Software Design Document for project News Aggregator

Alaa Mohamed
Marwan Ibrahim
Mayar Yasser
Mohamed Ayman

Supervised by
Dr. Walaa Hassan
Eng. Mennatullah Gamil

March 2020

1 Introduction

1.1 Purpose of this document

The main purpose of software design document is to fully describe the architecture of our system. The system aims to aggregate relevant news from several ranked sites and platforms based on the user's input keywords or sentences and summarize them to reduce user's reading time. Furthermore, the document will present the details of the system's diagrams,functions and features. This Software Design Document (SDD) is needed for the graduation project at Misr International University (MIU).

1.2 Scope of this project

- The system will:
    1. Aggregate news and articles all in one place.
    2. Crawl the relevant articles when user input some keywords or sentences to read about it.
    3. Summarize the aggregated information from various sources in one page in order to reduce reading time.
    4. Recommend articles for the user according to his reading interests.

1.3 Overview

In the last few years, the world had an incredible and huge growth of the rate of News.This project focuses on gathering news from many sources with brief summarized output for the reader in order to reduce time reading full article. It also aims to aggregate news from many sources as possible in one place.

Our document have 8 sections as follows:

1. Introduction: is about introduction that explain the purpose and scope of the project
2. System Overview: is system overview that describe how the system work.
3. System Architecture: which architectural design will be used in our system.
4. Data Design: Explains our data design including the database.
5. Component Design
6. Human Interface Design: Describes the system from the user's perspective with screenshots.
7. Requirement Matrix: shows which components satisfy each of the functional requirements
8. References

1.4 Definitions and Acronyms

Term	Definition
Software Design Document (SDD)	Used as the primary medium for communicating software design information.
NLP	Natural language processing is a subfield of computer science,information engineering, and artificial intelligence concerned with the interactions between computers and human languages.
NLTK	Natural Language Toolkit is a platform used for building Python programs that work with human language data for applying in statistical natural language processing
Sequence Diagram	It is a diagram that shows object/actor interactions arranged in time sequence.

2 System Overview

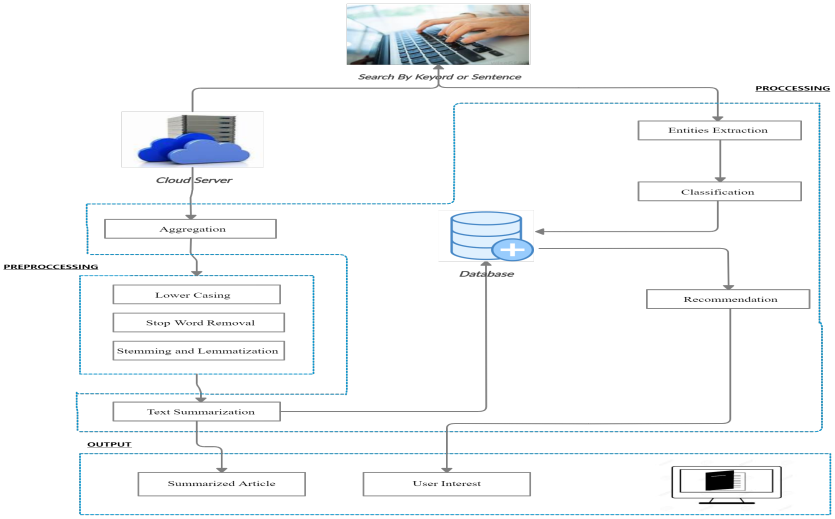


Figure 1: System overview

- When user enters a new keyword or some small sentences, it will face many stages in order such as:

  1. Aggregating relevant articles to this keyword or sentence, then it will go through the pre-processing stage which consists of:

     – Lowercase: used to reduce the size of the vocabulary in our data that cause multiple copies of the same word meaning.

     – Stop-word Removal: Done to remove small information of a text in order to focus on important words.

     – Lemmatization and Stemming: Remove inflection and map the word into the original/root form.

     After aggregation, all of this aggregated content will be taken to our algorithm to be summarized and then we will have the Output phase which contains the summarized article which is also saved in our database.

  2. Another stage that can be done to the keyword/ key phrases is:

     – Entities Extraction: Classifying key elements from a text into predefined categories.

     – Classification: Finding out which label is similar to the input keyword.

     – Recommendation: Showing the user top recommended topics according to his reading interests.

3. The final state is Output which is:
    - Summarized and well-classified news articles for the user to read.
    - Recommended Articles for the same topics that might interest this logged in user.

- Cloud Service: It contains a collection of ranked news articles.

- Database: The labeled articles are saved in it.

3 System Architecture

3.1 Architectural Design

For the Architectural design, functional and non-functional requirements were analyzed and it was shown that the system must insure the Model View Controller system architecture as shown below:

