San José State
UNIVERSITY

CMPE 256: Large Scale Analytics

Summer2019

Team Project Report

Project Name:

Web Based Research Paper Recommendation System using Hybrid and Multimodel

Advisor:

Professor Shih Yu Chang

Submission By:

Project Group #2

Team Members:

Name	Email	SJSU ID
Jay Kumar Patel	jaykumar.patel@sjsu.edu	013756210
Nithya Kuchadi	nithya.kuchadi@sjsu.edu	013769665
Premal Dattatray Samale	premaldattatray.samale@sjsu.edu	012566333
Pranjalkumar Patel	pranjalkumar.patel@sjsu.edu	013748709
Rama Tejaswini Thotapalli	ramatejaswini.thotapalli@sjsu.edu	013785681

Table of Contents

1. Introduction:

The overabundance of information which is available over the internet makes information seeking a very difficult task. Researchers find it difficult to access and keep track of the most relevant and promising research papers of their interest. The known approach to get research papers is to follow the list of references from the documents they already possessed Even though this approach might be quite effective in some instances, it does not guarantee full coverage of recommending research papers and cannot trace papers published after the possessed paper.

In this project, a pipelined hybridization approach for research paper recommender system is developed where users will be provided with top most recommendation with high accuracy. Here we have implemented content-based approaches where algorithms will compare the content and in addition to that mining the hidden associations between a target paper and its references using collaborative approaches.

Our aim is to identify the latent associations that exist between research papers based on the perspective of paper-citation relations. Firstly, we will find the similarity between target paper and list of papers, where we get top similar papers and then these papers are taken into account to find the list of candidate papers using references and citations. A candidate paper is qualified for consideration in if it cited any of the target paper’s references, recommend the top-N most similar papers based on the assumption that if there exist significant co-occurrence between the target paper and the qualified candidate papers, then there exist some extent of similarities between them. This strictness in qualifying a candidate paper helps in enhancing the overall performance of the approach and the ability to return relevant and useful recommendations at the top of the recommendation list.

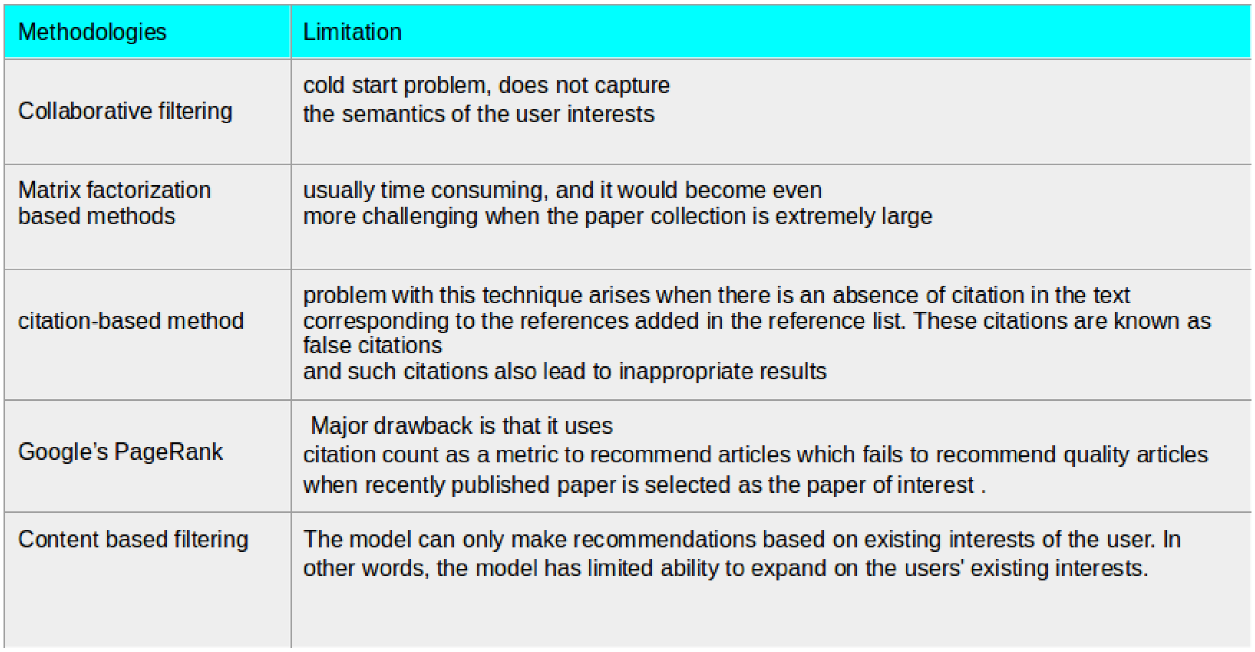
2. Problem Statement:

Researchers spend too much time and struggle to find the suitable article they are looking for. The problem becomes worse when a researcher with insufficient knowledge of searching research articles. How to formalize and solve the recommendation problem? In the traditional recommendation approaches, the results of the query miss many high-quality papers, which are either published recently or have low citation count. In this paper our aim is to develop a web-based application for research paper recommendation system with improved quality.

3. Literature Survey:

As a part of literature survey, we have gone through around 15 research papers and understood the various methodologies of research paper recommendation systems. We analyzed that different methods have their own advantages and limitations. Recent technologies are using hybrid based recommendation, content based filtering, collaborative filtering, citation based, using google PageRank.

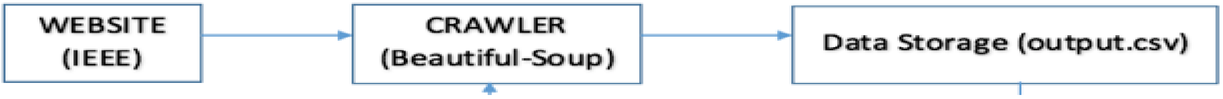
We have studied in detail limitation of each method and proposed our own web-based research paper recommendation system using hybrid based and multi model based approach. Multi model-based approach includes total 6 models of recommendation and its comparative study. Below table describes the summary of our literature survey.



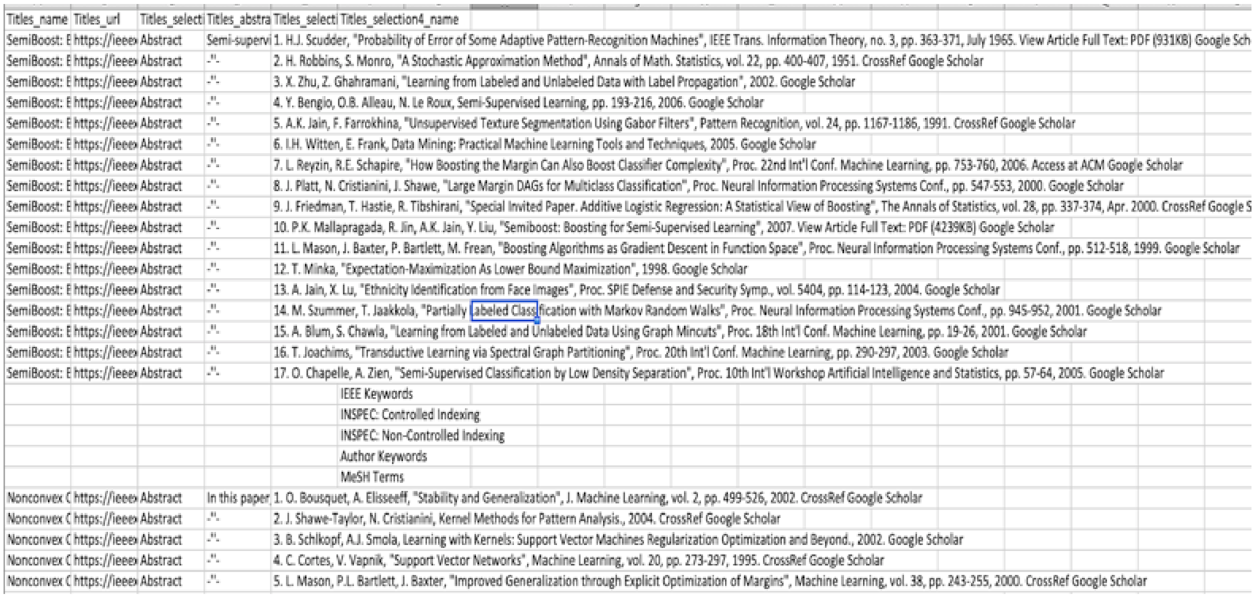
Methodologies	Limitation
Collaborative filtering	cold start problem, does not capture the semantics of the user interests
Matrix factorization based methods	usually time consuming, and it would become even more challenging when the paper collection is extremely large
citation-based method	problem with this technique arises when there is an absence of citation in the text corresponding to the references added in the reference list. These citations are known as false citations and such citations also lead to inappropriate results
Google's PageRank	Major drawback is that it uses citation count as a metric to recommend articles which fails to recommend quality articles when recently published paper is selected as the paper of interest .
Content based filtering	The model can only make recommendations based on existing interests of the user. In other words, the model has limited ability to expand on the users' existing interests.

4. Dataset Preparation:

We have scraped the research papers a dataset from ww.ieee.org, where we have titles, abstract, authors, URL's of the papers, references, citations and keywords of the paper. This web scraping is done using the beautiful soup and html parser framework. Firstly, we have downloaded all the html pages with the required information, and from those html pages we have inspected the class and id of the tables containing the required information. This information is retrieved using find all API and saved into CSV files.



Scraped data looks like:



