CMPE 256
Large Scale Analytics

Research Paper based recommendation -based on Hybrid and Multi model

Presented By: Team 02

Rama Tejaswini Thotapalli(013785681)
Nithya Kuchadi (013769665)
Pranjal Kumar Patel (013748709)
Jaykumar Patel (013756210)
Premal Dattatray Samale(012566333)

Agenda

- Problem Statement
- Approach
- Data Characteristics
- Data Preprocessing
- Solution Implementation
- Solution Evaluation

Website Link: http://ec2-34-217-84-100.us-west-2.compute.amazonaws.com/

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.

Literature Survey

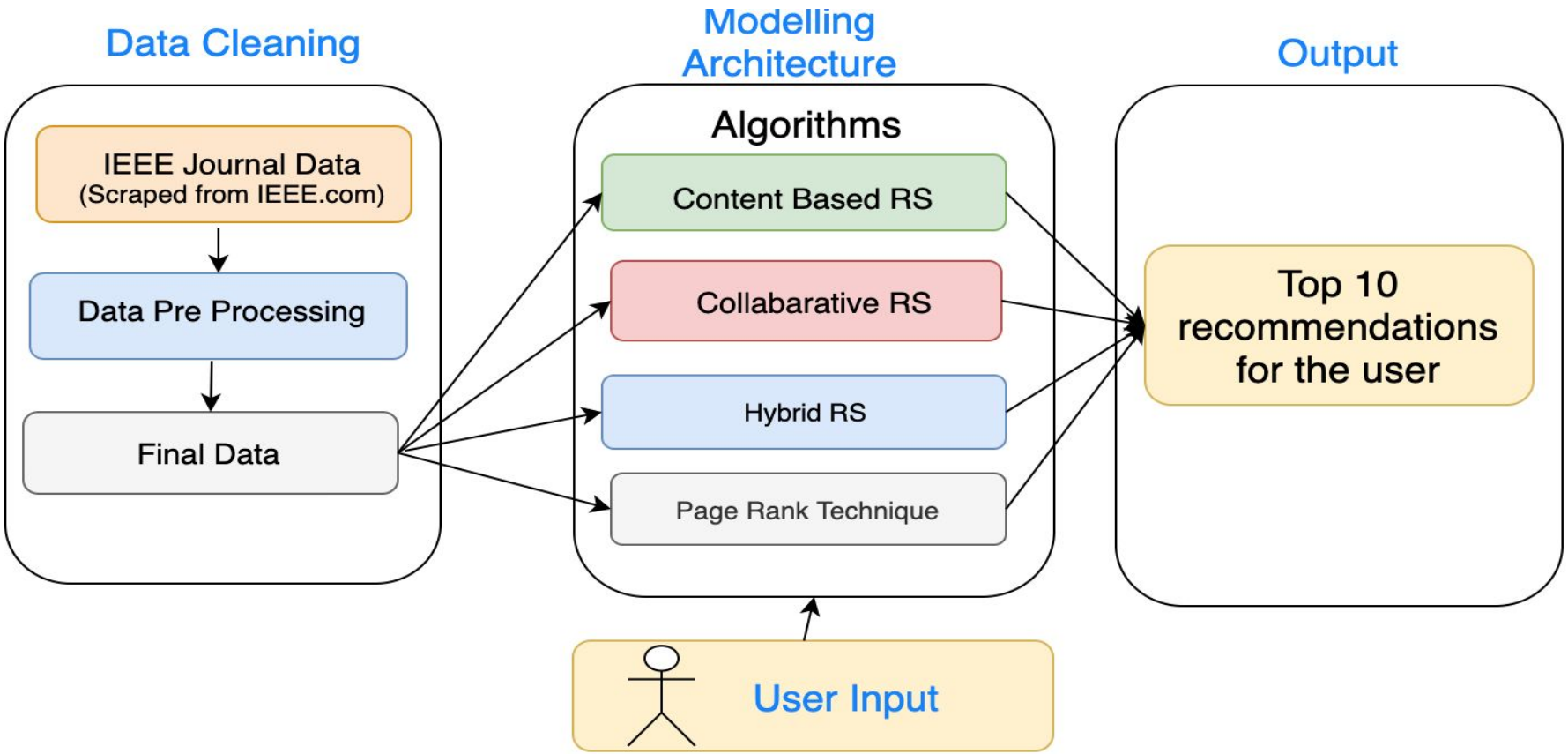
Methodologies	Limitation
Collaborative filtering	cold start problem, does not capture the semantics of the user interests
Matrix factorization based methods	usually time consuming, and it would become even more challenging when the paper collection is extremely large
citation-based method	problem with 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.

Our Approach

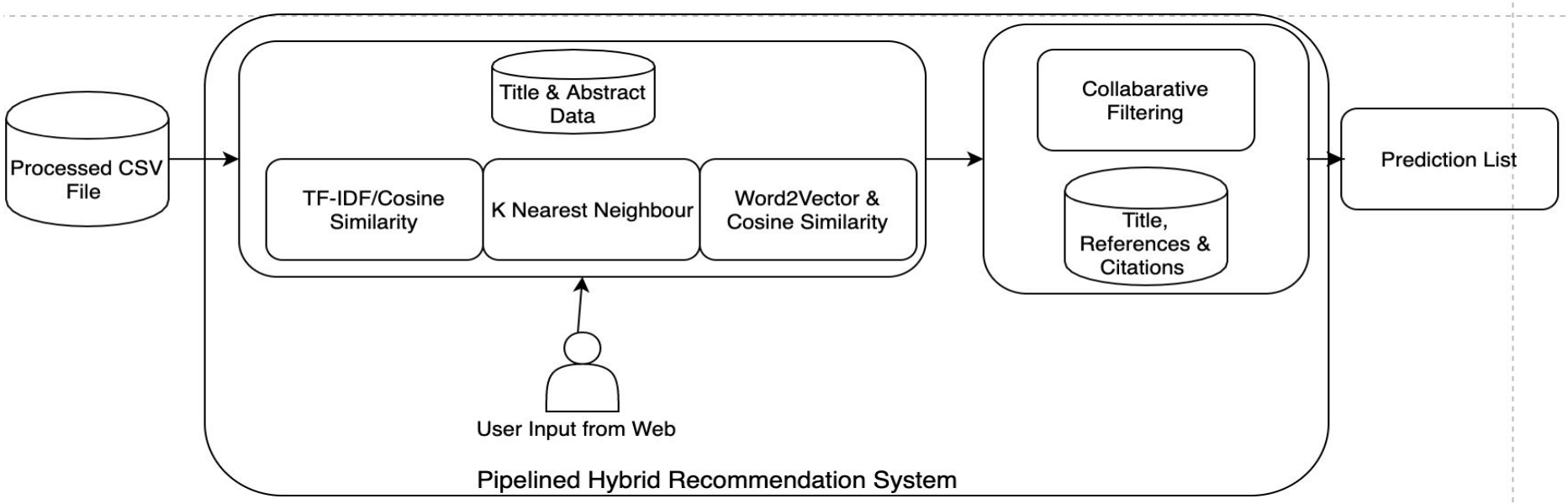
To overcome all the limitations of the above mentioned approaches and as we have large set of data, we are using the below mentioned models with features like title and abstract, References, Citations and Author.