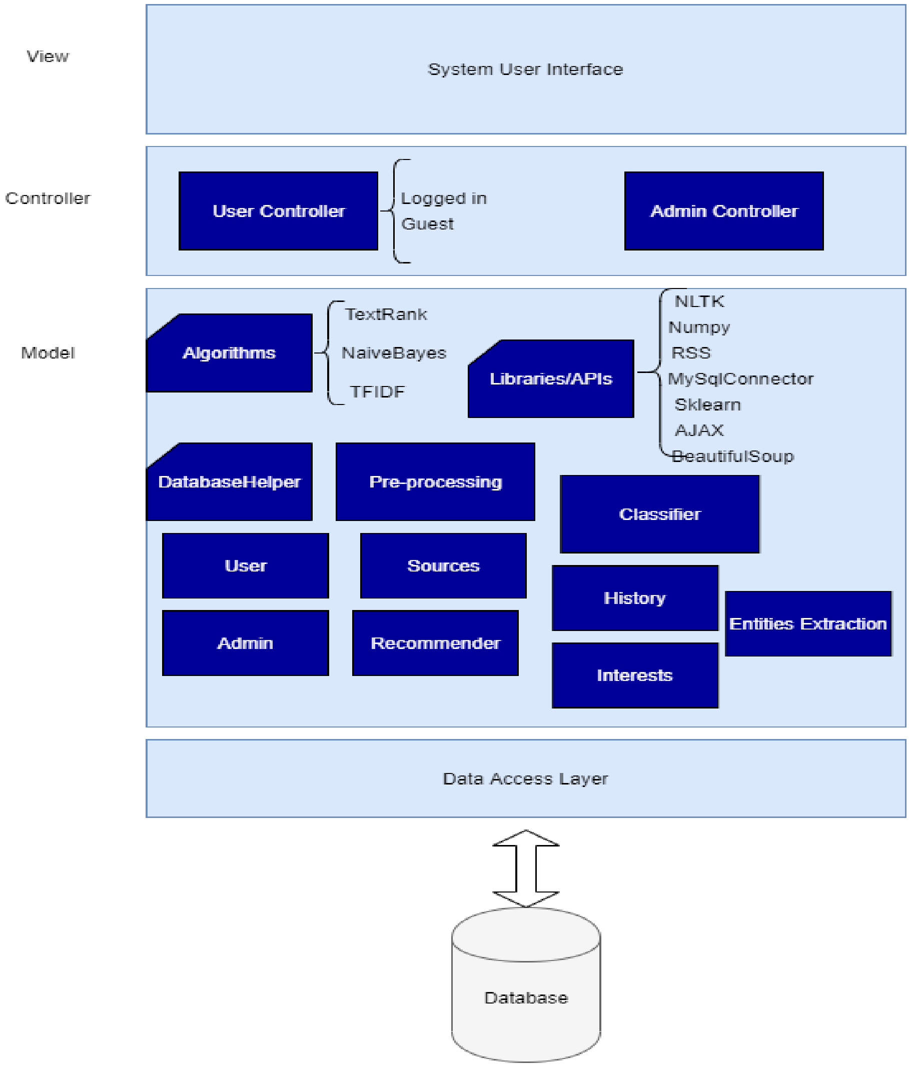


Figure 2: Architecture diagram

3.1.1 View

View is responsible for the representation of information and User Interface (UI).System has two different interfaces, one for admin with his operations and the other one is for the logged in user or guest in order to execute the core part which is aggregating and summarizing.

3.1.2 Controller

Controller is responsible for connecting or binding the model and view. The requests made within the view from the admin or user are taken and sent to database to retrieve data by used models and then data is passed to the view again to be shown. Examples of the controllers we have are: User controller that deals with how user interacts with the system, also the Admin controller deals with admin interactions and how they allow access to users.

3.1.3 Model

Model deals with simultaneous interpreters that will interact with the system either by executing the main and core functionalities or add and fetch data from the database. Moreover the controller is responsible for other functionalities like the input of text and the pre-processing done to those keywords.
Algorithm
TextRank: It starts with splitting the text into sentences.Then represent each word with a vector in a sentence using GloVe algorithm. Obtaining a similarity matrix for all sentences using cosine similarity. Converting it into a graph where the links determined by a similarity relation between them.Those links are used to obtain the vertices weight. ranking the weight un descending order to take the highest weight