Titles_name	Titles_url	Titles_selecti	Titles_abstra	Titles_selecti	Titles_selection4_name
SemiBoost: E	https://ieee	Abstract	Semi-supervi	1. H.J. Scudder, "Probability of Error of Some Adaptive Pattern-Recognition Machines", IEEE Trans. Information Theory, no. 3, pp. 363-371, July 1965. View Article Full Text: PDF (931KB) Google Sch	
SemiBoost: E	https://ieee	Abstract	-"-	2. H. Robbins, S. Monro, "A Stochastic Approximation Method", Annals of Math. Statistics, vol. 22, pp. 400-407, 1951. CrossRef Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	3. X. Zhu, Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation", 2002. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	4. Y. Bengio, O.B. Alleau, N. Le Roux, Semi-Supervised Learning, pp. 193-216, 2006. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	5. A.K. Jain, F. Farrokhina, "Unsupervised Texture Segmentation Using Gabor Filters", Pattern Recognition, vol. 24, pp. 1167-1186, 1991. CrossRef Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	6. I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2005. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	7. L. Reyzin, R.E. Schapire, "How Boosting the Margin Can Also Boost Classifier Complexity", Proc. 22nd Int'l Conf. Machine Learning, pp. 753-760, 2006. Access at ACM Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	8. J. Platt, N. Cristianini, J. Shawe, "Large Margin DAGs for Multiclass Classification", Proc. Neural Information Processing Systems Conf., pp. 547-553, 2000. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	9. J. Friedman, T. Hastie, R. Tibshirani, "Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting", The Annals of Statistics, vol. 28, pp. 337-374, Apr. 2000. CrossRef Google S	
SemiBoost: E	https://ieee	Abstract	-"-	10. P.K. Mallapragada, R. Jin, A.K. Jain, Y. Liu, "Semiboost: Boosting for Semi-Supervised Learning", 2007. View Article Full Text: PDF (4239KB) Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	11. L. Mason, J. Baxter, P. Bartlett, M. Frean, "Boosting Algorithms as Gradient Descent in Function Space", Proc. Neural Information Processing Systems Conf., pp. 512-518, 1999. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	12. T. Minka, "Expectation-Maximization As Lower Bound Maximization", 1998. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	13. A. Jain, X. Lu, "Ethnicity Identification from Face Images", Proc. SPIE Defense and Security Symp., vol. 5404, pp. 114-123, 2004. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	14. M. Szummer, T. Jaakkola, "Partially Labeled Classification with Markov Random Walks", Proc. Neural Information Processing Systems Conf., pp. 945-952, 2001. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	15. A. Blum, S. Chawla, "Learning from Labeled and Unlabeled Data Using Graph Mincuts", Proc. 18th Int'l Conf. Machine Learning, pp. 19-26, 2001. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	16. T. Joachims, "Transductive Learning via Spectral Graph Partitioning", Proc. 20th Int'l Conf. Machine Learning, pp. 290-297, 2003. Google Scholar	
SemiBoost: E	https://ieee	Abstract	-"-	17. O. Chapelle, A. Zien, "Semi-Supervised Classification by Low Density Separation", Proc. 10th Int'l Workshop Artificial Intelligence and Statistics, pp. 57-64, 2005. Google Scholar	
					IEEE Keywords
					INSPEC: Controlled Indexing
					INSPEC: Non-Controlled Indexing
					Author Keywords
					MeSH Terms
Nonconvex C	https://ieee	Abstract	In this paper,	1. O. Bousquet, A. Elisseeff, "Stability and Generalization", J. Machine Learning, vol. 2, pp. 499-526, 2002. CrossRef Google Scholar	
Nonconvex C	https://ieee	Abstract	-"-	2. J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis., 2004. CrossRef Google Scholar	
Nonconvex C	https://ieee	Abstract	-"-	3. B. Schlkopf, A.J. Smola, Learning with Kernels: Support Vector Machines Regularization Optimization and Beyond., 2002. Google Scholar	
Nonconvex C	https://ieee	Abstract	-"-	4. C. Cortes, V. Vapnik, "Support Vector Networks", Machine Learning, vol. 20, pp. 273-297, 1995. CrossRef Google Scholar	
Nonconvex C	https://ieee	Abstract	-"-	5. L. Mason, P.L. Bartlett, J. Baxter, "Improved Generalization through Explicit Optimization of Margins", Machine Learning, vol. 38, pp. 243-255, 2000. CrossRef Google Scholar	

4.1. Data Characteristics:

Characteristics of the data are the features and attributes of how the data looks like. Here authors, n_citations, title, venue, id and abstract of research papers.

{"authors": ["Leon A. Sakkal", "Kyle Z. Rajkowski", "Roger S. Armen"],

"n_citation": 0, "references": ["4f4f200c-0764-4fef-9718-b8bccf303dba", "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"],

"title": "Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosine and muscarinic acetylcholine receptors",

"venue": "Journal of Computational Chemistry","year": 2017,

"id": "013ea675-bb58-42f8-a423-f5534546b2b1"

"abstract": "In this paper, a kind of novel jigsaw EBG structure is designed and applied into conformal antenna array. The stop band of jigsaw EBG structure is simulated. And the 4-element antenna array with Taylor amplitude distribution is considered. The effect of EBG structure on the side-lobe level of antenna array radiation pattern is investigated."

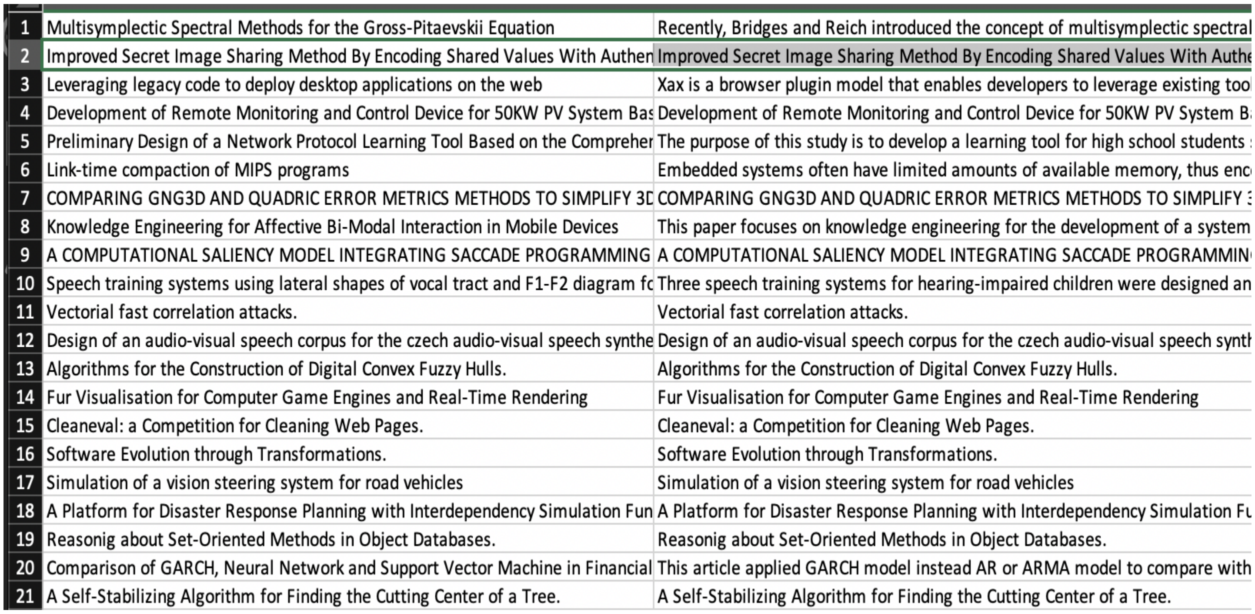
}

4.2. Data Preprocessing:

Data preprocessing is a data mining technique that transforms raw data into an understandable format. Real-world data will be inconsistent, incomplete and/or lacking in certain behaviors or trends, and is likely to contain many errors, as users will provide implicit and explicit feedbacks. Data preprocessing is a method of resolving such issues and get the data, which is used to train the algorithms for prediction and recommendations. In our dataset, we have removed the null rows of the data

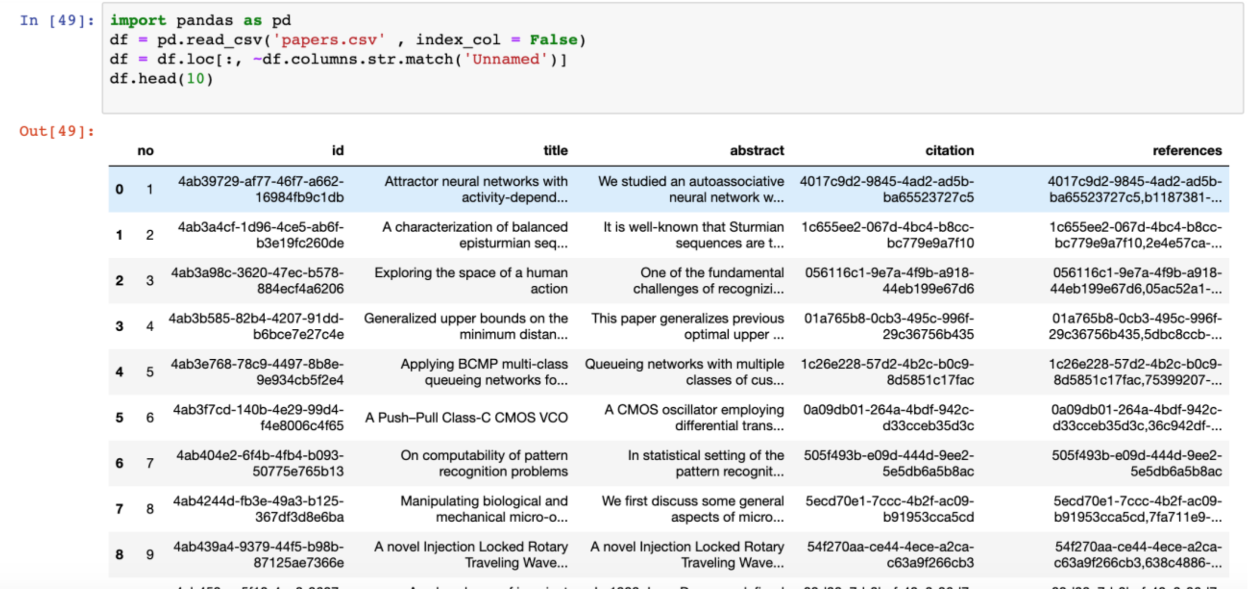
- The fields 'paper authors', 'citation number', 'paper venue', 'published year' are dropped from the csv file.
- Removing the records, which do not have title
- Replacing the records' null abstract with title where there is no abstract
- The extracted text is cleaned such that only English words are contained in the csv file.

In order to maintain 3 characteristics volume, velocity and variety of the data, we have done these data processing steps. After preprocessing the data looks like.



1	Multisymplectic Spectral Methods for the Gross-Pitaevskii Equation	Recently, Bridges and Reich introduced the concept of multisymplectic spectral
2	Improved Secret Image Sharing Method By Encoding Shared Values With Authen	Improved Secret Image Sharing Method By Encoding Shared Values With Authe
3	Leveraging legacy code to deploy desktop applications on the web	Xax is a browser plugin model that enables developers to leverage existing too
4	Development of Remote Monitoring and Control Device for 50KW PV System Bas	Development of Remote Monitoring and Control Device for 50KW PV System B
5	Preliminary Design of a Network Protocol Learning Tool Based on the Compreher	The purpose of this study is to develop a learning tool for high school students
6	Link-time compaction of MIPS programs	Embedded systems often have limited amounts of available memory, thus enc
7	COMPARING GNG3D AND QUADRIC ERROR METRICS METHODS TO SIMPLIFY 3D	COMPARING GNG3D AND QUADRIC ERROR METRICS METHODS TO SIMPLIFY 3
8	Knowledge Engineering for Affective Bi-Modal Interaction in Mobile Devices	This paper focuses on knowledge engineering for the development of a system
9	A COMPUTATIONAL SALIENCY MODEL INTEGRATING SACCADE PROGRAMMING	A COMPUTATIONAL SALIENCY MODEL INTEGRATING SACCADE PROGRAMMIN
10	Speech training systems using lateral shapes of vocal tract and F1-F2 diagram fc	Three speech training systems for hearing-impaired children were designed an
11	Vectorial fast correlation attacks.	Vectorial fast correlation attacks.
12	Design of an audio-visual speech corpus for the czech audio-visual speech synthe	Design of an audio-visual speech corpus for the czech audio-visual speech synth
13	Algorithms for the Construction of Digital Convex Fuzzy Hulls.	Algorithms for the Construction of Digital Convex Fuzzy Hulls.
14	Fur Visualisation for Computer Game Engines and Real-Time Rendering	Fur Visualisation for Computer Game Engines and Real-Time Rendering
15	Cleaneval: a Competition for Cleaning Web Pages.	Cleaneval: a Competition for Cleaning Web Pages.
16	Software Evolution through Transformations.	Software Evolution through Transformations.
17	Simulation of a vision steering system for road vehicles	Simulation of a vision steering system for road vehicles
18	A Platform for Disaster Response Planning with Interdependency Simulation Fun	A Platform for Disaster Response Planning with Interdependency Simulation Fu
19	Reasonig about Set-Oriented Methods in Object Databases.	Reasonig about Set-Oriented Methods in Object Databases.
20	Comparison of GARCH, Neural Network and Support Vector Machine in Financial	This article applied GARCH model instead AR or ARMA model to compare with
21	A Self-Stabilizing Algorithm for Finding the Cutting Center of a Tree.	A Self-Stabilizing Algorithm for Finding the Cutting Center of a Tree.

