

**CS 6359 Section 001**  
**Object-Oriented Analysis & Design**  
**Spring 2022**

**PROJECT-1 REPORT**

***TITLE: SMART SEARCH ENGINE***

**TEAM MEMBERS:**

YELISETTI KRISHNA TEJA (NET-ID: KXY200016)  
SATYA SOMEPALLI (NET-ID: SXS190436)  
SANJANA PENMETSU (NET-ID: SXP190149)  
PREETHAM RAO GOTTUMUKULA (NET-ID: PXG210001)  
HARSHITH RAVIPROLU (NET-ID: HXR180005)  
SAKETH DASAVATHINI (NET-ID: SXD190016)  
B MOUNIKA (NET-ID: MXB210007)  
THOTA JAYASHREE SANTHOSHI (NET-ID: JXT210011)  
SAI PRANAV REDDY DONTNIDI (NET-ID: SXD200125)  
APUROOP PARAVADA (NET-ID: AXP210033)  
YOGESH BALU (NET-ID: YXB200007)  
SRINATH REDDY MAVILLAPALLY (NET-ID: SXM210047)  
KIRAN DEVARAJ RAJ (NET-ID: KXR190038)

## **FUNCTIONAL REQUIREMENTS:**

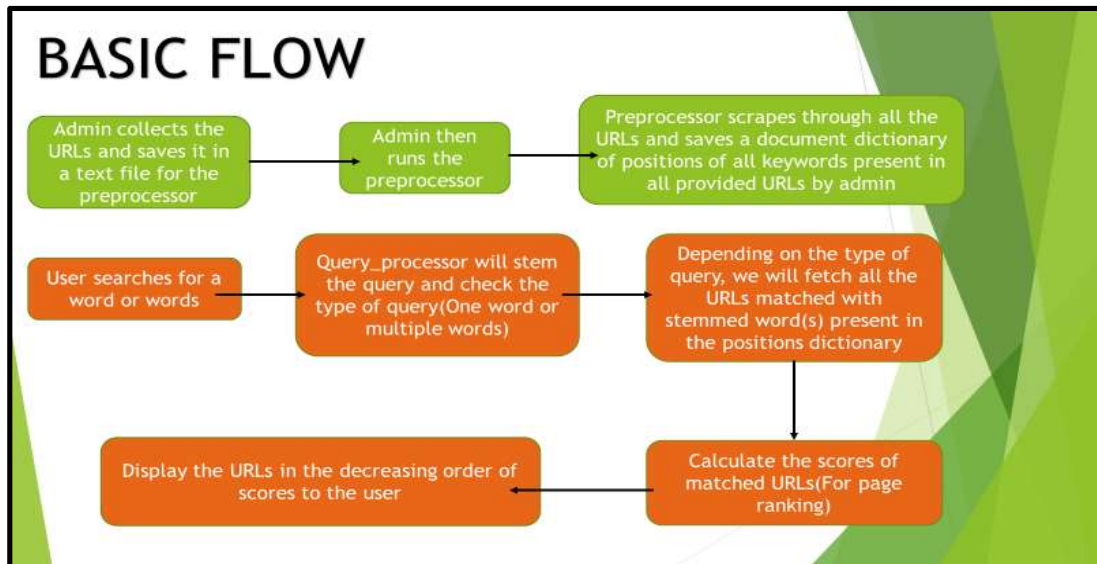
- **Logical Operators (AND & OR) Search:**
  - Implemented OR logically search for multi phrase query.
  - Will Implement AND logically search for multi phrase query in the next phase of the project.
- **Domain-Specific Support:**
  - Implemented search engine on movie database.
- **Ranking/Ordering URLs:**
  - After query search, results are returned based on scores of TF-IDF in decreasing order.
- **Filter Stop Words:**
  - Filtered Stop Words from the query for accurate results.
- **Locally Hosted:**
  - Hosted Search Engine locally using Flask Application.
- **AutoFill Query Support:**
  - In the next phase of the project, we will Implement AutoFill operation at search box for input query.
- **URLs Addition:**
  - Will add more URLs and expand the movie database in the next phase of the project.
- **Pagination:**
  - Will implement the number of results returned per page functionality in the next phase of the project.
- **Removal of Outdated URLs:**
  - In the next phase of the project, we are planning to deal with any outdated URLs present in the database.
- **Page Not Found (404):**
  - For the next phase of the project, we will be implementing and displaying a special page for those results which cannot be found according to our movie database.

## **NON- FUNCTIONAL REQUIREMENTS:**

- **Friendly and Readable User Interface:**
  - As of now, we are using the Flask application to display our basic output. In the next phase of the project, we are planning to implement a better User Interface
- **Portability and usability:**
  - Code Implementation is made Understandable and Portable as an app for any further development.
- **Localization:**
  - We are maintaining a requirements text file that will let us run our application on any machine.
- **Low Latency & High Throughput:**
  - Our search engine is efficient because we are using pre-processed dictionaries to retrieve the results for the searched query quickly.