3.2 Decomposition Description

3.2.1 Class Diagram

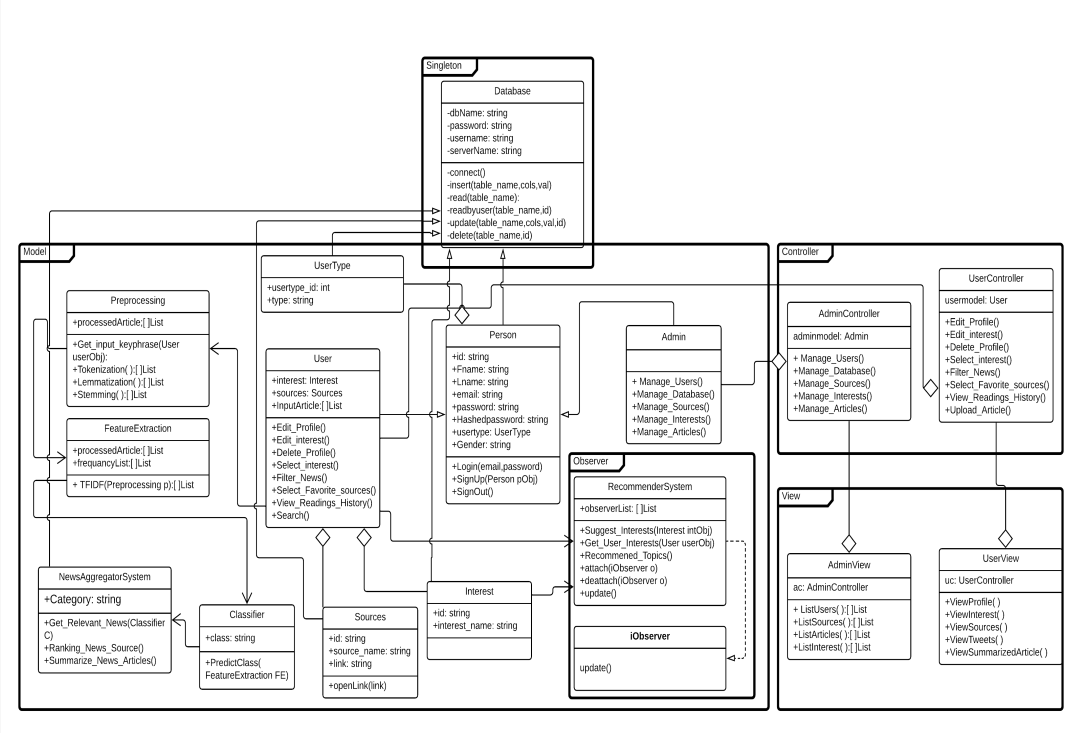


Figure 3: Class Diagram

3.2.2 Activity Diagram

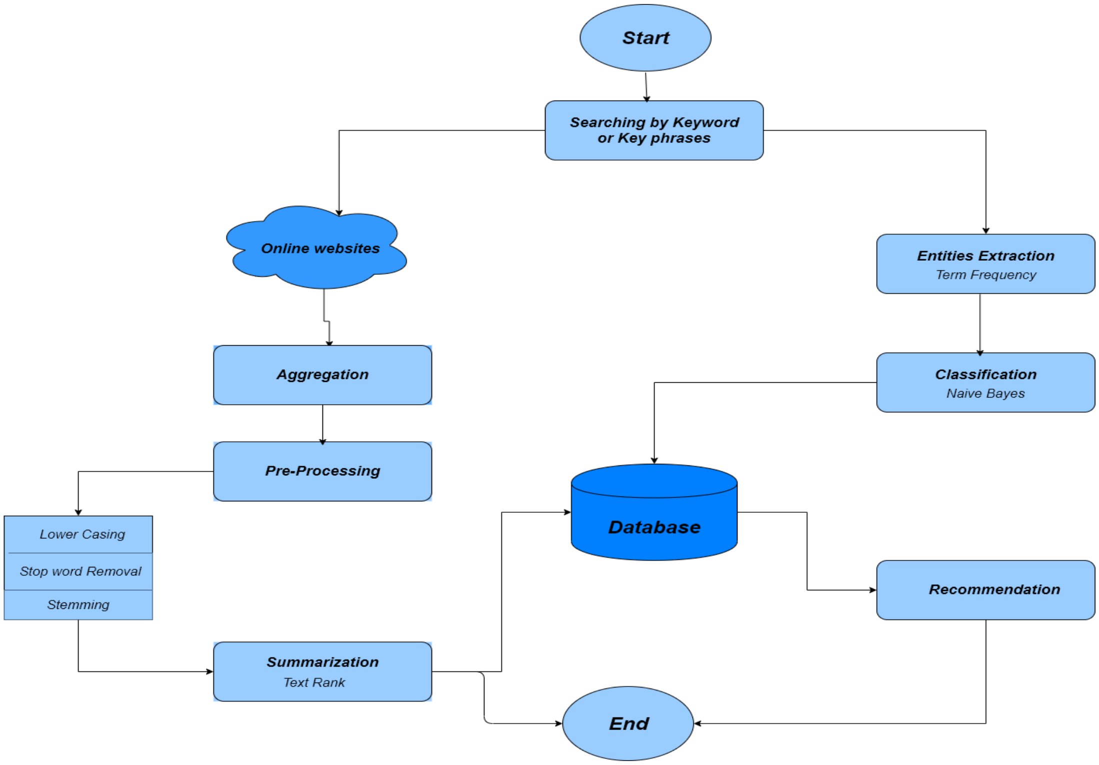


Figure 4: Activity Diagram

3.2.3 Sequence Diagram

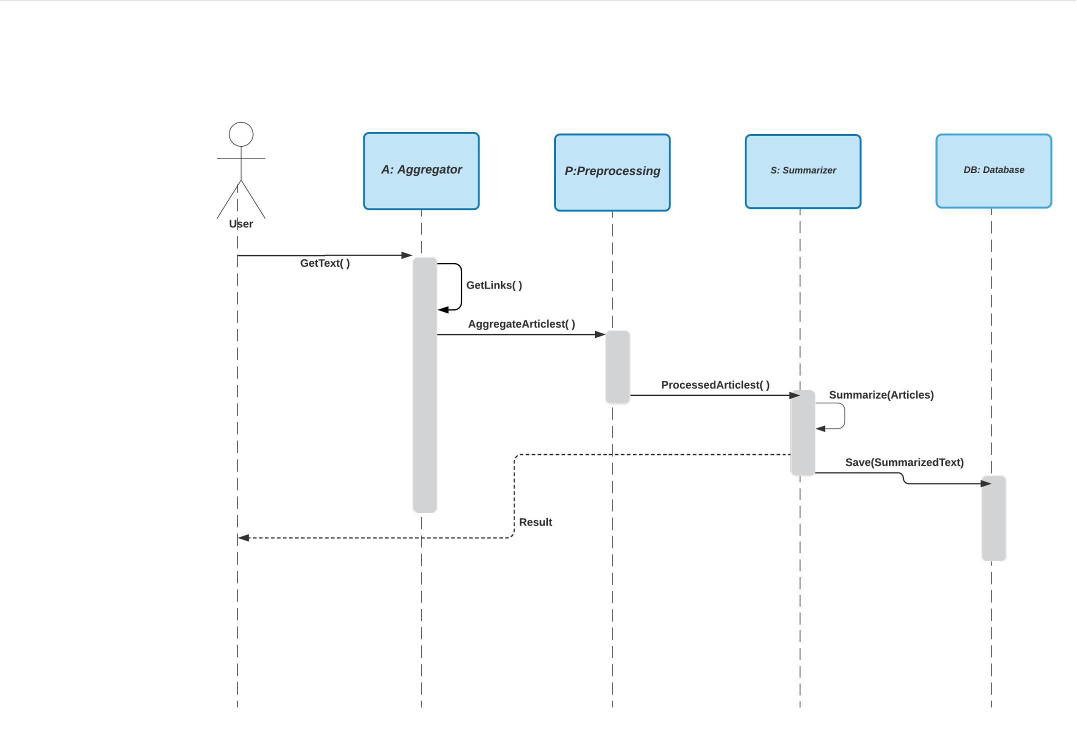


Figure 5: Summarization Sequence diagram

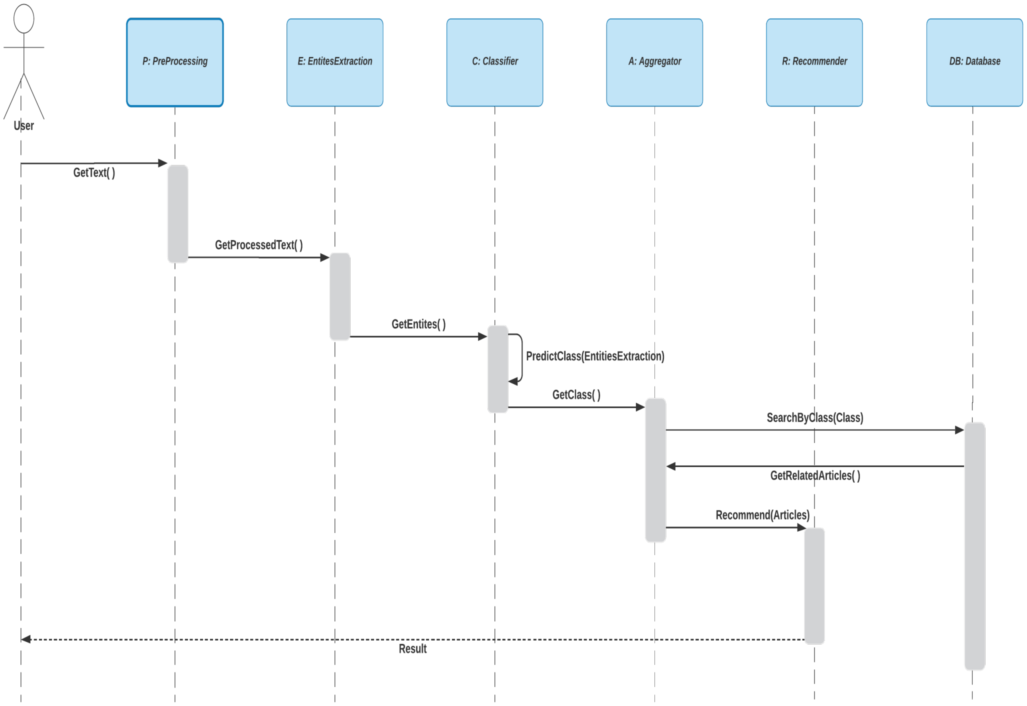


Figure 6: Recommendation Sequence diagram

3.2.4 Use Case Diagram

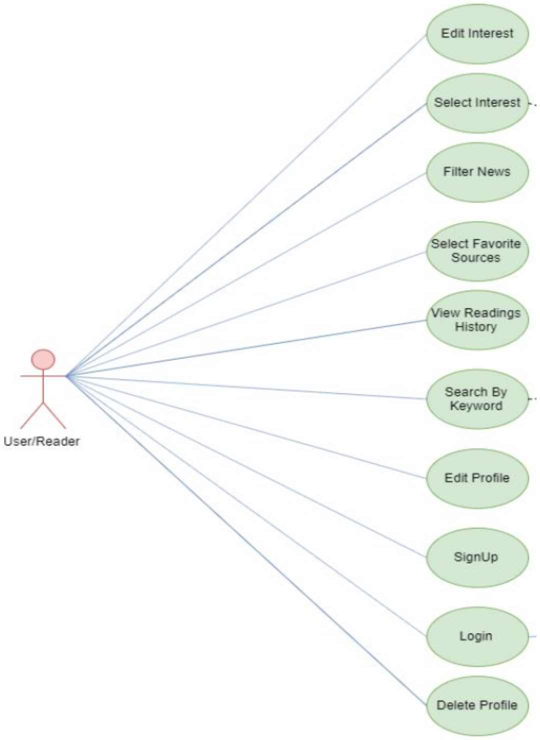


Figure 7: User Functions

The user which is defined as the reader can register an account and edit the profile or delete it.While registering he can select interests and select the favourite news sources.When the user starts to read, he can search