Final data after preprocessing.



```
In [49]: import pandas as pd
         df = pd.read_csv('papers.csv' , index_col = False)
         df = df.loc[:, ~df.columns.str.match('Unnamed')]
         df.head(10)
```

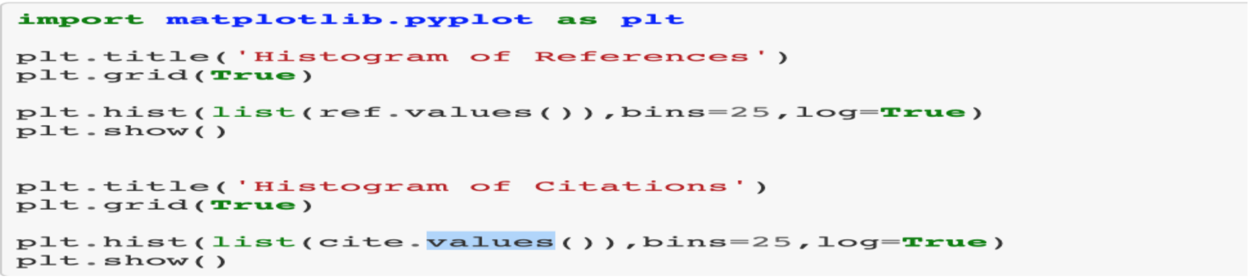
Out[49]:

	no	id	title	abstract	citation	references
0	1	4ab39729-af77-46f7-a662-16984fb9c1db	Attractor neural networks with activity-depend...	We studied an autoassociative neural network w...	4017c9d2-9845-4ad2-ad5b-ba65523727c5	4017c9d2-9845-4ad2-ad5b-ba65523727c5,b1187381-...
1	2	4ab3a4cf-1d96-4ce5-ab6f-b3e19fc260de	A characterization of balanced episturmian seq...	It is well-known that Sturmian sequences are t...	1c655ee2-067d-4bc4-b8cc-bc779e9a7f10	1c655ee2-067d-4bc4-b8cc-bc779e9a7f10,2e4e57ca-...
2	3	4ab3a98c-3620-47ec-b578-884ecf4a6206	Exploring the space of a human action	One of the fundamental challenges of recognizi...	056116c1-9e7a-4f9b-a918-44eb199e67d6	056116c1-9e7a-4f9b-a918-44eb199e67d6,05ac52a1-...
3	4	4ab3b585-82b4-4207-91dd-b6bce7e27c4e	Generalized upper bounds on the minimum distan...	This paper generalizes previous optimal upper ...	01a765b8-0cb3-495c-996f-29c36756b435	01a765b8-0cb3-495c-996f-29c36756b435,5dbc8ccb-...
4	5	4ab3e768-78c9-4497-8b8e-9e934cb5f2e4	Applying BCMP multi-class queueing networks fo...	Queueing networks with multiple classes of cus...	1c26e228-57d2-4b2c-b0c9-8d5851c17fac	1c26e228-57d2-4b2c-b0c9-8d5851c17fac,75399207-...
5	6	4ab3f7cd-140b-4e29-99d4-f4e8006c4f65	A Push–Pull Class-C CMOS VCO	A CMOS oscillator employing differential trans...	0a09db01-264a-4bdf-942c-d33cceb35d3c	0a09db01-264a-4bdf-942c-d33cceb35d3c,36c942df-...
6	7	4ab404e2-6f4b-4fb4-b093-50775e765b13	On computability of pattern recognition problems	In statistical setting of the pattern recognit...	505f493b-e09d-444d-9ee2-5e5db6a5b8ac	505f493b-e09d-444d-9ee2-5e5db6a5b8ac
7	8	4ab4244d-fb3e-49a3-b125-367df3d8e6ba	Manipulating biological and mechanical micro-o...	We first discuss some general aspects of micro...	5ecd70e1-7ccc-4b2f-ac09-b91953cca5cd	5ecd70e1-7ccc-4b2f-ac09-b91953cca5cd,7fa711e9-...
8	9	4ab439a4-9379-44f5-b98b-87125ae7366e	A novel Injection Locked Rotary Traveling Wave...	A novel Injection Locked Rotary Traveling Wave...	54f270aa-ce44-4ece-a2ca-c63a9f266cb3	54f270aa-ce44-4ece-a2ca-c63a9f266cb3,638c4886-...

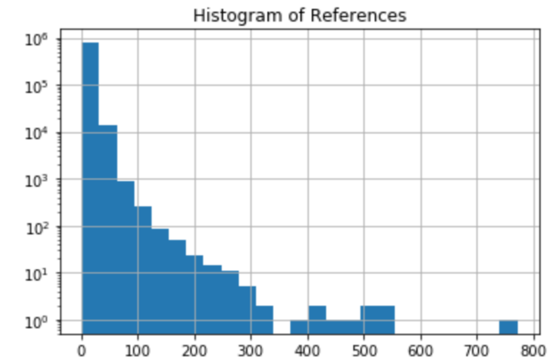
4.3. Exploratory data analysis:

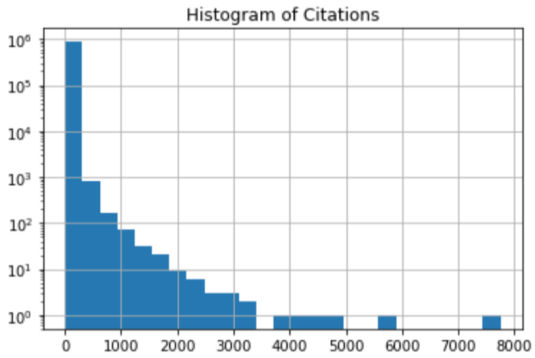
Exploratory data Analysis is a technique which employs a variety of techniques. It consists of various techniques like below.

1.Plotting raw data
2. Plotting simple statistics such as mean, standard deviation, etc.
3. Positioning such plots, to maximize our natural pattern recognition.

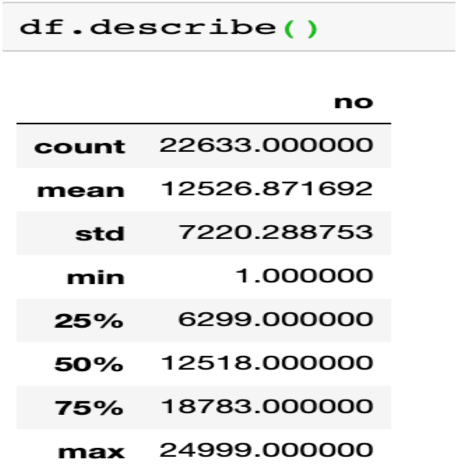


```
import matplotlib.pyplot as plt

plt.title('Histogram of References')
plt.grid(True)

plt.hist(list(ref.values()),bins=25,log=True)
plt.show()


plt.title('Histogram of Citations')
plt.grid(True)

plt.hist(list(cite.values()),bins=25,log=True)
plt.show()
```





Description of the data:



```
df.describe()
```

```
                 no
count  22633.000000
mean   12526.871692
std     7220.288753
min        1.000000
25%     6299.000000
50%    12518.000000
75%    18783.000000
max    24999.000000
```

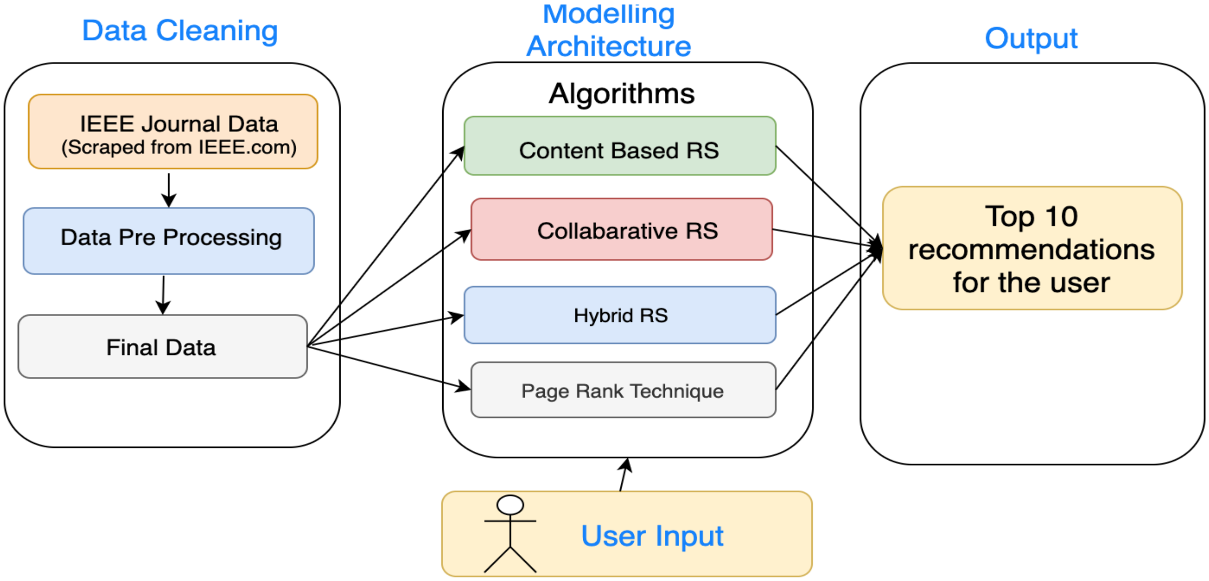
5. Analysis and Methodology of project:

This project is implemented using the pipelined hybrid approach where the output of content-based implementation is provided to collaborative filtering implementation as input.

To overcome all the limitations of the above-mentioned approaches and as we have a large set of data, we are using the below mentioned models with features like title and abstract, References, Citations and Author.

- Content Based
  - TF-IDF & cosine similarity
  - K Nearest Neighbor using Bag of words model and Euclidean Distance.
  - Word2Vec model, Cosine Similarity
- Collaborative Filtering
- Pipelined Hybrid System (Sending the result of the Content based Filtering to Collaborative Filtering)
- Page Rank Algorithm using Neo4j

Block diagram:



6. Implementation:

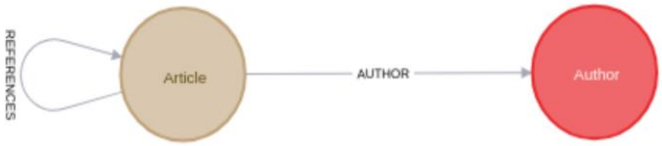
6.1 Page Rank Technique:

PageRank is a way of measuring the importance of website pages. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. In this implementation we have used below tools and frameworks.