- Content Based
  - TF-IDF & cosine similarity
  - K Nearest Neighbour 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

Implementation Block Diagram

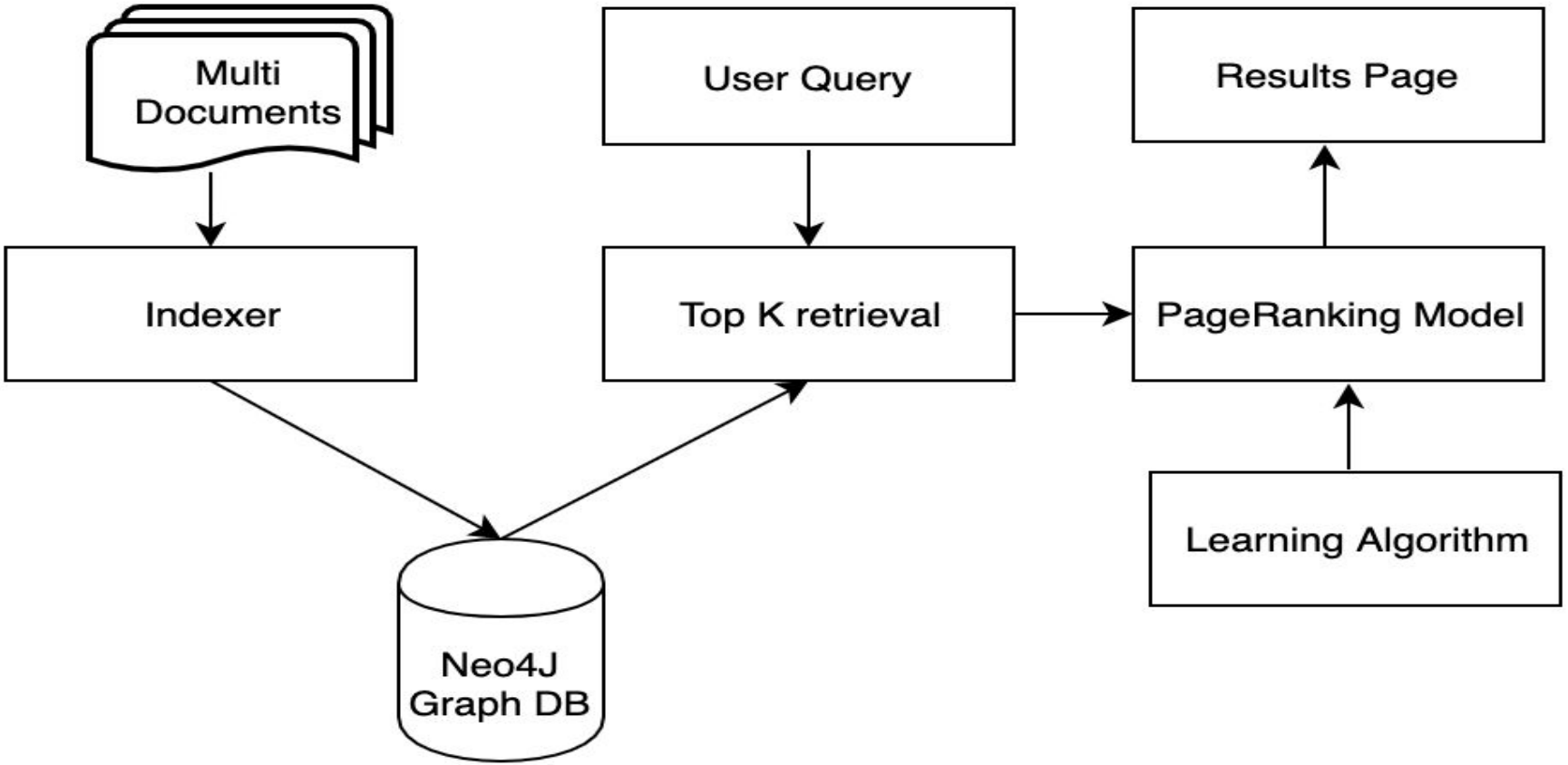


Hybrid RS Architecture Diagram



Content based Recommendation system output is sent to Collaborative Filtering approach

Page Rank Architecture



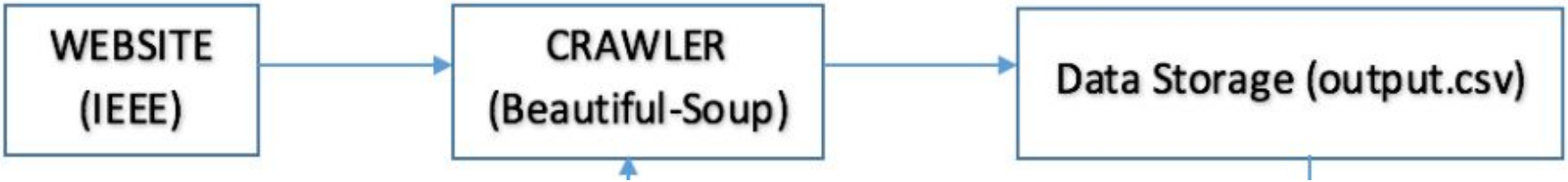
Acquire Data

Web Scraping

We have scraped the research paper data from the http://www.ieeeorg.com where we have extracted the titles , abstract , References, Keywords, id of the papers. Some of the data we have scraped using Parse hub tool.