by keyword in order to get the relevant articles from news sources.In addition, he can view his reading history or filter the news search result.

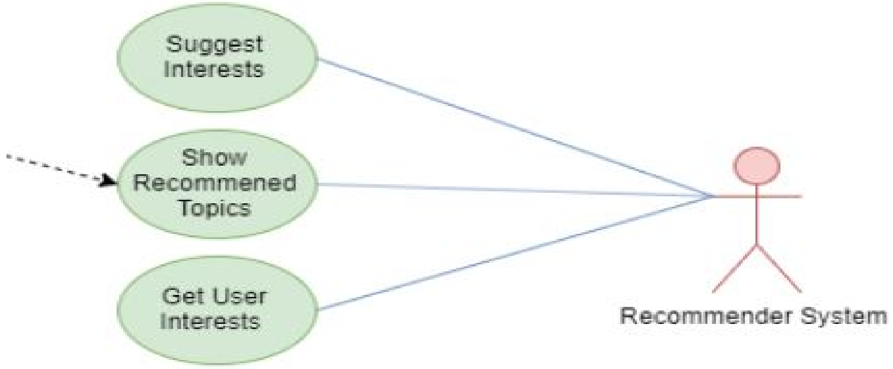


Figure 8: Recommender System Functions

According to the user, this system can show recommended topics upon user interests. Beside that it can also suggest the user some new interests regardless his chosen interests.

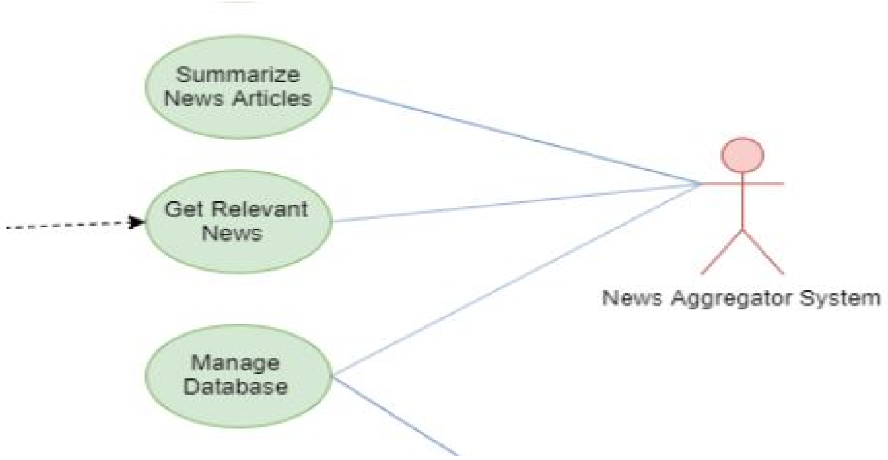


Figure 9: News Aggregator System Core Functions

This is our core system that its core functionality is getting relevant articles after getting some keywords from the user. It also summarize the output relevant articles to these keywords in order to save user's reading time.