- Used below tools and library for the implementation of recommendation system
  - Neo4j
  - Neo4j graph algorithms
  - Neo4j APOC
  - Graphaware's NLP plugin

Ne4j is the graph database where it can store the data in graph formats with nodes and lins Inout project the nodes are the articles and links are links to articles.

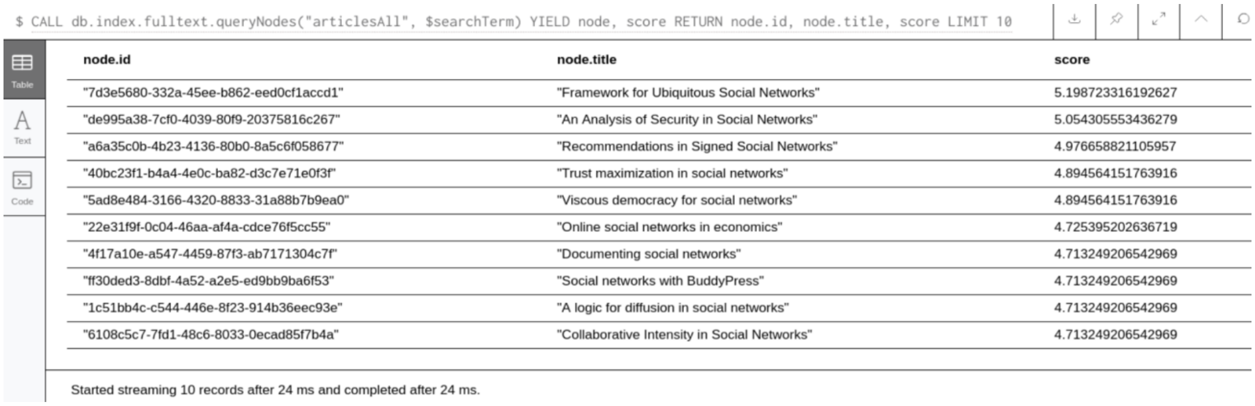
- Taken the below mentioned graph for the connection between the articles and references



Algorithm:
Step 1: Store the data into Neo4j graph database.
Step 2: Search the database with the titles of the search word
Step 3: If the records found then the records will be sent to the pagerank algorithm of networkX which will evaluate the page rank of the articles.
Step 4: The articles which are more relevant and has the highest Pagerank will be provided to the User.

Results:

This is the similar articles with provided input as social networks.



```
$ CALL db.index.fulltext.queryNodes("articlesAll", $searchTerm) YIELD node, score RETURN node.id, node.title, score LIMIT 10
```

node.id	node.title	score
"7d3e5680-332a-45ee-b862-eed0cf1accd1"	"Framework for Ubiquitous Social Networks"	5.198723316192627
"de995a38-7cf0-4039-80f9-20375816c267"	"An Analysis of Security in Social Networks"	5.054305553436279
"a6a35c0b-4b23-4136-80b0-8a5c6f058677"	"Recommendations in Signed Social Networks"	4.976658821105957
"40bc23f1-b4a4-4e0c-ba82-d3c7e71e0f3f"	"Trust maximization in social networks"	4.894564151763916
"5ad8e484-3166-4320-8833-31a88b7b9ea0"	"Viscous democracy for social networks"	4.894564151763916
"22e31f9f-0c04-46aa-af4a-cdce76f5cc55"	"Online social networks in economics"	4.725395202636719
"4f17a10e-a547-4459-87f3-ab7171304c7f"	"Documenting social networks"	4.713249206542969
"ff30ded3-8dbf-4a52-a2e5-ed9bb9ba6f53"	"Social networks with BuddyPress"	4.713249206542969
"1c51bb4c-c544-446e-8f23-914b36eec93e"	"A logic for diffusion in social networks"	4.713249206542969
"6108c5c7-7fd1-48c6-8033-0ecad85f7b4a"	"Collaborative Intensity in Social Networks"	4.713249206542969

Started streaming 10 records after 24 ms and completed after 24 ms.

After getting the similar articles, the articles will be sent to pagerank technique and sorted with the descending order of page ranks. The top 5 articles will be provided to the user as recommendations. The result is provided as article name and page rank of that article.

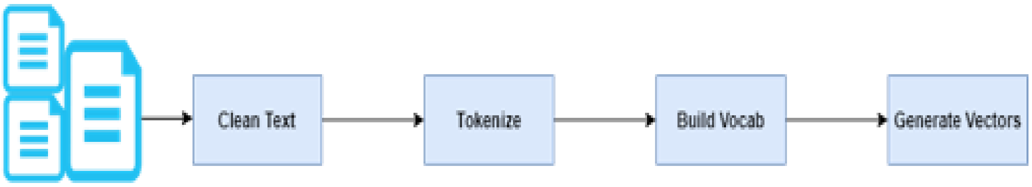
6.2 Content-Based Filtering

The content-based technique is adopted because of its suitability in domains or situations where items are more than the users. Content-based recommenders provide recommendations by comparing the representation of contents describing an item or a product to the representation of the content describing the interest of the user (User's profile of interest). Unlike Collaborative Filtering, if the items have enough descriptions, we avoid the “new item problem”. Content representations are varied, and they open up the options to use different approaches like text processing techniques, the use of semantic information, inferences, etc.

In this project, for content-based filtering, we have considered the title content of the research papers. We have implemented 3 models for transforming all this data into a Vector Space Model, an algebraic representation of titles.

6.2.1. K Nearest Neighbors:

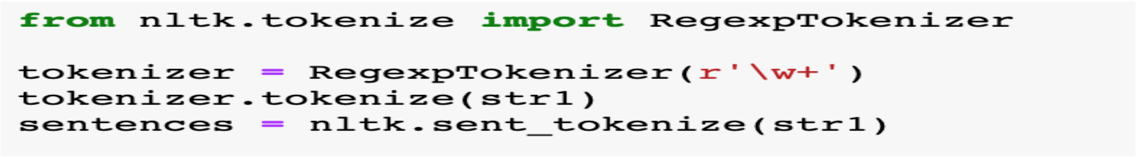
KNN is implemented by using Bag of words model. The bag-of-words model is a way of extracting features from the text for use in machine learning algorithms.
In this approach, tokenized words are used for each title and the frequency of each token is determined. The process of converting NLP text into numbers is called vectorization. Bag-of-words model works on Terms Frequency, i.e. counting the occurrences of tokens and building a sparse matrix of titles x tokens. This model represents documents ignoring the order of the words. In this model, each document looks like a bag containing some words. Therefore, this method allows word modeling based on dictionaries, where each bag contains a few words from the dictionary. The sparse matrix is then feeded to KNN and it recommends n nearest neighbors by calculating the Euclidean distance between test data and each title in the corpus.



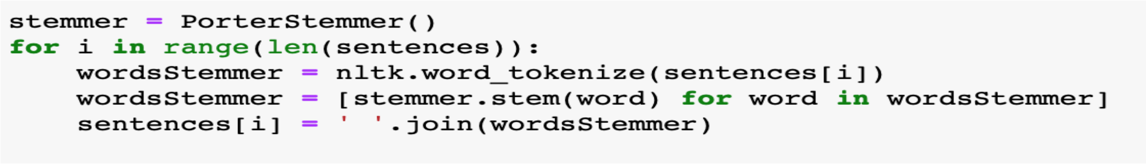
Algorithm is stated as below.

1. The input text data is cleaned and tokenized using nltk library.
2. Feature vectors are generated using Bag of words model
3. KNN model is trained by using these feature vectors
4. Initialized the value of k
5. For getting recommendations, iterate from 1 to number of trained data
6. Calculate distance between test data and each row
7. Sort the distances in ascending order
8. Get nearest k rows and recommend to the user

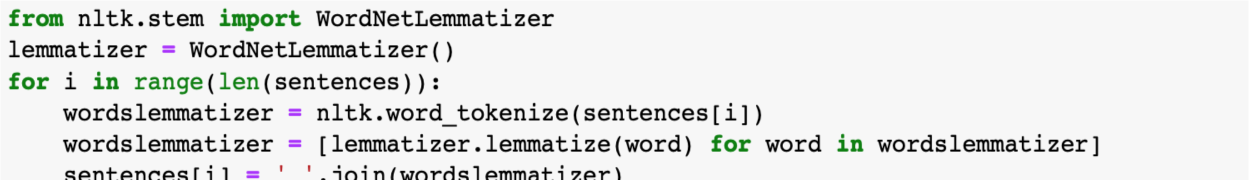
Implementation Screenshots:



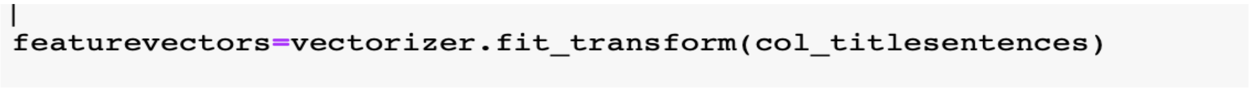
```
from nltk.tokenize import RegexpTokenizer

tokenizer = RegexpTokenizer(r'\w+')
tokenizer.tokenize(str1)
sentences = nltk.sent_tokenize(str1)
```



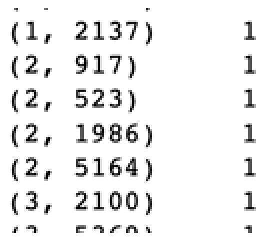
```
stemmer = PorterStemmer()
for i in range(len(sentences)):
    wordsStemmer = nltk.word_tokenize(sentences[i])
    wordsStemmer = [stemmer.stem(word) for word in wordsStemmer]
    sentences[i] = ' '.join(wordsStemmer)
```



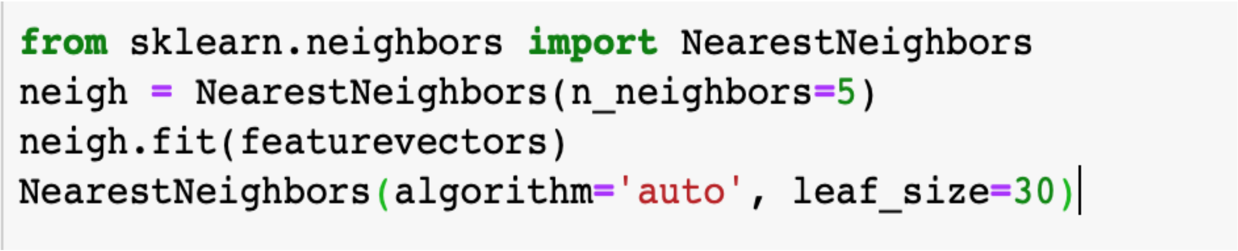
```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
for i in range(len(sentences)):
    wordslemmatizer = nltk.word_tokenize(sentences[i])
    wordslemmatizer = [lemmatizer.lemmatize(word) for word in wordslemmatizer]
    sentences[i] = ' '.join(wordslemmatizer)
```



```
|
featurevectors=vectorizer.fit_transform(col_titlesentences)
```