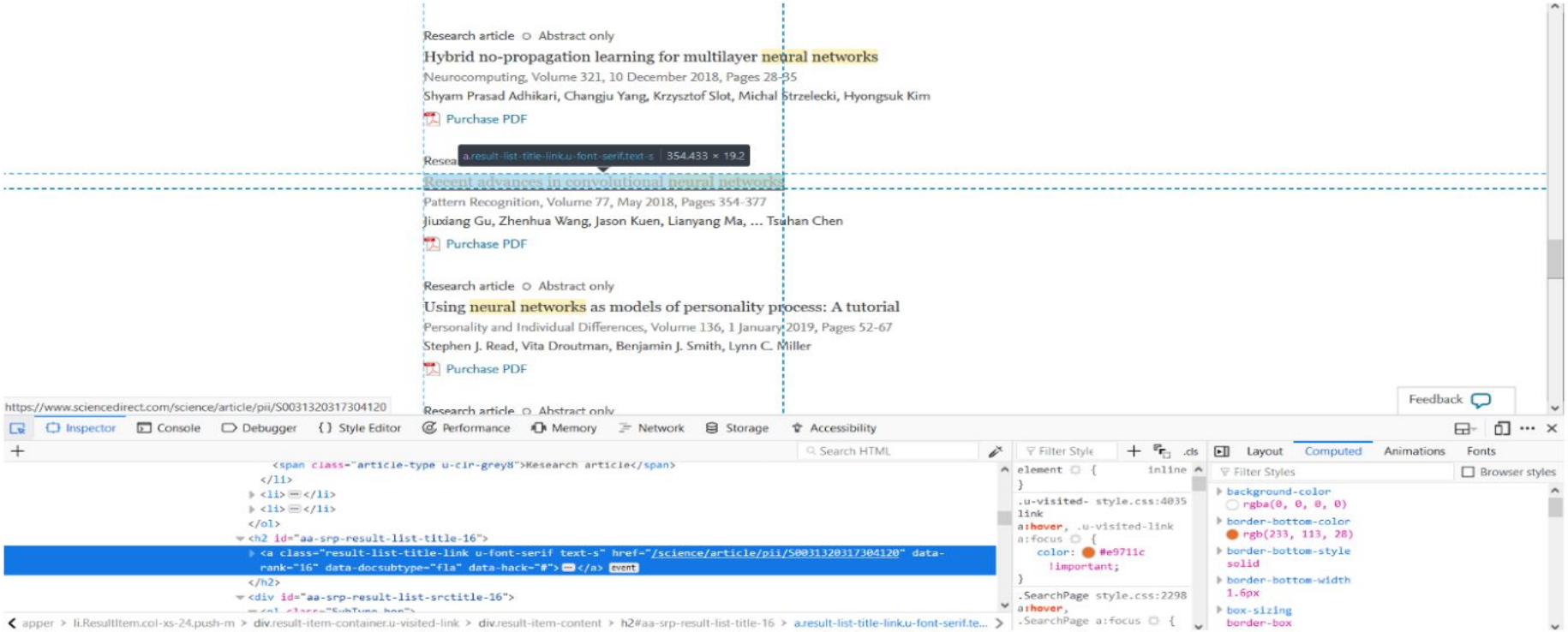
- Python (3.5)
- *BeautifulSoup* library for handling the text extraction from the web page's source code (HTML and CSS)
- *requests* library for handling the interaction with the web page (Using HTTP requests)

Web scraped data



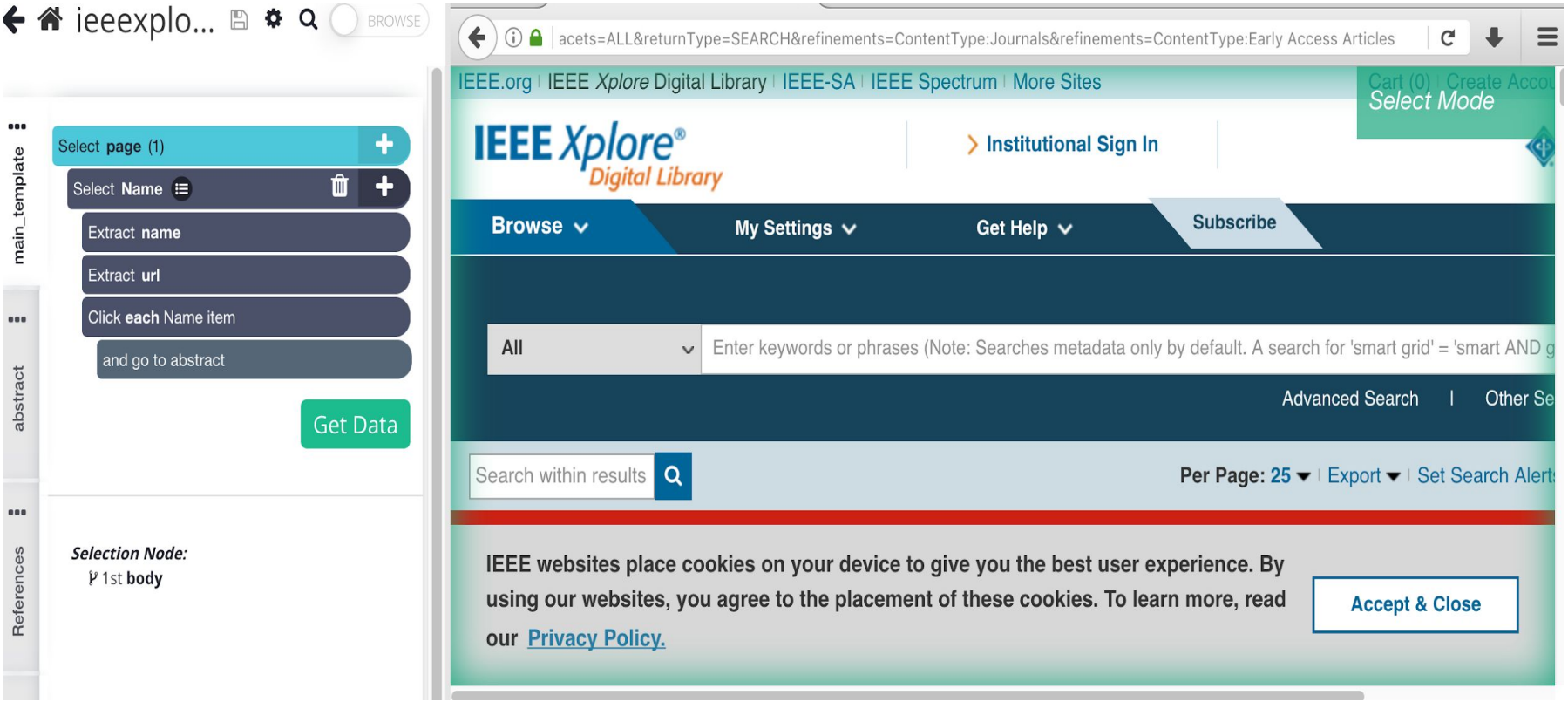
- Using the Beautiful soup framework we will get the entire data into a html where we can search the data using find_ all API of rows, columns and a links.
- Once the data is extracted we will store it into the csv file.

Inspecting the rows and column of the required data



Inspect element of a web page

Web Scraping using the Parse hub tool



Scraped Data

Titles_name	Titles_url	Titles_selecti	Titles_abstra	Titles_selecti	Titles_selection4_name
SemiBoost: E	https://ieeex	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://ieeex	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://ieeex	Abstract	-"-	3. X. Zhu, Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation", 2002. Google Scholar	
SemiBoost: E	https://ieeex	Abstract	-"-	4. Y. Bengio, O.B. Alleau, N. Le Roux, Semi-Supervised Learning, pp. 193-216, 2006. Google Scholar	
SemiBoost: E	https://ieeex	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://ieeex	Abstract	-"-	6. I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2005. Google Scholar	
SemiBoost: E	https://ieeex	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://ieeex	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://ieeex	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://ieeex	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://ieeex	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://ieeex	Abstract	-"-	12. T. Minka, "Expectation-Maximization As Lower Bound Maximization", 1998. Google Scholar	
SemiBoost: E	https://ieeex	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://ieeex	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://ieeex	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://ieeex	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://ieeex	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://ieeex	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://ieeex	Abstract	-"-	2. J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis., 2004. CrossRef Google Scholar	
Nonconvex C	https://ieeex	Abstract	-"-	3. B. Schlkopf, A.J. Smola, Learning with Kernels: Support Vector Machines Regularization Optimization and Beyond., 2002. Google Scholar	
Nonconvex C	https://ieeex	Abstract	-"-	4. C. Cortes, V. Vapnik, "Support Vector Networks", Machine Learning, vol. 20, pp. 273-297, 1995. CrossRef Google Scholar	
Nonconvex C	https://ieeex	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	

Data Characteristics

Below is an example:

{"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."