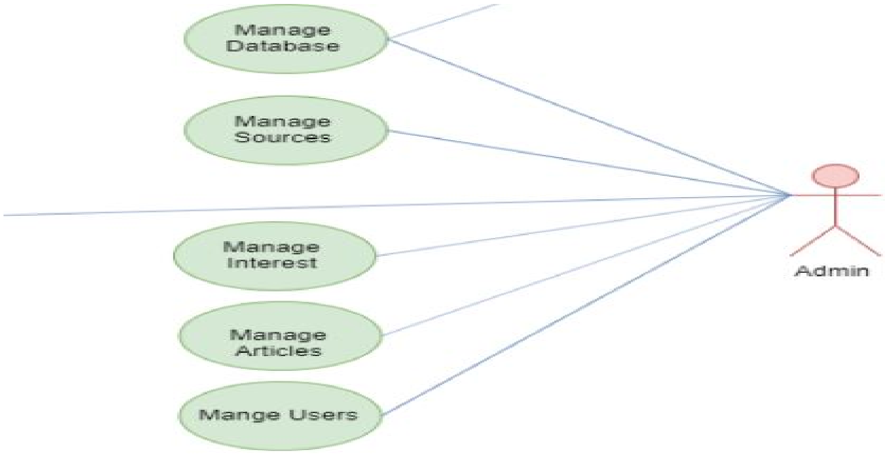


Figure 10: Admin Functions

The admin will be the one who can control the system where he can manage the sources,database,interests,sources and news articles. Admin can add, delete and modify the users, he can also modify and choose the news sources.Articles can be modified by the admin.

3.3 Design Rationale

As mention, previously, we have used Model View Controller (MVC) as architecture design for out system as it helped us to separate the functionality of the system and the actual User Interface (UI) of the system. So the modifications are more easier.

4 Data Design

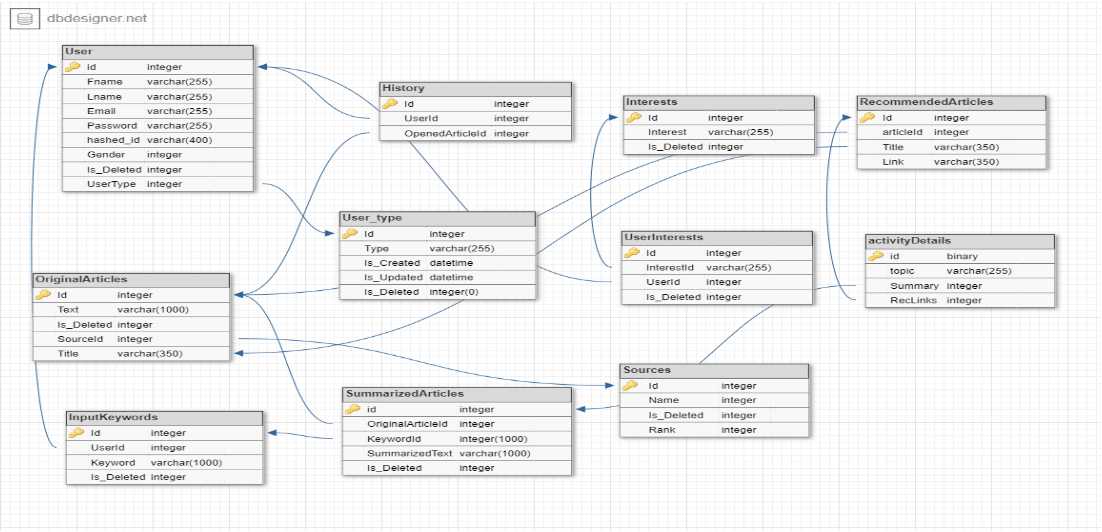


Figure 11: System ERD

4.1 Data Description

- User : This table stores user's information such as his Id, name, email, password.
- History: This table contains every user with his reading history.
- InputKeywords: This table contains every user with his input keywords to search for articles and summarized paragraphs about a specific topic.
- SummarizedArticles: This table contains summarized articles that have the highest rate in order to recommend them to other users.
- RecommendedArticles: This table contains the article that will be recommended to the user based on his interest.

5 Component Design

5.1 Pre-Processing

5.1.1 Stemming

Stemming is the algorithm of taking the word and return this word to its root.We used NLTK Library for stemming.

5.1.2 Lower Casing

Lower Casing is the process of making all of the words in the given article is in the same format which is lower cased.

5.1.3 Stop words removal

Stop words removal is the process of taking every sentence in the article or in the given text and remove all of the stop words from it like "this","the","and".This process was done by NLTK Library.

5.2 Feature Extraction

TFIDF was used in Feature Extraction which is short for term frequency inverse document frequency, is a numerical statistic that aims to reflect how important a word is to a document.

5.3 Articles Aggregation

5.3.1 Rich Site Summary (RSS)

For the guest home page, We are using Rich Site Summary (RSS) to aggregate updated articles from credible sources.

5.3.2 Beautiful Soup

Beautiful Soup is a python package that is used in our project for web scraping and parsing HTML and XML documents.

5.4 Text Classification

Text Classification was actually done by Naive Bayes approach.

5.5 Text Summarization

5.5.1 TextRank Algorithm

We are using TextRank algorithm [1] for articles summarization. TextRank is a graph based ranking algorithm that is used for summarization.The methodology of how this algorithm works in our project is showed in the shown figure:

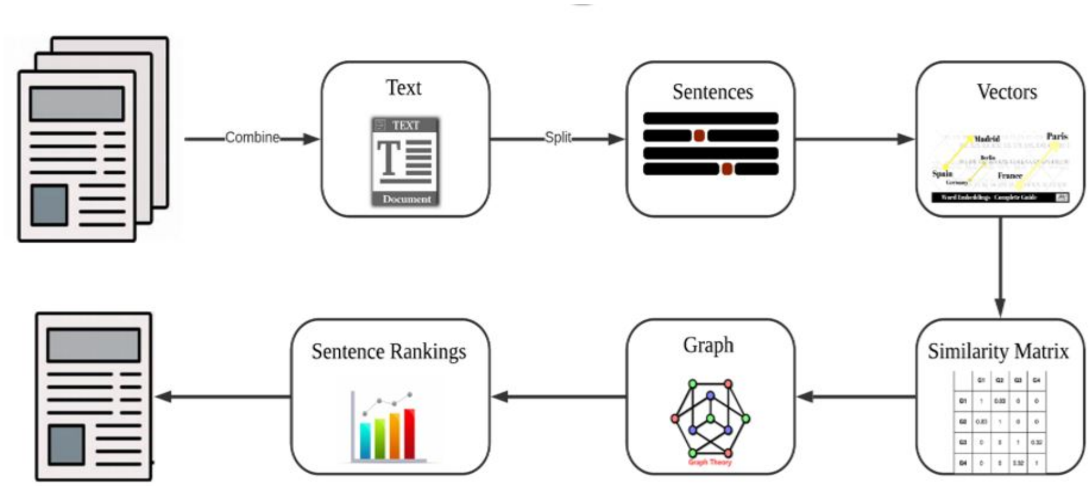


Figure 12: How TextRank Algorithm works.