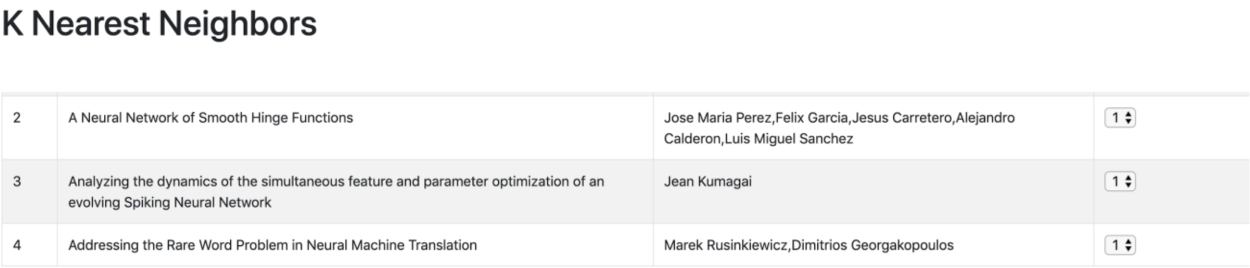


```
(1, 2137)     1
(2, 917)      1
(2, 523)      1
(2, 1986)     1
(2, 5164)     1
(3, 2100)     1
```



```
from sklearn.neighbors import NearestNeighbors
neigh = NearestNeighbors(n_neighbors=5)
neigh.fit(featurevectors)
NearestNeighbors(algorithm='auto', leaf_size=30)
```

Result Screenshot:



K Nearest Neighbors

2	A Neural Network of Smooth Hinge Functions	Jose Maria Perez,Felix Garcia,Jesus Carretero,Alejandro Calderon,Luis Miguel Sanchez	1
3	Analyzing the dynamics of the simultaneous feature and parameter optimization of an evolving Spiking Neural Network	Jean Kumagai	1
4	Addressing the Rare Word Problem in Neural Machine Translation	Marek Rusinkiewicz,Dimitrios Georgakopoulos	1

6.2.2. TF-IDF & Cosine Similarity:

A specific implementation of a Bag of Words is the TF-IDF representation, where TF is for Term Frequency and IDF is Inverse Document Frequency.

Tf-idf is a transformation that is applied on texts to get two real-valued vectors.

●TF: This is simply the frequency of a word in a document. A word that occurs frequently is probably important to that document’s meaning.

●IDF: This is the universe of document frequency among the whole corpus of documents. This tells us the common words that just appear everywhere no matter what the topic is like (a, the,an,is,and..etc..)

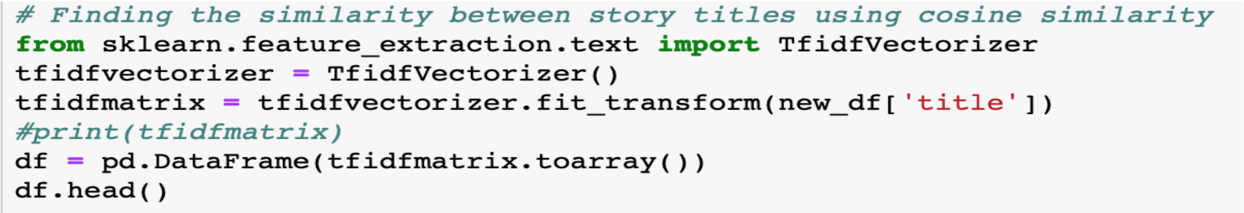
The product of TF and IDF gives us the idea how often the word appears in a document, over how often it just appears everywhere. This gives a measure of how important and unique the word is for the document.

Cosine similarity of any pair of vectors by taking their dot product and dividing that by the product of their norms. That yields the cosine of the angle between the vectors. If d2 and q are tf-idf vectors.

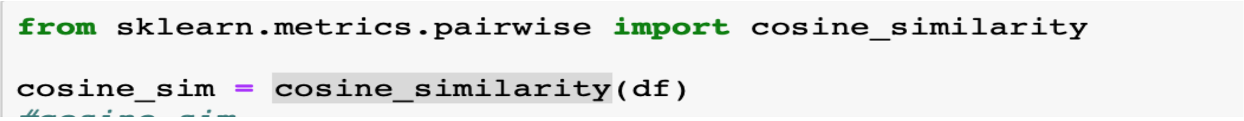
$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

The text pre-processing steps are common for all models in Content-based Filtering. For generating TF-idf vectors, we have used TfidfVectorizer from Sklearn library.

Implementation Screenshots:

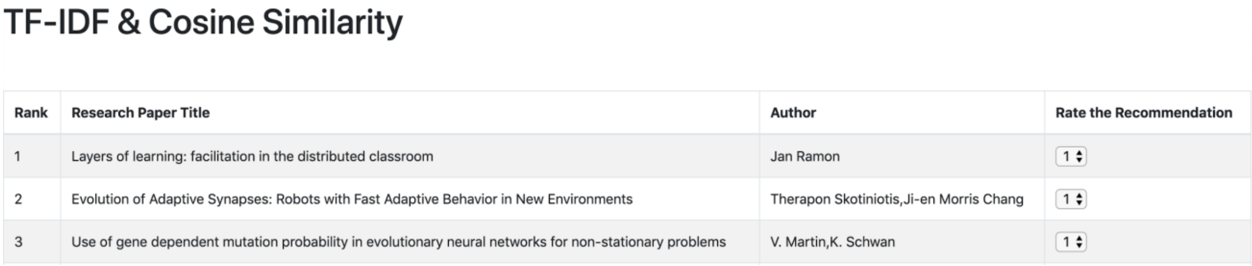


```
# Finding the similarity between story titles using cosine similarity
from sklearn.feature_extraction.text import TfidfVectorizer
tfidfvectorizer = TfidfVectorizer()
tfidfmatrix = tfidfvectorizer.fit_transform(new_df['title'])
#print(tfidfmatrix)
df = pd.DataFrame(tfidfmatrix.toarray())
df.head()
```



```
from sklearn.metrics.pairwise import cosine_similarity

cosine_sim = cosine_similarity(df)
#cosine_sim
```

Result Screenshot:



TF-IDF & Cosine Similarity

Rank	Research Paper Title	Author	Rate the Recommendation
1	Layers of learning: facilitation in the distributed classroom	Jan Ramon	1 ⬍
2	Evolution of Adaptive Synapses: Robots with Fast Adaptive Behavior in New Environments	Therapon Skotiniotis,Ji-en Morris Chang	1 ⬍
3	Use of gene dependent mutation probability in evolutionary neural networks for non-stationary problems	V. Martin,K. Schwan	1 ⬍

The Bag of Words representation does not consider the context of words. For research paper recommendation, it is also important for us to capture the semantic content representation. For this reason, we have implemented Word2Vec model.

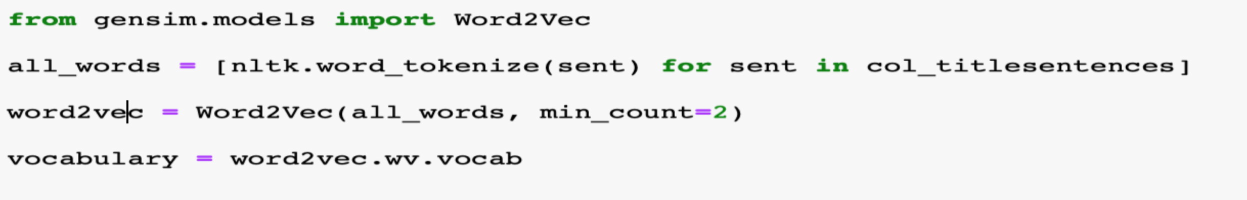
6.2.3. Word2Vec:

Word2vec is a popular machine learning model for generating word embeddings. Word2vec is a shallow, two-layer neural network model that is trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space.

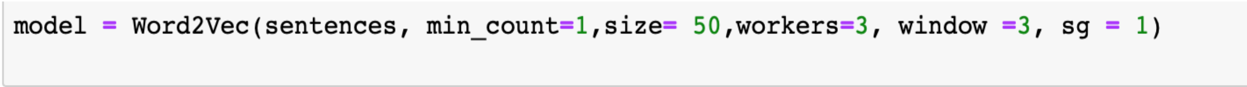
Word vectors are positioned in the vector space in such a way that words that share common contexts in the corpus are located in close proximity to one another in the space. Word2vec model preserves the relationship between words. It deals with the addition of new words in the vocabulary.

The text data is cleaned and tokenized. The tokenized data is stemmed, lemmatized and stop words and punctuation marks are removed in pre-processing steps. Implemented Word2Vec model by using gensim library.

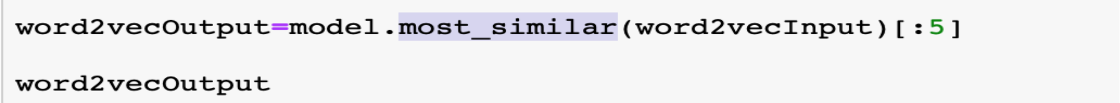
Implementation Screenshots:



```
from gensim.models import Word2Vec
all_words = [nltk.word_tokenize(sent) for sent in col_titlesentences]
word2vec = Word2Vec(all_words, min_count=2)
vocabulary = word2vec.wv.vocab
```

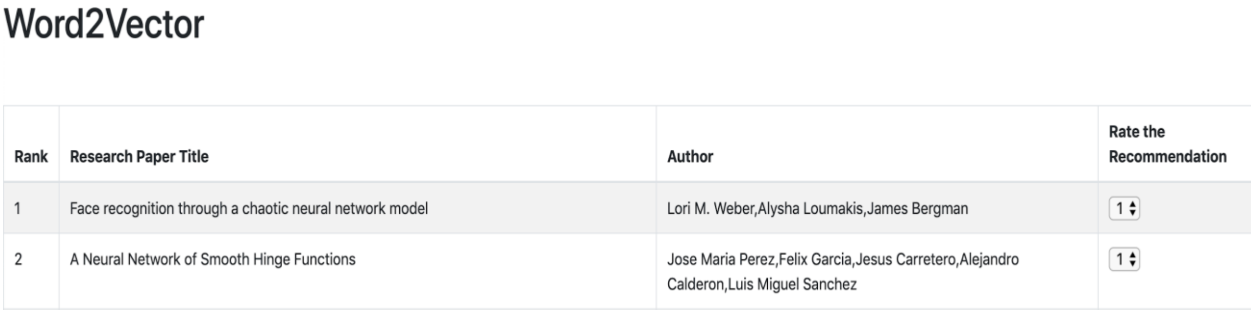


```
model = Word2Vec(sentences, min_count=1,size= 50,workers=3, window =3, sg = 1)
```



```
word2vecOutput=model.most_similar(word2vecInput)[:5]
word2vecOutput
```

Result Screenshot:



Word2Vector

Rank	Research Paper Title	Author	Rate the Recommendation
1	Face recognition through a chaotic neural network model	Lori M. Weber,Alysha Loumakis,James Bergman	1
2	A Neural Network of Smooth Hinge Functions	Jose Maria Perez,Felix Garcia,Jesus Carretero,Alejandro Calderon,Luis Miguel Sanchez	1