}

Data Pre-Processing Steps

- We need to extract different data for different algorithms, But before that the complete data should be processed.
- Cleaning of the data should be done before sending this data to models as some of the columns has

  1. Null Values

  2. Sparse Data

  3. Some papers do not have title and abstract; some have title only without abstract and some have abstract without title

Data Pre-Processing Steps

- 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.

	A	B
1	Multisymplectic Spectral Methods for the Gross-Pitaevskii Equation	Recently, Bridges and Reich introduced the concept of multisymplectic sp
2	Improved Secret Image Sharing Method By Encoding Shared Values With .	0
3	Leveraging legacy code to deploy desktop applications on the web	Xax is a browser plugin model that enables developers to leverage existir
4	Development of Remote Monitoring and Control Device for 50KW PV Syste	0
5	Preliminary Design of a Network Protocol Learning Tool Based on the Com	The purpose of this study is to develop a learning tool for high school stu
6	Link-time compaction of MIPS programs	Embedded systems often have limited amounts of available memory, thu
7	COMPARING GNG3D AND QUADRIC ERROR METRICS METHODS TO SIMP	0
8	Knowledge Engineering for Affective Bi-Modal Interaction in Mobile Devic	This paper focuses on knowledge engineering for the development of a s
9	A COMPUTATIONAL SALIENCY MODEL INTEGRATING SACCADE PROGRAM	0
10	Speech training systems using lateral shapes of vocal tract and F1-F2 diag	Three speech training systems for hearing-impaired children were desigr
11	Vectorial fast correlation attacks.	0
12	Design of an audio-visual speech corpus for the czech audio-visual speech	0
13	Algorithms for the Construction of Digital Convex Fuzzy Hulls.	0
14	Fur Visualisation for Computer Game Engines and Real-Time Rendering	0
15	Cleaneval: a Competition for Cleaning Web Pages.	0
16	Software Evolution through Transformations.	0
17	Simulation of a vision steering system for road vehicles	0
18	A Platform for Disaster Response Planning with Interdependency Simulati	0
19	Reasonig about Set-Oriented Methods in Object Databases.	0
20	Comparison of GARCH, Neural Network and Support Vector Machine in Fi	This article applied GARCH model instead AR or ARMA model to compare
21	A Self-Stabilizing Algorithm for Finding the Cutting Center of a Tree.	0
22	A methodology for the physically accurate visualisation of roman polychro	This paper describes the design and implementation of a methodology fo
23	Identifying Psychological Theme Words from Emotion Annotated Interviev	Recent achievements in Natural Language Processing (NLP) and Psycholo
24	A pedestrian navigation method for user's safe and easy wayfinding	In recent years, most of mobile phones have a function of pedestrian nav

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 tool
4	Development of Remote Monitoring and Control Device for 50KW PV System Bas	Development of Remote Monitoring and Control Device for 50KW PV System Ba
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 fo	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

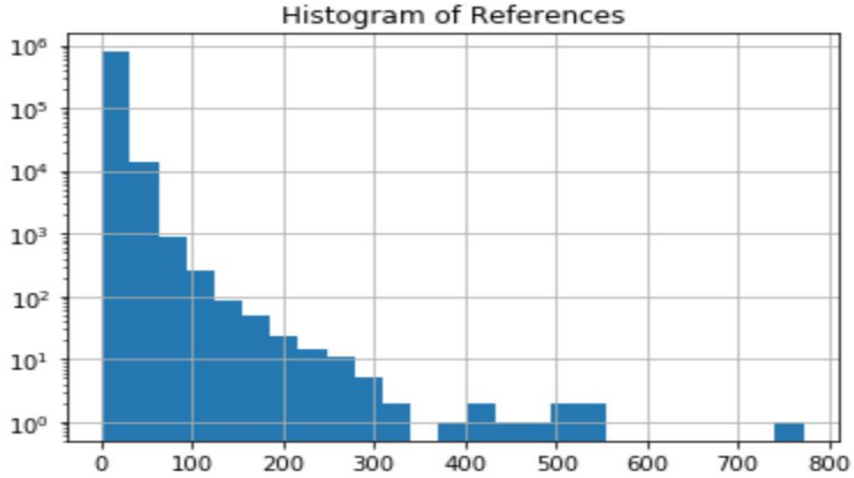
In [49]:
```
import pandas as pd
df = pd.read_csv('papers.csv' , index_col = False)
df = df.loc[:, ~df.columns.str.match('Unnamed')]
df.head(10)
```

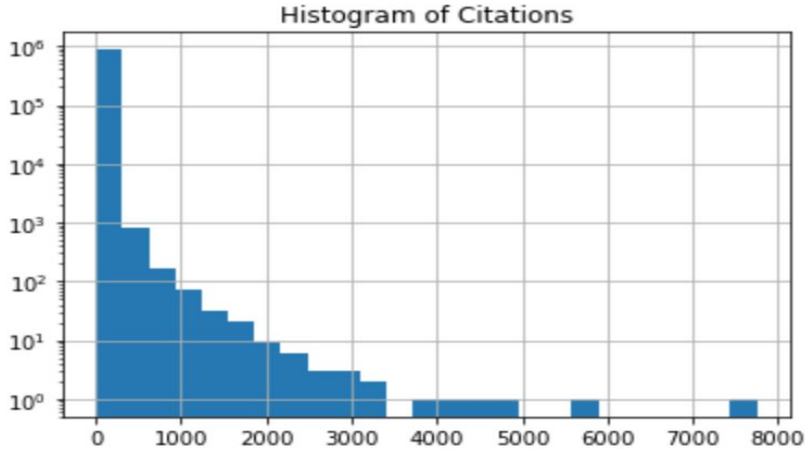
Out[49]:

	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-...