There are six phases as shown in fig 11.

1. Choosing 3 articles from the database to combine as one original text as shown in figure 3.

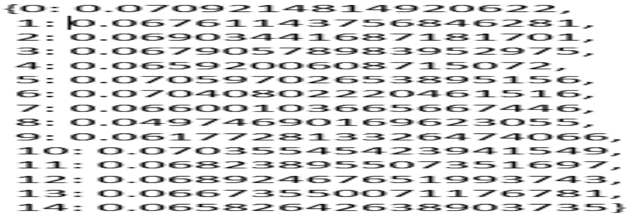
2. Tokenizing the original text as shown in figure 3 into sentences.

3. The Third phase is vectorization.In this phase w represent each word by a vector based on the co-occurrence of a word with the others in a single sentence using Global vector algorithm(GloVe) [2]. Then we represent each sentence by a vector calculated from the mean of words vectors in a sentence.

4. In this phase we obtain a similarity matrix for all sentences using cosine similarity [3].The similarity here refers to common content in sentences. Cos$\theta = \frac{a.b}{||a||*||b||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i}\sqrt{\sum_{i=1}^{n} b_i}}$
   where,a.b= $\sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$
   $isthedotproductofthetwovectors.$

5. After obtaining the similarity matrix in the previous phase.we convert it into a graph where the links determined by a similarity relation between them.Those links are used to obtain the vertices weight. The importance of a sentence is based on the number of links that represented as score for each vertex as shown in figure 6 using PageRank algorithm [4]. Let the directed graph, G=(V,E) where V represents set of vertices and E represents set of edges. The vertex score Vi is defined as follows: S(V$_i$) = $(1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$ where d is a damping factor that can be set between 0 and 1, which has the role of integrating into the model

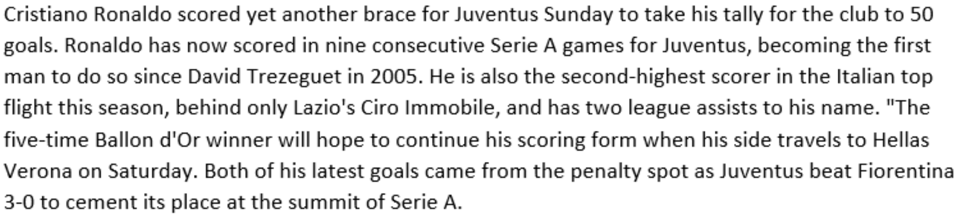
the probability of jumping from a given vertex to another random vertex in the graph



```
{0: 0.07092148814920622,
 1: 0.06761143756846281,
 2: 0.06903441687181701,
 3: 0.06790578983952975,
 4: 0.0659200608715072,
 5: 0.07059702653895156,
 6: 0.07040802220461516,
 7: 0.06600103665667446,
 8: 0.04974690169623055,
 9: 0.061772813326474066,
 10: 0.07035545423941549,
 11: 0.06823895507351697,
 12: 0.06892467651993743,
 13: 0.06673550071176781,
 14: 0.06582642638903735}
```

Figure 13: Sentence Scores

6. The final phase is ranking the scores shown in figure 6 in descending order. The highest scores create the final summary shown in figure 13.



Cristiano Ronaldo scored yet another brace for Juventus Sunday to take his tally for the club to 50 goals. Ronaldo has now scored in nine consecutive Serie A games for Juventus, becoming the first man to do so since David Trezeguet in 2005. He is also the second-highest scorer in the Italian top flight this season, behind only Lazio's Ciro Immobile, and has two league assists to his name. "The five-time Ballon d'Or winner will hope to continue his scoring form when his side travels to Hellas Verona on Saturday. Both of his latest goals came from the penalty spot as Juventus beat Fiorentina 3-0 to cement its place at the summit of Serie A.