7. Collaborative Filtering

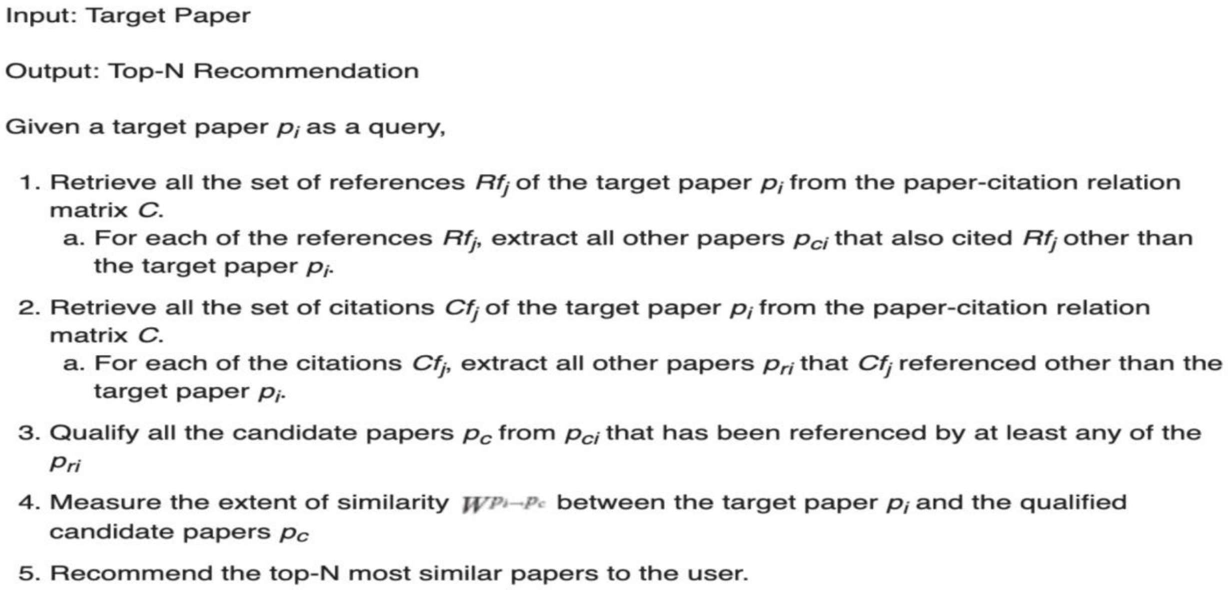
Collaborative filtering recommends items based on how similar users have rated the item. This paradigm assumes that users with similar interests in the past will have similar interests in the future.

Item-Based Collaborative Filtering:

In 1998, Amazon proposed item-based collaborative filtering to address some of the limitations of the user-based approach. Rather than matching the active user to similar users, item-based algorithms match items the active user has rated to similar items. The algorithm then aggregates and recommends similar items, i.e.: users who liked this item also liked. Like user-based collaborative filtering, similarity between two items can be calculated using any similarity measure. Item-based algorithms return top k recommendations, but many approaches also simply return all items with a similarity score above a certain threshold. With user-based collaborative filtering, pre-computing the user neighborhood can lead to poor predictions because user similarity is a dynamic measure and changes constantly. Therefore, all computations must be completed online. Item-based collaborative filtering avoids this problem because item similarity is more static. This allows for pre-computation of the item-item similarity and leads to vast improvements in performance. By leveraging the advantages of collaborative filtering approach, we utilize the publicly available contextual metadata (citations and references for research paper) to infer the hidden associations that exist between research papers in order to personalize recommendations.

In our Collaborative Filtering approach, a candidate paper is qualified for consideration if and only if it cited any of the target paper's references and there exists another paper which cited both the candidate and the target papers simultaneously. Then similarity between the target paper and the qualified candidate papers is calculated and recommend the top-N most similar papers.

Algorithm is given below:



Input: Target Paper

Output: Top-N Recommendation

Given a target paper p_i as a query,

1. Retrieve all the set of references Rf_j of the target paper p_i from the paper-citation relation matrix C.
    a. For each of the references Rf_j, extract all other papers p_{ci} that also cited Rf_j other than the target paper p_i.
2. Retrieve all the set of citations Cf_j of the target paper p_i from the paper-citation relation matrix C.
    a. For each of the citations Cf_j, extract all other papers p_{ri} that Cf_j referenced other than the target paper p_i.
3. Qualify all the candidate papers p_c from p_{ci} that has been referenced by at least any of the p_{ri}
4. Measure the extent of similarity $W^{p_i \rightarrow p_c}$ between the target paper p_i and the qualified candidate papers p_c
5. Recommend the top-N most similar papers to the user.

8. Hybrid Pipelined Approach:

Pipelined hybrids implement a staged process in which several techniques sequentially build on each other before the final one produces recommendations for the user. The pipelined hybrid variants differentiate themselves mainly according to the type of output they produce for the next stage.

In other words, a preceding component may either preprocess input data to build a model that is exploited by the subsequent stage or deliver a recommendation list for further refinement.

In our implementation, we have used output of content-based filtering to collaborative filtering as input. The KNN algorithm, for content-based filtering find out the nearest neighbors for an input query by using title and abstract of research papers from the dataset. Then output is given to collaborative filtering where research papers' citation and references are used for training. The output is top N recommendation using hybrid Pipelined Approach.

9. System Evaluation:

In the web application user can provide ratings to each research paper he was referred to.
Users will give a rating to the recommendations generated by algorithms through a web interface and the algorithm with the highest average rating is considered as the best algorithm

10. Web Application Screenshots:

Front end is developed using bootstrap and HTML5 and CSS. We have used python Flask REST API for communication from front end to backend model.

The Web Application Consists of the Content based and Collaborative based and Pipeline Hybrid models (KNN, TFIDF, Word2Vector, Hybrid)

When User provides input and click on Get recommendations it will provide recommendations based on different models.

For system evaluation users are also allowed to submit a rating for each recommendation and then based on user rating recommendations are evaluated.

This application is deployed in Amazon AWS. link:

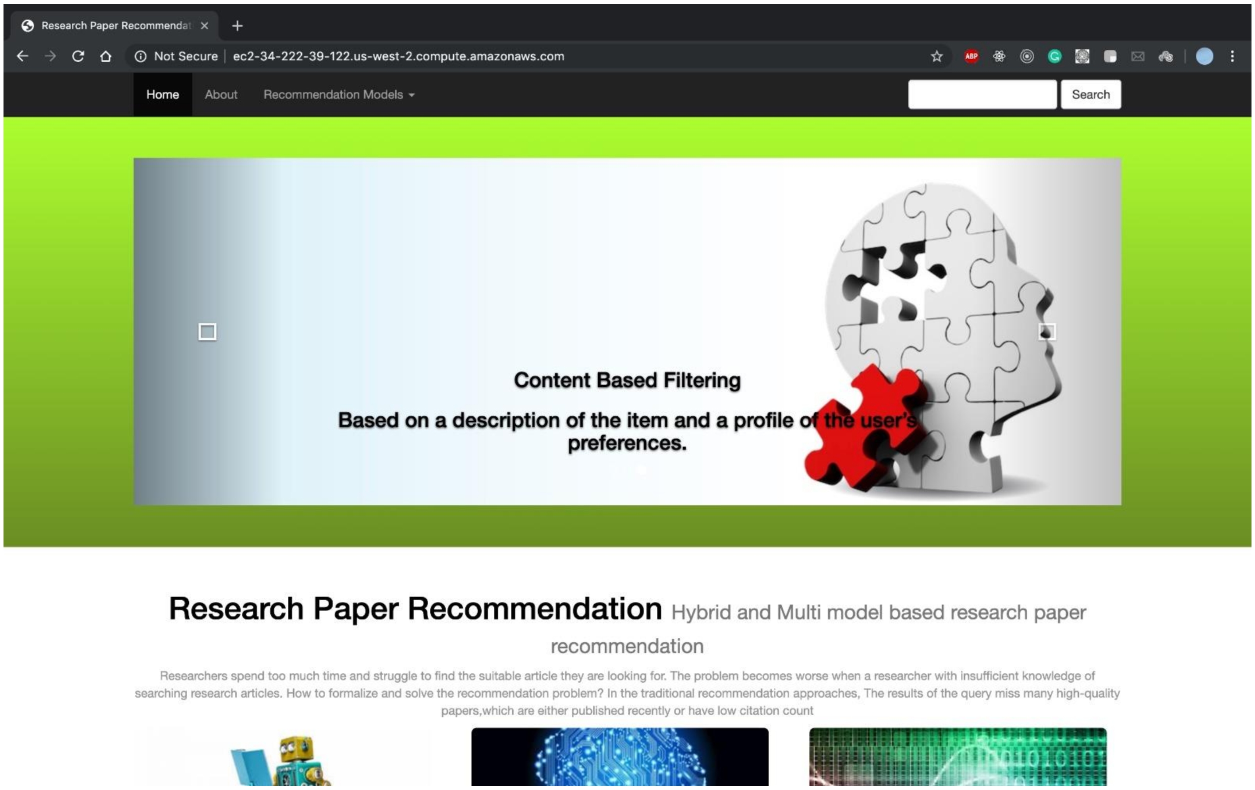
http://ec2-54-71-205-30.us-west-2.compute.amazonaws.com/

Two Demo videos Presentation demo: https://youtu.be/Ra8Y6hcaJKc

Web Application demo: https://youtu.be/PKIZtyGzR3I

Screenshots:

Screenshot1: You should be able to view the homepage as below:


Research Paper Recommendat
Not Secure | ec2-34-222-39-122.us-west-2.compute.amazonaws.com
Home
About
Recommendation Models
Search

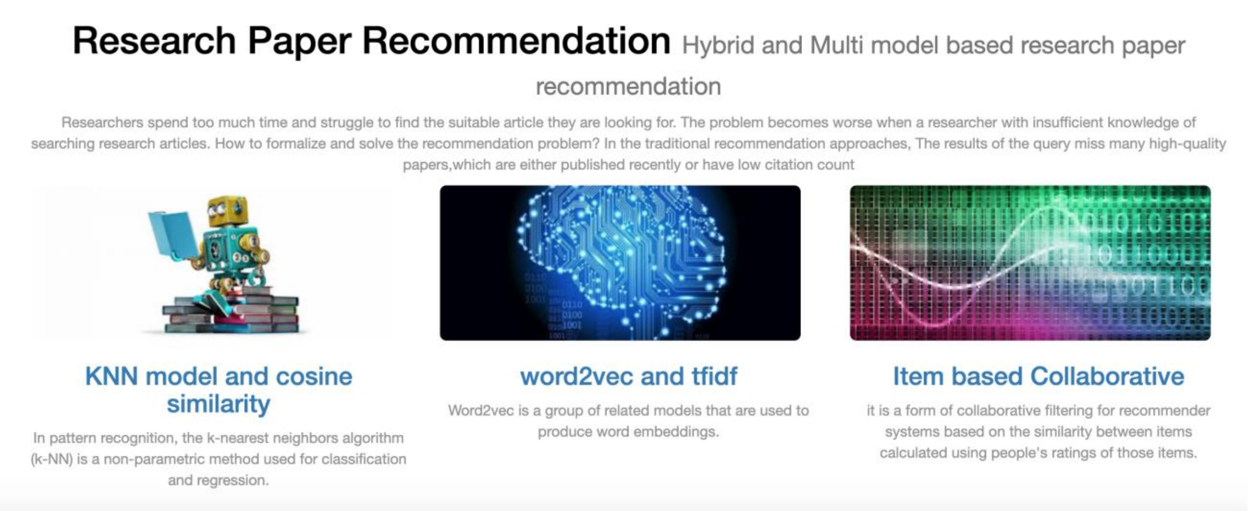
Content Based Filtering

Based on a description of the item and a profile of the user's preferences.