Exploratory data Analysis

```
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()
```





- In Exploratory data Analysis

we will describe the data Characteristics and data statistics in plots and provide visual understanding of 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

Data for final model for Content Based Approach

```
new_df= df[['id','title']]
```

	id	title
0	4ab39729-af77-46f7-a662-16984fb9c1db	Attractor neural networks with activity-depend...
1	4ab3a4cf-1d96-4ce5-ab6f-b3e19fc260de	A characterization of balanced episturmian seq...
2	4ab3a98c-3620-47ec-b578-884ecf4a6206	Exploring the space of a human action
3	4ab3b585-82b4-4207-91dd-b6bce7e27c4e	Generalized upper bounds on the minimum distan...
4	4ab3e768-78c9-4497-8b8e-9e934cb5f2e4	Applying BCMP multi-class queueing networks fo...
5	4ab3f7cd-140b-4e29-99d4-f4e8006c4f65	A Push–Pull Class-C CMOS VCO

Data taken for Collaborative Approaches

```
ref_df= df[['id','title', 'references', 'citation']]
ref_df.head(10)
```

	id	title	references	citation
0	4ab39729-af77-46f7-a662-16984fb9c1db	Attractor neural networks with activity-depend...	4017c9d2-9845-4ad2-ad5b-ba65523727c5,b1187381-...	4017c9d2-9845-4ad2-ad5b-ba65523727c5
1	4ab3a4cf-1d96-4ce5-ab6f-b3e19fc260de	A characterization of balanced episturmian seq...	1c655ee2-067d-4bc4-b8cc-bc779e9a7f10,2e4e57ca-...	1c655ee2-067d-4bc4-b8cc-bc779e9a7f10
2	4ab3a98c-3620-47ec-b578-884ecf4a6206	Exploring the space of a human action	056116c1-9e7a-4f9b-a918-44eb199e67d6,05ac52a1-...	056116c1-9e7a-4f9b-a918-44eb199e67d6
3	4ab3b585-82b4-4207-91dd-b6bce7e27c4e	Generalized upper bounds on the minimum distan...	01a765b8-0cb3-495c-996f-29c36756b435,5dbc8ccb-...	01a765b8-0cb3-495c-996f-29c36756b435
4	4ab3e768-78c9-4497-8b8e-9e934cb5f2e4	Applying BCMP multi-class queueing networks fo...	1c26e228-57d2-4b2c-b0c9-8d5851c17fac,75399207-...	1c26e228-57d2-4b2c-b0c9-8d5851c17fac
5	4ab3f7cd-140b-4e29-99d4-f4e8006c4f65	A Push–Pull Class-C CMOS VCO	0a09db01-264a-4bdf-942c-d33cceb35d3c,36c942df-...	0a09db01-264a-4bdf-942c-d33cceb35d3c
6	4ab404e2-6f4b-4fb4-b093-50775e765b13	On computability of pattern recognition problems	505f493b-e09d-444d-9ee2-5e5db6a5b8ac	505f493b-e09d-444d-9ee2-5e5db6a5b8ac
7	4ab4244d-fb3e-49a3-b125-367df3d8e6ba	Manipulating biological and mechanical micro-o...	5ecd70e1-7ccc-4b2f-ac09-b91953cca5cd,7fa711e9-...	5ecd70e1-7ccc-4b2f-ac09-b91953cca5cd

Data taken for Pagerank Approach

The data taken into page rank algorithm is the Json file of the same data which is directly loaded into the Neo4j Graph Database

{"authors": ["Tegegne Marew", "Doo-Hwan Bae"], "n_citation": 1, "references": ["2134bf3b-fd89-4724-90ce-5993b4fa3218", "906c17e0-db09-407b-b760-41df5a3f02
"94f4382e-cfa6-4aec-92b8-3711fc55da54", "9f172585-8d42-4fce-b6ae-aede321f3fd4", "a3aee287-efd0-4b9d-9cda-d47dd192c9f4", "a9a7fd07-ef71-4b3c-8fcf-d7fe114d2
"d63dd4ae-4b30-484b-8ffc-88d21839ddad"], "title": "Using Classpects for Integrating Non-Functional and Functional Requirements.", "venue": "international
on software engineering", "year": 2006, "id": "01f1d231-80ae-4cce-b56c-9d821e0924d0"}
{"authors": ["Lei Zhang", "Xuan Zhang", "Meiping Chai", "Yibing Tan", "Shigeru Miyake", "Yoji Taniguchi", "Jun Hosoya", "Ryota Mibe"], "n_citation": 2, "re
["3e3b524c-70c5-4008-b349-fd7ae950e655", "4929a7b3-0d81-4123-973a-82b075304713", "563bdfaf-91c2-4440-b146-54954bf7ee48", "6e9dad6f-50db-467b-959c-18881450e
"a1485664-bc1c-4cc7-9cc2-6234571d4a70", "de8bc699-183a-4e37-8883-c01565ccfe4f", "e41c93b8-75b8-48b5-a5ad-5c2833f7cd0d", "e8218c23-52bb-4ec8-9574-98181f7e3
"title": "Solution Proposals for Japan-Oriented Offshore Software Development in China", "venue": "international conference on software engineering", "yea
"id": "0e6ce7a9-6456-437b-9f3f-4bda192a6fae"}
{"authors": ["Dongyun Liu", "Hong Mei"], "n_citation": 39, "references": ["4b837f17-7e38-4175-82bc-daa37f162933", "65acba26-3449-48c8-ac84-83465430d11b",
5f2b-44b4-977a-9755a6484ac7"], "title": "Mapping Requirements to Software Architecture by Feature-Orientation.", "venue": "international conference on soft
engineering", "year": 2003, "id": "10c7185a-f2b7-4810-b1d6-1340c2949922"}
{"abstract": "IEEE 802.11e Medium Access Control (MAC) is an emerging extension of the IEEE 802.11 Wireless Local Area Network (WLAN) standard to support
Service (QoS). The IEEE 802.11e uses both centrally-controlled as well as contention-based channel access mechanisms to transfer data across the wireless
also provides the mechanism to specify and negotiate the resource based on the user's QoS requirement. This paper presents a MAC-level QoS signaling for IE
WLAN and addresses its interaction with higher layer signaling protocols including Resource ReSerVation Protocol (RSVP) and Subnet Bandwidth Manager (SBM).
explain a novel way of setting up sidestream connections for direct station-to-station streaming within an 802.11e WLAN.", "authors": ["N. Sai Shankar", "S
Choi"], "n_citation": 50, "title": "QoS Signaling for Parameterized Traffic in IEEE 802.11e Wireless LANs", "venue": "Lecture Notes in Computer Science",
2002, "id": "11f0bd37-ae5a-43e6-b14a-a59bc00fdd90"}

PageRank Implementation

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 Simple , If B, C, D links to A then page rank of A is

$$PR(A) = PR(B) + PR(C) + PR(D).$$

Detailed Example

Suppose if B had a link to pages C and A, page C had a link to page A, and page D had links to all three pages. Then the page rank of A is

B-->A,C; C-->A; D-->A, B, C

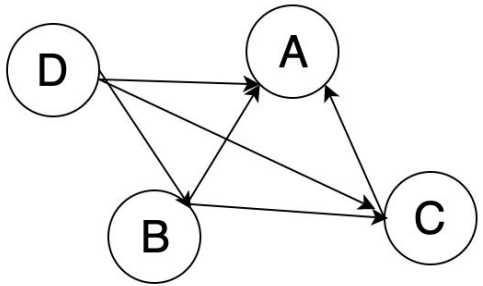
$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$

In other words, the PageRank conferred by an outbound link is equal to the document's own PageRank score divided by the number of outbound links L( ).