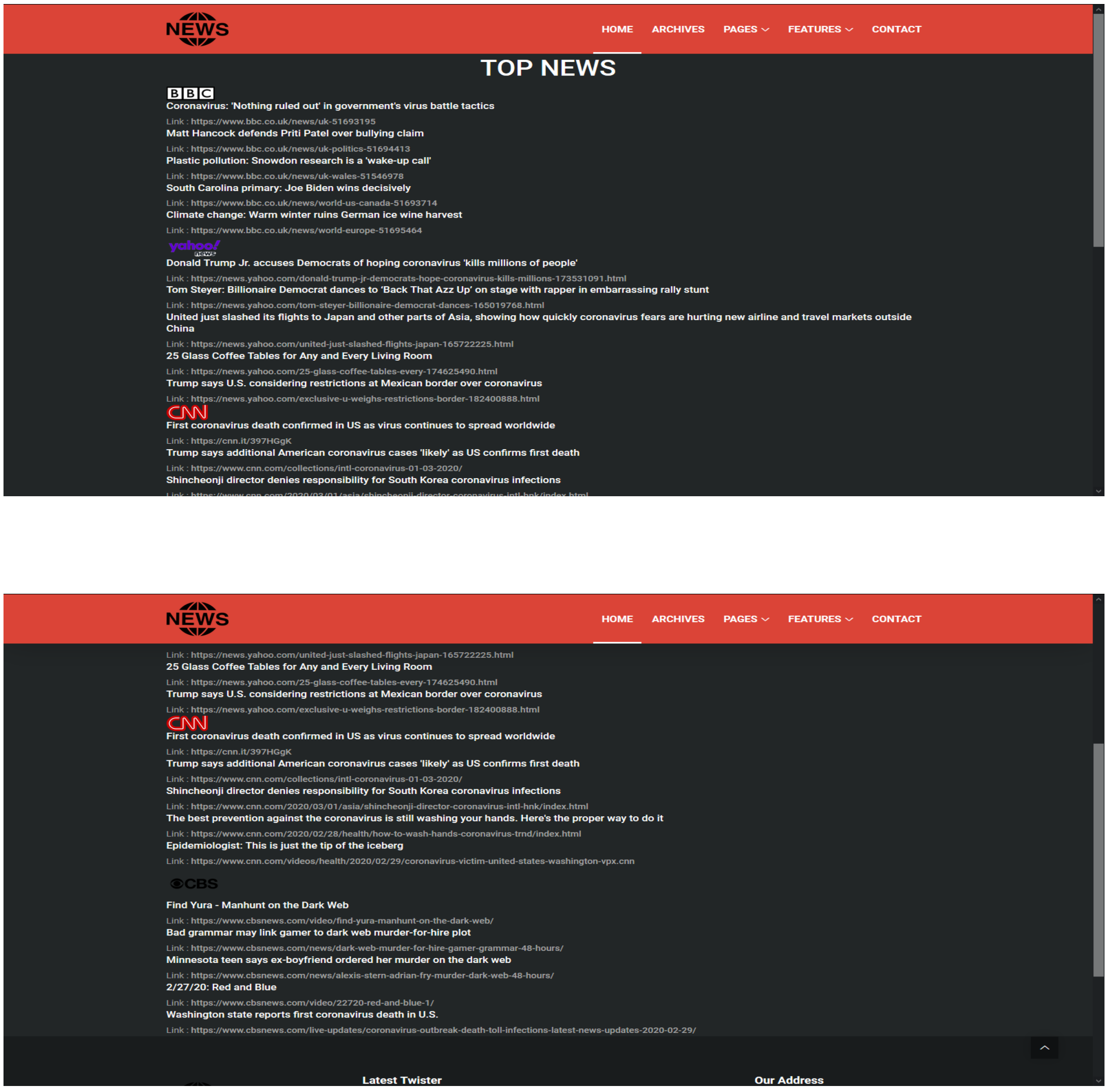
Figure 14: Summary using TextRank

6 Human Interface Design

6.1 Overview of User Interface

Describe the functionality of the system from the user's perspective. Explain how the user will be able to interact with the system. Information that will be displayed to the user is also involved.

6.2 Screen Images



TOP NEWS

BBC

Coronavirus: 'Nothing ruled out' in government's virus battle tactics

Link : https://www.bbc.co.uk/news/uk-51693195

Matt Hancock defends Priti Patel over bullying claim

Link : https://www.bbc.co.uk/news/uk-politics-51694413

Plastic pollution: Snowdon research is a 'wake-up call'

Link : https://www.bbc.co.uk/news/uk-wales-51546978

South Carolina primary: Joe Biden wins decisively

Link : https://www.bbc.co.uk/news/world-us-canada-51693714

Climate change: Warm winter ruins German ice wine harvest

Link : https://www.bbc.co.uk/news/world-europe-51695464

yahoo! news

Donald Trump Jr. accuses Democrats of hoping coronavirus 'kills millions of people'

Link : https://news.yahoo.com/donald-trump-jr-democrats-hope-coronavirus-kills-millions-173531091.html

Tom Steyer: Billionaire Democrat dances to 'Back That Azz Up' on stage with rapper in embarrassing rally stunt

Link : https://news.yahoo.com/tom-steyer-billionaire-democrat-dances-165019768.html

United just slashed its flights to Japan and other parts of Asia, showing how quickly coronavirus fears are hurting new airline and travel markets outside China

Link : https://news.yahoo.com/united-just-slashed-flights-japan-165722225.html

25 Glass Coffee Tables for Any and Every Living Room

Link : https://news.yahoo.com/25-glass-coffee-tables-every-174625490.html

Trump says U.S. considering restrictions at Mexican border over coronavirus

Link : https://news.yahoo.com/exclusive-u-weighs-restrictions-border-182400888.html

CNN

First coronavirus death confirmed in US as virus continues to spread worldwide

Link : https://cnn.it/397HGgK

Trump says additional American coronavirus cases 'likely' as US confirms first death

Link : https://www.cnn.com/collections/intl-coronavirus-01-03-2020/

Shincheonji director denies responsibility for South Korea coronavirus infections

Link : https://www.cnn.com/2020/03/01/asia/shincheonji-director-coronavirus-intl-hnk/index.html

NEWS
HOME ARCHIVES PAGES FEATURES CONTACT

Link : https://news.yahoo.com/united-just-slashed-flights-japan-165722225.html

25 Glass Coffee Tables for Any and Every Living Room

Link : https://news.yahoo.com/25-glass-coffee-tables-every-174625490.html

Trump says U.S. considering restrictions at Mexican border over coronavirus

Link : https://news.yahoo.com/exclusive-u-weighs-restrictions-border-182400888.html

CNN

First coronavirus death confirmed in US as virus continues to spread worldwide

Link : https://cnn.it/397HGgK

Trump says additional American coronavirus cases 'likely' as US confirms first death

Link : https://www.cnn.com/collections/intl-coronavirus-01-03-2020/

Shincheonji director denies responsibility for South Korea coronavirus infections

Link : https://www.cnn.com/2020/03/01/asia/shincheonji-director-coronavirus-intl-hnk/index.html

The best prevention against the coronavirus is still washing your hands. Here's the proper way to do it

Link : https://www.cnn.com/2020/02/28/health/how-to-wash-hands-coronavirus-trnd/index.html

Epidemiologist: This is just the tip of the iceberg

Link : https://www.cnn.com/videos/health/2020/02/29/coronavirus-victim-united-states-washington-vpx.cnn

CBS

Find Yura - Manhunt on the Dark Web

Link : https://www.cbsnews.com/video/find-yura-manhunt-on-the-dark-web/

Bad grammar may link gamer to dark web murder-for-hire plot

Link : https://www.cbsnews.com/news/dark-web-murder-for-hire-gamer-grammar-48-hours/

Minnesota teen says ex-boyfriend ordered her murder on the dark web

Link : https://www.cbsnews.com/news/alexis-stern-adrian-fry-murder-dark-web-48-hours/

2/27/20: Red and Blue

Link : https://www.cbsnews.com/video/22720-red-and-blue-1/

Washington state reports first coronavirus death in U.S.

