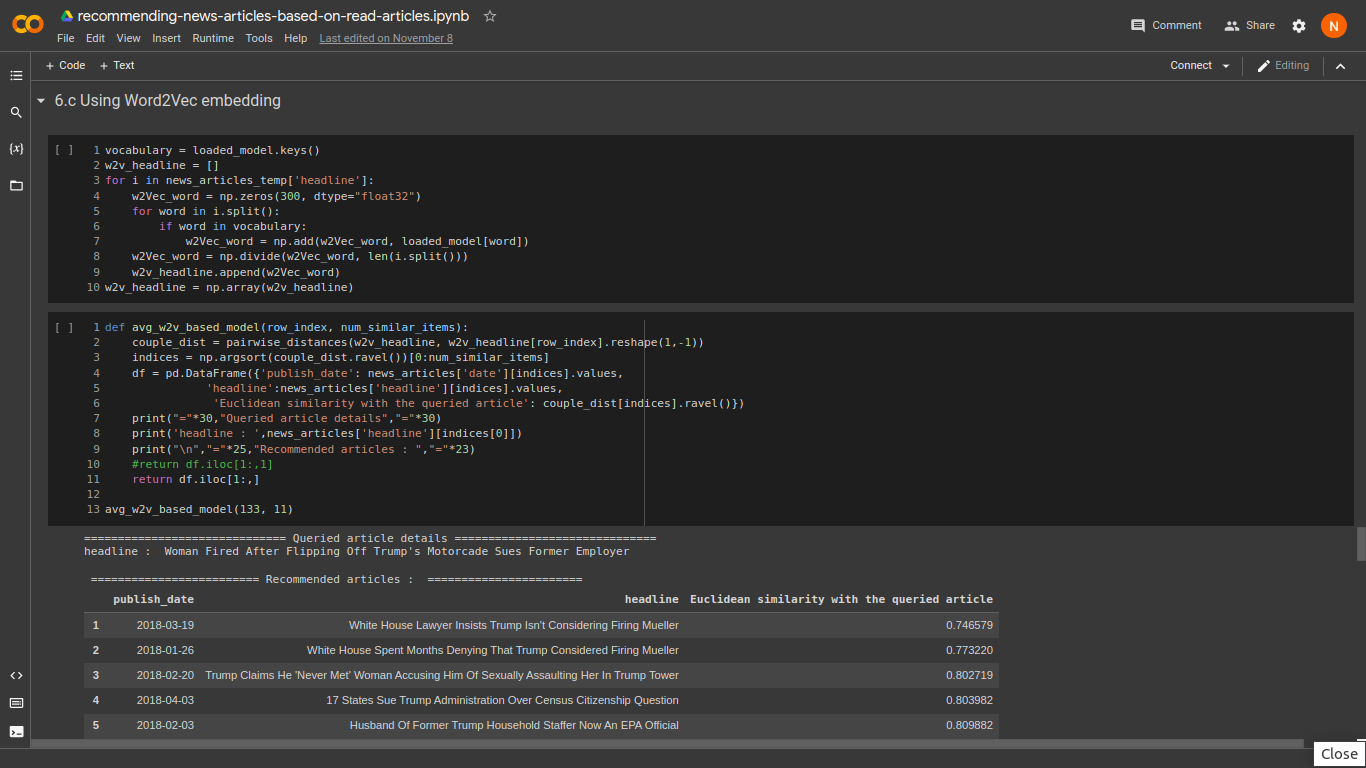


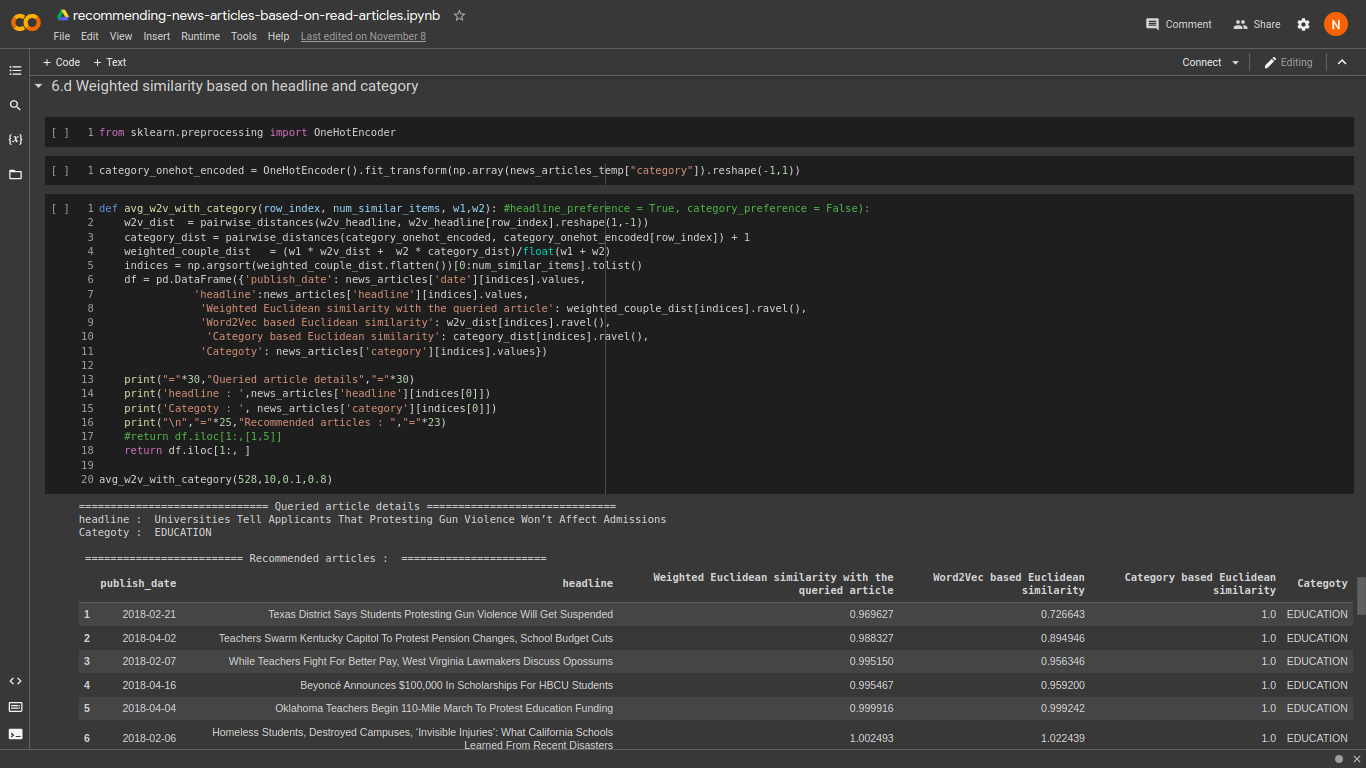












**Github Repository Link**

<https://github.com/Niranjan1112/Recommender-Systems>

**References**

[1]https://medium.com/web-mining-is688-spring-2021/article-recommendation-system-using-python-8b0fec6e6de8

[2]https://link.springer.com/article/10.1007/s10462-021-10043-x

[3]https://www.google.com/url?sa=t&source=web&rct=j&url=https://arxiv.org/pdf/1707.00506&ved=2ahUKEwiKyPfmpLz5AhXoRmwGHbCzASoQFnoECDsQAQ&usg=AOvVaw3F\_oBvH9GLpt5tIEG\_-2Vh

[4]https://arxiv.org/pdf/1904.10367v2.pdf

[5]https://www.researchgate.net/publication/346669020\_Research\_Progress\_of\_News\_Recommendation\_Methodsods

[6]https://www.researchgate.net/publication/338724860\_News\_Recommendation\_Systems\_-\_Accomplishments\_Challenges\_Future\_Directio