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

In the general case, the PageRank value for any page u can be expressed as:

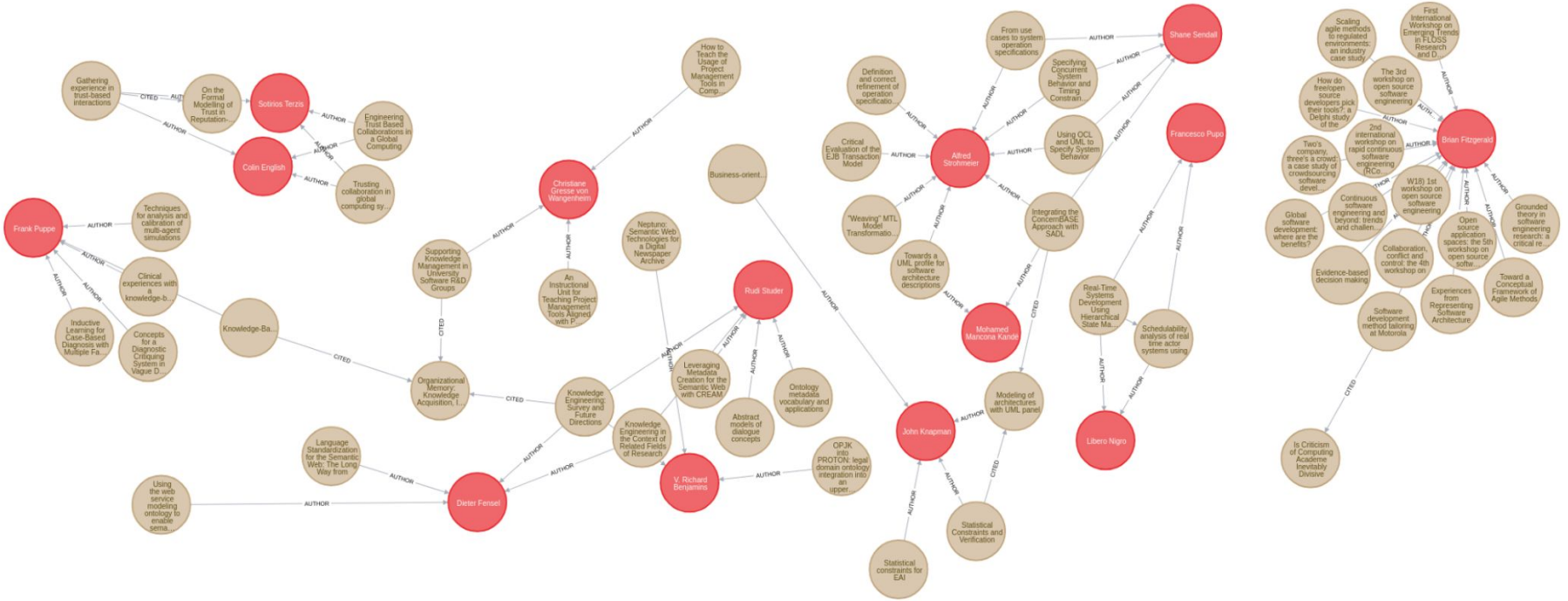
$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$



Implementation

- Used below tools and library for the implementation of recommendation system
    - Neo4j
    - Neo4j graph algorithms
    - Neo4j APOC
    - Graphaware's NLP plugin
- Taken the below mentioned graph for the connection between the articles and references

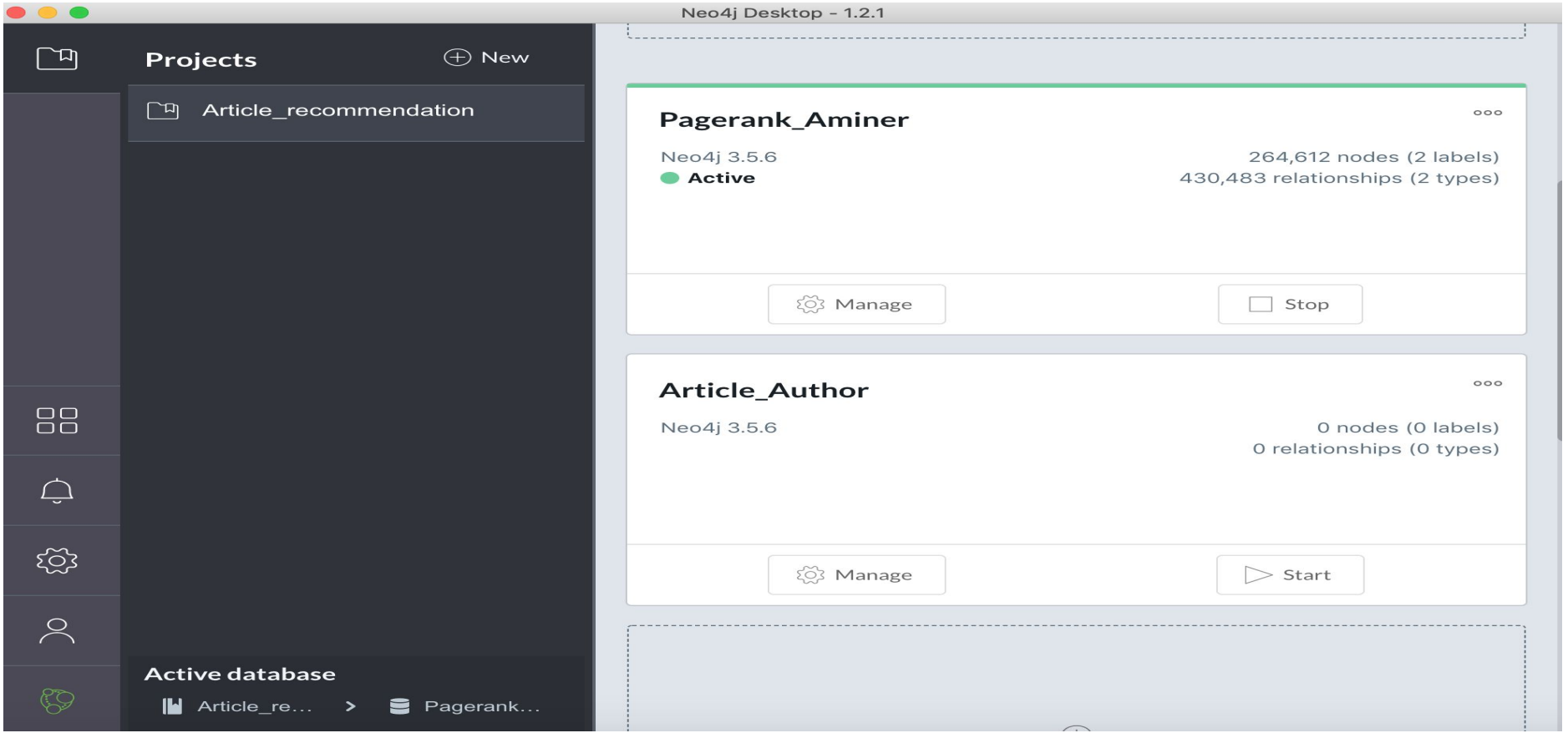
Article and Author Network Graph



PageRank Algorithm

- Store the data into Neo4j graph database.
- Search the database with the titles of the search word
- 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.
- The articles which are more relevent and has highest Pagerank will be provided to the User

Neo4j Desktop View



Queries given to Neo4j to search the word in the database.