Link : https://www.cbsnews.com/live-updates/coronavirus-outbreak-death-toll-infections-latest-news-updates-2020-02-29/

Latest Twister

Our Address

Figure 15: Guest Home Screen

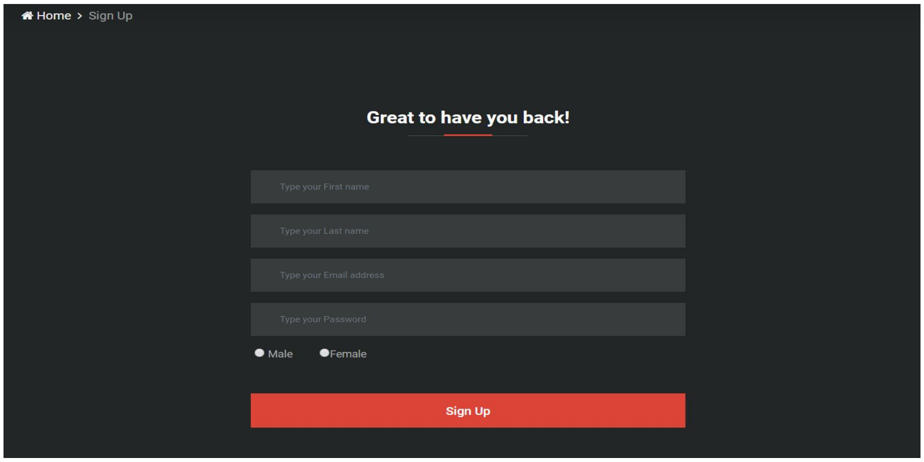


Figure 16: Sign Up Screen

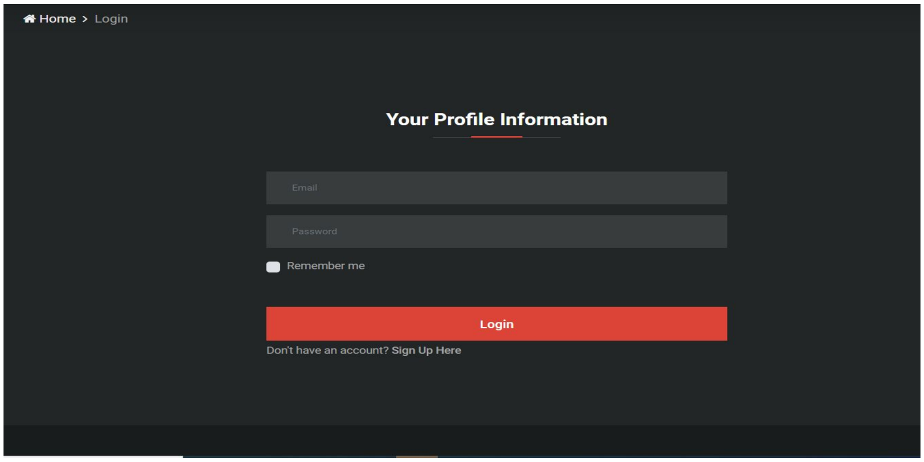
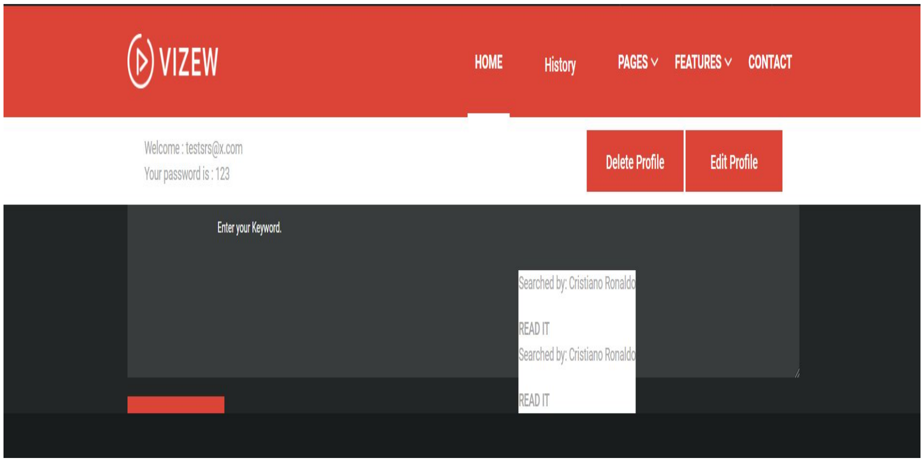


Figure 17: Login Screen



Welcome : testsrs@x.com
Your password is : 123

Enter your Keyword.

Searched by: Cristiano Ronaldo

READ IT

Searched by: Cristiano Ronaldo

READ IT

Figure 18: User History Screen

7 Requirements Matrix

Requirement id	Requirement type	Description	Status
1.Login	Required	User enters his username and password to be checked in the database.	Completed
2.Sign Up	Required	User enters his information to the system.	Completed
3.Input text	Required	User enters keywords to get relevant articles to them.	Completed
4.Aggregate News	Required	News are aggregated from more than one credible source to our website.	Completed
5.Summarize text	Required	Relevant articles to user's keywords are summarized to save user's reading time.	Completed
6.Recommend articles	Required	Recommend some other articles to the logged in user to read based on his interests or what he always reads.	In Progress
7.Get History	Required	Save user's reading history in order to be able to recommend him/her articles that match their interests.	In Progress

References

[1] Bartomiej Balcerzak, Wojciech Jaworski, and Adam Wierzbicki. Application of textrank algorithm for credibility assessment. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 451–454. IEEE, 2014.

[2] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference*

on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[3] Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6. IEEE, 2016.

[4] Peng Chen, Huafeng Xie, Sergei Maslov, and Sidney Redner. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.