Research Paper Recommendation Hybrid and Multi model based research paper recommendation

Researchers spend too much time and struggle to find the suitable article they are looking for. The problem becomes worse when a researcher with insufficient knowledge of searching research articles. How to formalize and solve the recommendation problem? In the traditional recommendation approaches, The results of the query miss many high-quality papers,which are either published recently or have low citation count

screenshot2:



Research Paper Recommendation Hybrid and Multi model based research paper recommendation

Researchers spend too much time and struggle to find the suitable article they are looking for. The problem becomes worse when a researcher with insufficient knowledge of searching research articles. How to formalize and solve the recommendation problem? In the traditional recommendation approaches, The results of the query miss many high-quality papers,which are either published recently or have low citation count

KNN model and cosine similarity

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.

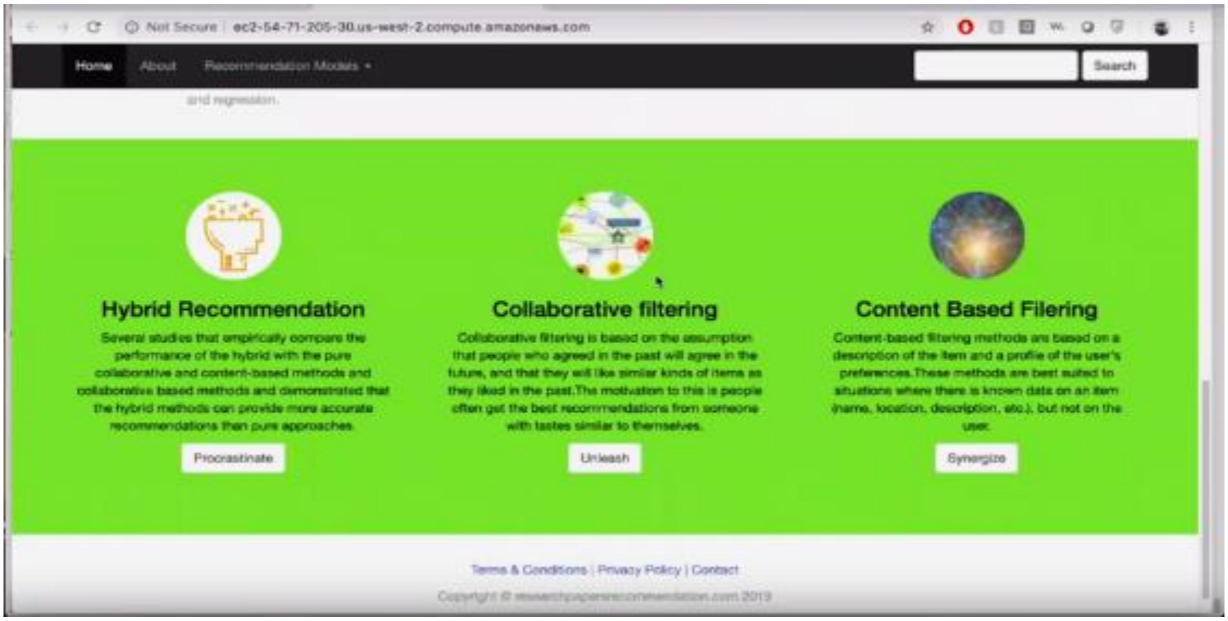
word2vec and tfidf

Word2vec is a group of related models that are used to produce word embeddings.

Item based Collaborative

it is a form of collaborative filtering for recommender systems based on the similarity between items calculated using people's ratings of those items.

Screenshot 3: after user scroll down homepage user can see below images.



Not Secure | ec2-54-71-205-30.us-west-2.compute.amazonaws.com
Home
About
Recommendation Models ▾
Search

and regression.

Hybrid Recommendation

Several studies that empirically compare the performance of the hybrid with the pure collaborative and content-based methods and collaborative based methods and demonstrated that the hybrid methods can provide more accurate recommendations than pure approaches.

Collaborative filtering

Collaborative filtering is based on the assumption that people who agreed in the past will agree in the future, and that they will like similar kinds of items as they liked in the past.The motivation to this is people often get the best recommendations from someone with tastes similar to themselves.

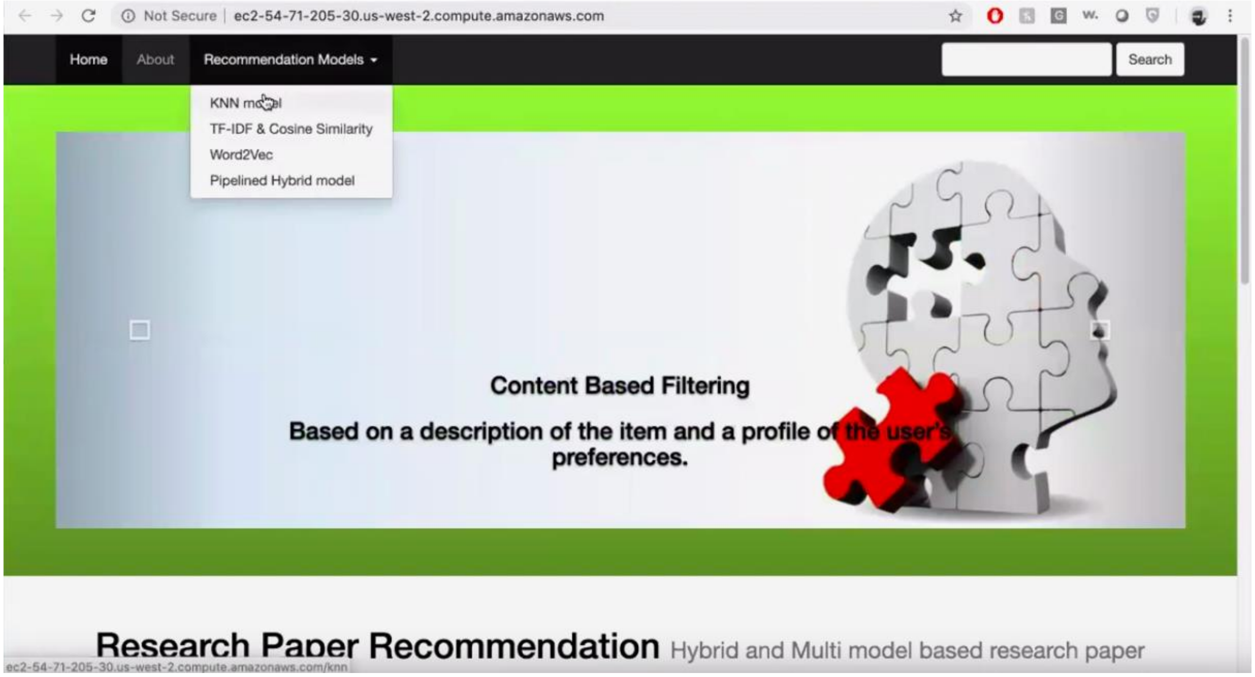
Content Based Filering

Content-based filtering methods are based on a description of the item and a profile of the user's preferences.These methods are best suited to situations where there is known data on an item (name, location, description, etc.), but not on the user.

Terms & Conditions | Privacy Policy | Contact

Copyright © [illegible] 2019

Screenshot 4: When use click on recommendation Models tab, user can see all models in drop down.

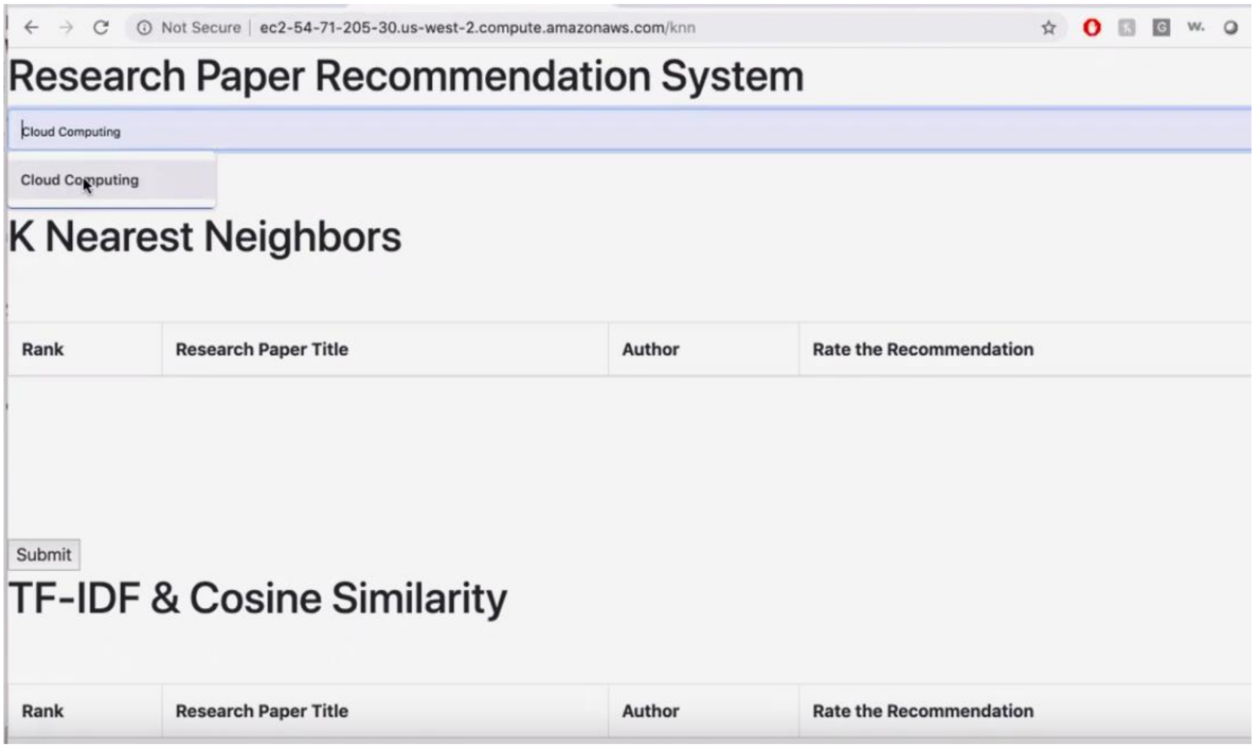

Not Secure | ec2-54-71-205-30.us-west-2.compute.amazonaws.com
Home
About
Recommendation Models
Search
KNN model
TF-IDF & Cosine Similarity
Word2Vec
Pipelined Hybrid model
ec2-54-71-205-30.us-west-2.compute.amazonaws.com/knn

Content Based Filtering

Based on a description of the item and a profile of the user's preferences.