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

Here the search term is suppose Social Networks

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

Table | Text | Code

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.

Now we will call the pagerank procedure to get the pagerank of each article

```
CALL db.index.fulltext.queryNodes("articlesAll",
$searchTerm)YIELD node
WITH collect(node) as articles
CALL algo.pageRank.stream('Article', 'REFERENCES', {
sourceNodes: articles})
YIELD nodeId, score
WITH nodeId,score
ORDER BY score DESC
LIMIT 10
RETURN algo.getNodeById(nodeId).title as article, score
```

Final Titles of articles of the search term ‘Social Networks’

```
$ CALL db.index.fulltext.queryNodes("articlesAll", $searchTerm) YIELD node WITH collect(node) as articles CALL algo.pageRank.st…
```

Table | Text | Code

article	score
"The anatomy of a large-scale hypertextual Web search engine"	22.183792999999998
"The Structure and Function of Complex Networks"	18.461975
"Statistical mechanics of complex networks"	16.600644499999998
"Mining the network value of customers"	14.802673500000001
"Introduction to Modern Information Retrieval"	12.52662
"Authoritative sources in a hyperlinked environment"	12.238283000000001
"Authoritative sources in a hyperlinked environment"	11.948126999999998
"On power-law relationships of the Internet topology"	11.932342499999997
"Measurement and analysis of online social networks"	11.6072775
"GroupLens: an open architecture for collaborative filtering of netnews"	11.502463500000001

Started streaming 10 records after 10308 ms and completed after 10310 ms.

Content Based Filtering

- To recommend papers based on the titles.
- Implemented 3 models
    1. TF-IDF & Cosine similarity
    2. K Nearest Neighbors
    3. Word2Vec

TF-IDF & Cosine Similarity

Tf-idf is a transformation you apply to texts to get two real-valued vectors.

- TF: Term frequency. This is simply the frequency of a word in a document.
- IDF: Inverse Document Frequency . This is the universe of document frequency among the whole corpus of documents.

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 $\mathbf{d}_2$ and **q** are tf-idf vectors.

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

TF-IDF & Cosine Similarity

```
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)
```

TF-IDF Implementation

```
# 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)
```

TF-IDF Result

Input:

```
recommendations('neural networks')
```

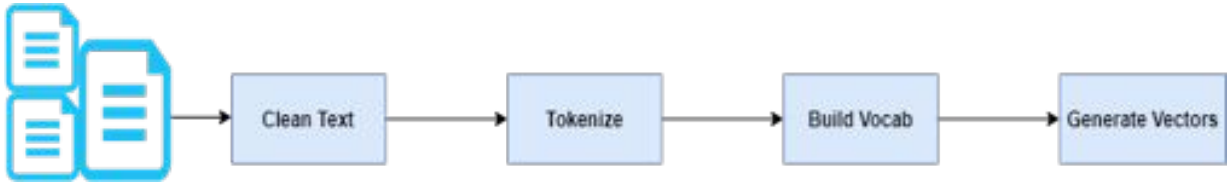
Output:

```
['Recurrent neural network as a linear attractor for pattern association',
 'Rounding Methods for Neural Networks with Low Resolution Synaptic Weights',
 'Layers of learning: facilitation in the distributed classroom',
 'Spike timing dependent synaptic plasticity in biological systems',
 'Sequential Memory: A Putative Neural and Synaptic Dynamical Mechanism']
```

K Nearest Neighbors

For a given target paper, KNN algorithm uses Bag of Words model and Euclidean distance to recommend K nearest papers from the dataset.

A bag-of-words model is a way of extracting features from text for use in modeling, such as with machine learning algorithms.



KNN Implementation

```
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)
```

KNN Result

Input:

```
test2=vectorizer.transform(["A novel Injection Locked Rotary Traveling Wave Oscillator"]).toarray()
```

Output:

```
Analysis of oscillator injection locking by harmonic balance method
Injection-Locked Clocking: A Low-Power Clock Distribution Scheme for High-Performance Microprocessors
Analysis of full-wave conductor system impedance over substrate using novel integration techniques
Predictable task migration for locked caches in multi-core systems
```

Word2Vec

- Popular machine learning model for generating word embeddings.
- It is a two layer neural network
- Word Embeddings- Mapping of words in a vector space
- Preserves relationship between words
- Deals with addition of new words in the vocabulary

Word2Vec Implementation

```
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
```

Word2Vec Result

Input:

```
word2vecInput='A novel Injection Locked Rotary Traveling Wave Oscillator'
```

Output:

```
[('Implementation of a directional beacon-based position location algorithm in a signal processing framework',
  0.5252764225006104),
 ('Data fusion algorithms for network anomaly detection: classification and evaluation',
  0.49821797013282776),
 ('Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies',
  0.4581085443496704),
 ('Comparative interactomics analysis of protein family interaction networks using PSIMAP (protein structural interac
tome map)',
  0.4025886356830597),
 ('Towards a real-time navigation strategy for a mobile robot',
  0.39102256298065186)]
```

Collaborative Filtering

- 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 exist 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.

major contributions

1. We utilized the advantages of publicly available contextual metadata to propose an independent research paper that does not require a priori user profile.
2. Our approach provides personalized recommendations regardless of the research field and regardless of user expertise.

Algorithm

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.

Pipelined Hybrid Model

To combine two approaches, content-based and collaborative filtering to get recommendation at the level of user interface

The cosine similarity is used to find out the research papers based on content (like., title and abstract) and this output is used for collaborative filtering using references of research papers.

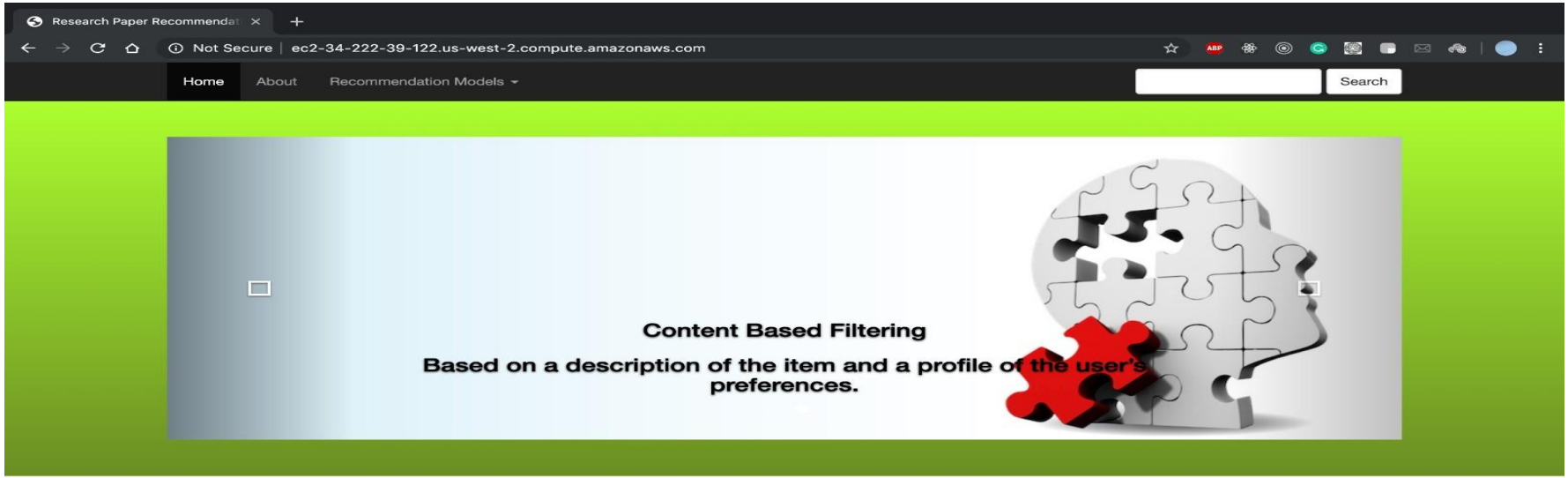
Web Application of Research paper Recommendation System