Research Paper Recommendation Hybrid and Multi model based research paper

Screenshot 5: After clicking on recommendation models users are redirected to below page which allow user to search paper.


Not Secure | ec2-54-71-205-30.us-west-2.compute.amazonaws.com/knn
Cloud Computing
Cloud Computing
Submit

Research Paper Recommendation System

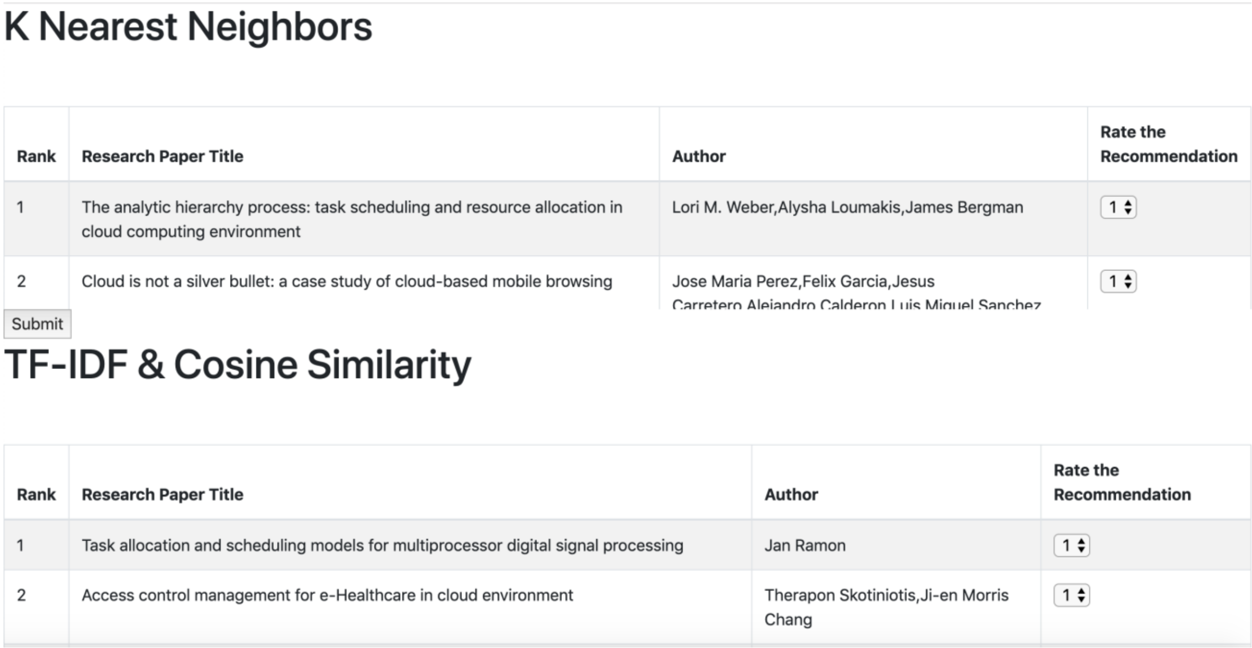
K Nearest Neighbors

Rank	Research Paper Title	Author	Rate the Recommendation

TF-IDF & Cosine Similarity

Rank	Research Paper Title	Author	Rate the Recommendation

Screenshot 6: After user search paper for ex. “Cloud computing”, user gets recommendation papers for all the models. Also, users can send rating for each recommendation.



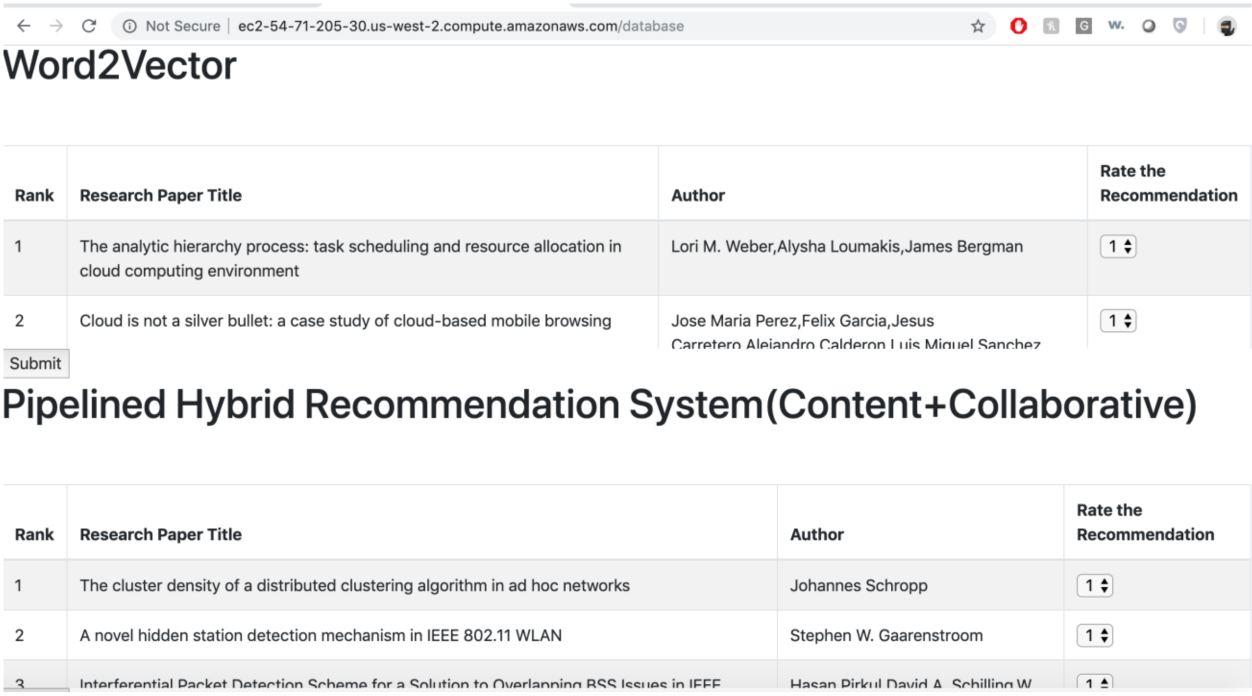
K Nearest Neighbors

Rank	Research Paper Title	Author	Rate the Recommendation
1	The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment	Lori M. Weber,Alysha Loumakis,James Bergman	1
2	Cloud is not a silver bullet: a case study of cloud-based mobile browsing	Jose Maria Perez,Felix Garcia,Jesus Carretero Alejandro Calderon Luis Miguel Sanchez	1

TF-IDF & Cosine Similarity

Rank	Research Paper Title	Author	Rate the Recommendation
1	Task allocation and scheduling models for multiprocessor digital signal processing	Jan Ramon	1
2	Access control management for e-Healthcare in cloud environment	Therapon Skotiniotis,Ji-en Morris Chang	1

If user provides the input and click on Get Recommendations, then the below recommendations will be provided to User.



Not Secure | ec2-54-71-205-30.us-west-2.compute.amazonaws.com/database

Word2Vector

Rank	Research Paper Title	Author	Rate the Recommendation
1	The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment	Lori M. Weber,Alysha Loumakis,James Bergman	1
2	Cloud is not a silver bullet: a case study of cloud-based mobile browsing	Jose Maria Perez,Felix Garcia,Jesus Carretero Alejandro Calderon Luis Miguel Sanchez	1

Pipelined Hybrid Recommendation System(Content+Collaborative)

Rank	Research Paper Title	Author	Rate the Recommendation
1	The cluster density of a distributed clustering algorithm in ad hoc networks	Johannes Schropp	1
2	A novel hidden station detection mechanism in IEEE 802.11 WLAN	Stephen W. Gaarenstroom	1
3	Interferential Packet Detection Scheme for a Solution to Overlapping BSS Issues in IEEE	Hasan Pirkul David A. Schilling W.	1

Users can also provide rating to each recommendation which will be saved in the database and helpful in calculating the efficiency of the algorithms.

11. Conclusion:

We have successfully implemented the web-based research paper recommendation based on hybrid and multi model approaches. Moreover, provided the detail analysis and comparison report of total 4 models (KNN, TFIDF, Word-Vector and Hybrid). Our experimental evaluation and study show that hybrid-based approach provides better recommendations.

12. Application Links:

- This is the GitHub link : https://github.com/tramatejaswini/ResearchPaper_Recommendation_System
- This application is deployed in Amazon AWS. link: http://ec2-54-71-205-30.us-west-2.compute.amazonaws.com/
- videos Presentation demo: https://youtu.be/Ra8Y6hcaJKc
- Web Application demo: https://youtu.be/PKIZtyGzR3I

13. Acknowledgement

Thanks to Professor Shih Yu Chang for his timely guidance and the motivation he provided to go beyond expectations.

14. References

1. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184516
2. https://aip.scitation.org/doi/10.1063/1.5041583
3. https://www.researchgate.net/publication/200610399_Scienstein_A_Research_Paper_Recommender_System
4. https://link.springer.com/article/10.1007/s00799-015-0156-0
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5628815/
6. https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50
7. https://www.quackit.com/html/templates/business_website_templates.cfm
8. Bulut, B., Kaya, B., Alhajj, R., & Kaya, M. (2018, August). A Paper Recommendation System Based on User's Research Interests. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*(pp. 911-915). IEEE.
9. Beel, J., Langer, S., Genzmehr, M., & Nürnberger, A. (2013, July). Introducing Docear's research paper recommender system. In *JCDL* (pp. 459-460).

10. Waheed, W., Imran, M., Raza, B., Malik, A. K., & Khattak, H. A. (2019). A Hybrid Approach Toward Research Paper Recommendation Using Centrality Measures and Author Ranking. *IEEE Access*, *7*, 33145-33158.
11. Neethukrishnan, K. V., & Swaraj, K. P. (2017, February). Ontology based research paper recommendation using personal ontology similarity method. In *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-4). IEEE.
12. Xue, H., Guo, J., Lan, Y., & Cao, L. (2014, August). Personalized paper recommendation in online social scholar system. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 612-619). IEEE Press.
13. Nishioka, C., & Ogata, H. (2018, May). Research Paper Recommender System for University Students on the E-Book System. In *Proceedings of the 18th ACM/IEEE on* Joint Conference on Digital Libraries (pp. 369-370). ACM.
14. Beel, J., Genzmehr, M., Langer, S., Nürnberger, A., & Gipp, B. (2013, October). A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation (pp. 7-14). ACM.
15. Xue, H., Guo, J., Lan, Y., & Cao, L. (2014, August). Personalized paper recommendation in online social scholar system. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 612-619). IEEE Press.
16. Hong, K., Jeon, H., & Jeon, C. (2012, August). UserProfile-based personalized research paper recommendation system. In 2012 8th International Conference on Computing and Networking Technology (INC, ICCIS and ICMIC) (pp. 134-138). IEEE.

==============================END================================