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

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

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

Web Application-home page



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

Research Paper Recommendat

Not Secure | ec2-34-222-39-122.us-west-2.compute.amazonaws.com

Home About Recommendation Models Search

Research Paper Recommendation

Hybrid and Multi model based research paper recommendation

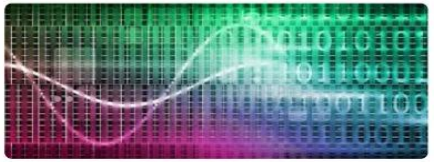
Researchers spend too much time and struggle to find the suitable article they are looking for. The problem becomes worse when a researcher with insufficient knowledge of searching research articles. How to formalize and solve the recommendation problem? In the traditional recommendation approaches, 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

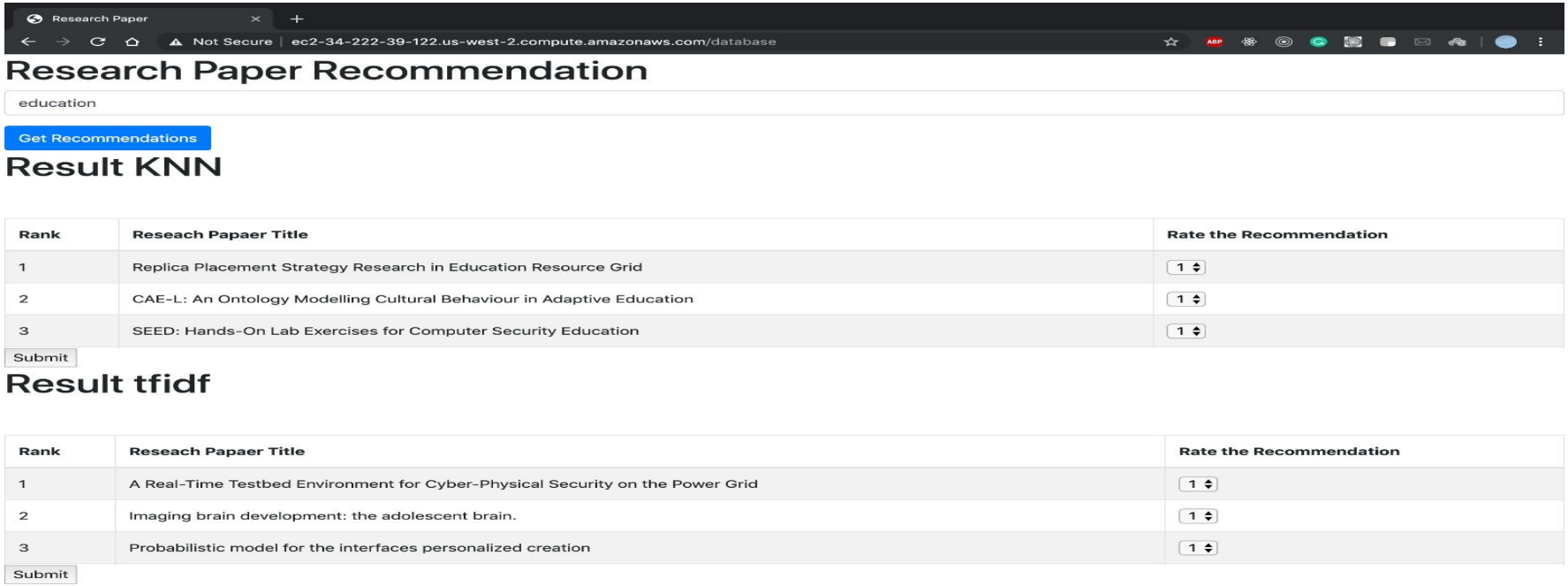
Word2vec is a group of related models that are used to produce word embeddings.

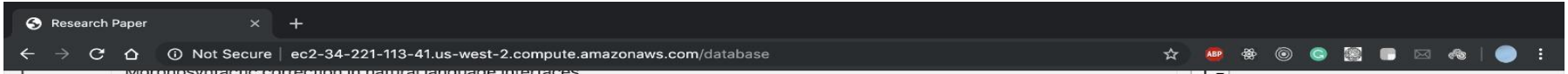


Item based Collaborative

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

Result of the KNNand TF-IDF Model





1	Morphosyntactic correction in natural language interfaces	1
2	Automatic construction of semantic lexicons for learning natural language interfaces	1
3	A controlled natural language layer for the semantic web	1

Submit

Result: Word-vector

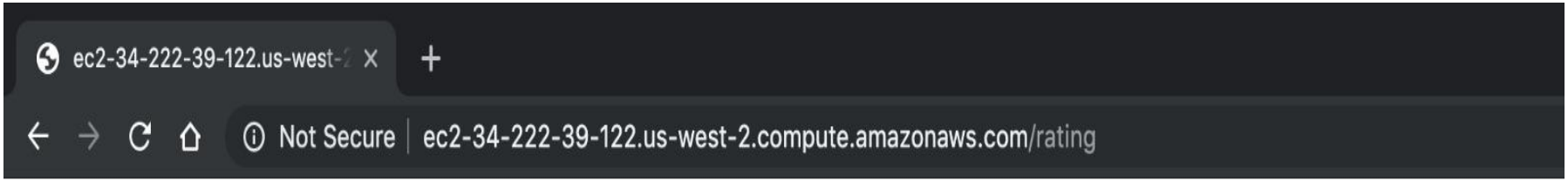
2	Cost-effective and low-power memory address bus encodings	1
3	ON-LINE: an architecture for modelling legal information	1
4	sGAL: a computational method for finding surface exposed sites in proteins suitable for Cys-mediated cross-linking	1
5	Development of a Novel	1

Submit

Result: Hybrid

Rank	Reseach Papaer Title	Rate the Recommendation
1	Mendeley readership counts: An investigation of temporal and disciplinary differences	1

Submit



Thank you.Your ratings are submitted successfully!

Home

EVALUATION

In the web application user can provide ratings to the each and every research paper he was referred to.

Users will give 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.

Conclusion

We have successfully implemented the 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 shows that hybrid based approach provides better recommendation.

Thank you