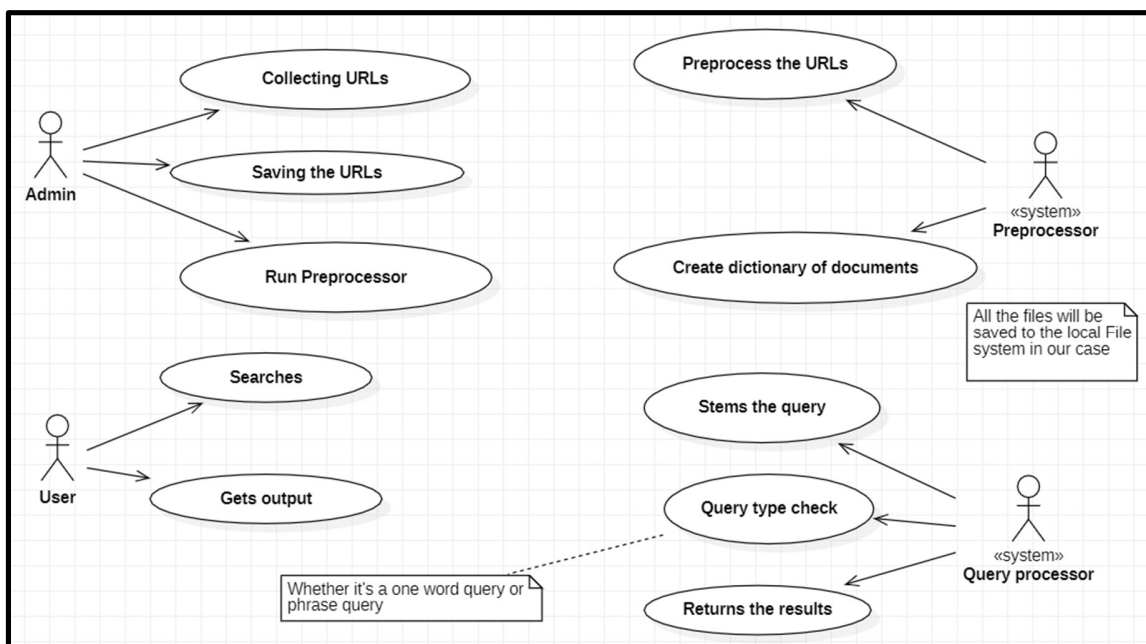
## DIAGRAMS:

To understand the UML diagrams, a basic flow of the project is attached below for reference:

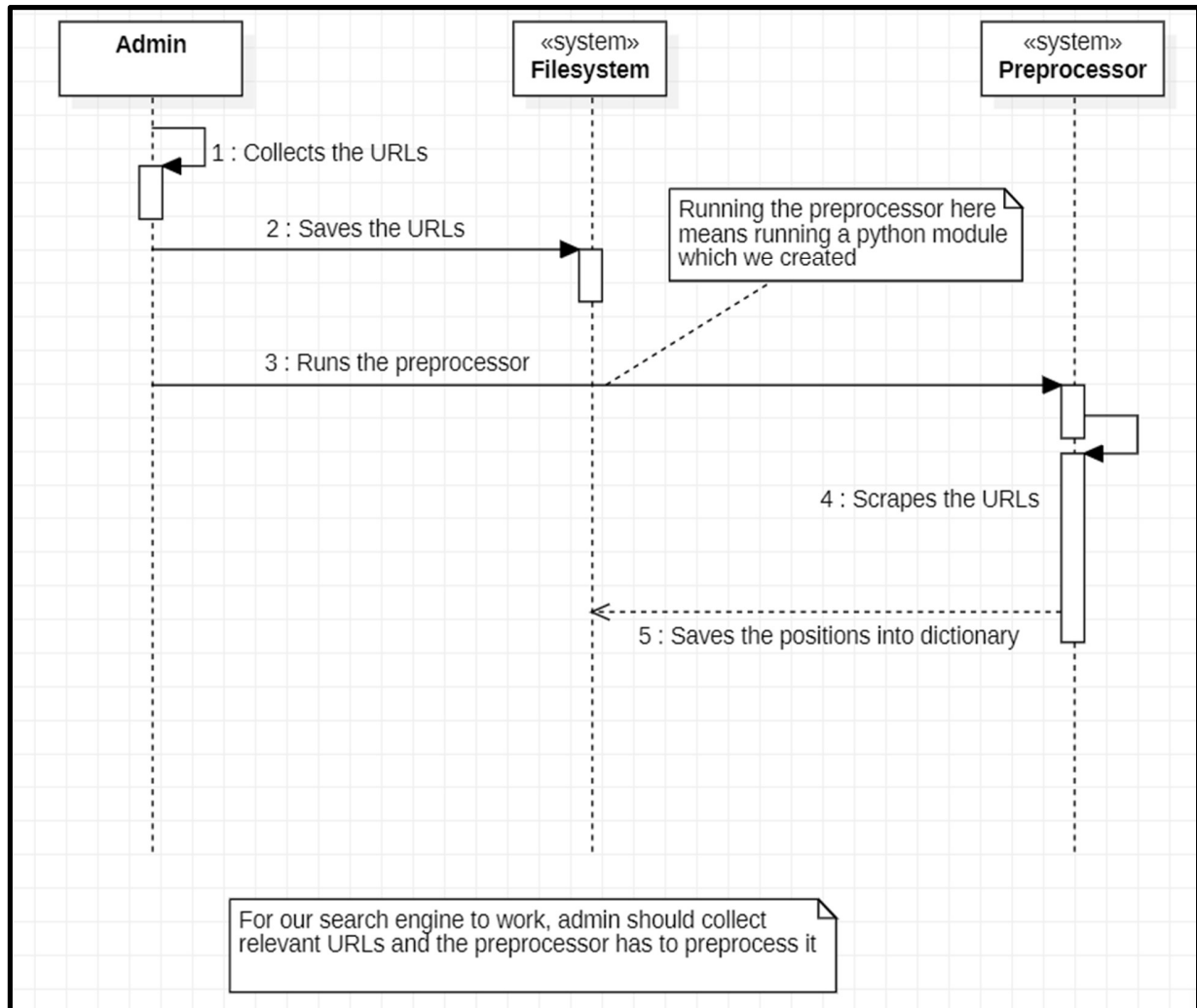


## UML DIAGRAMS:

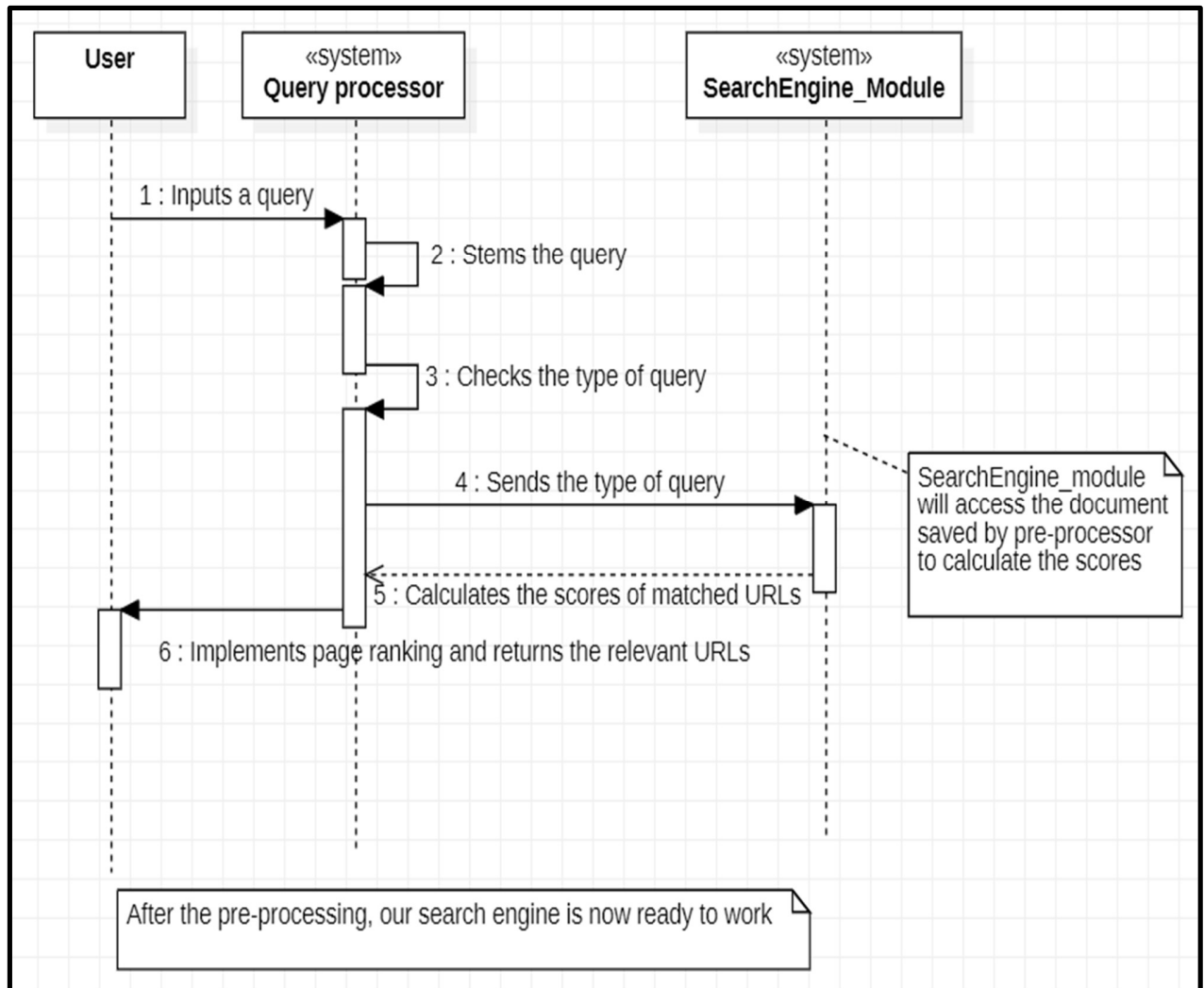
Following is our use case diagram where external actors are the User and the Admin and internal actors (inside the system) are Preprocessor and Query processor.



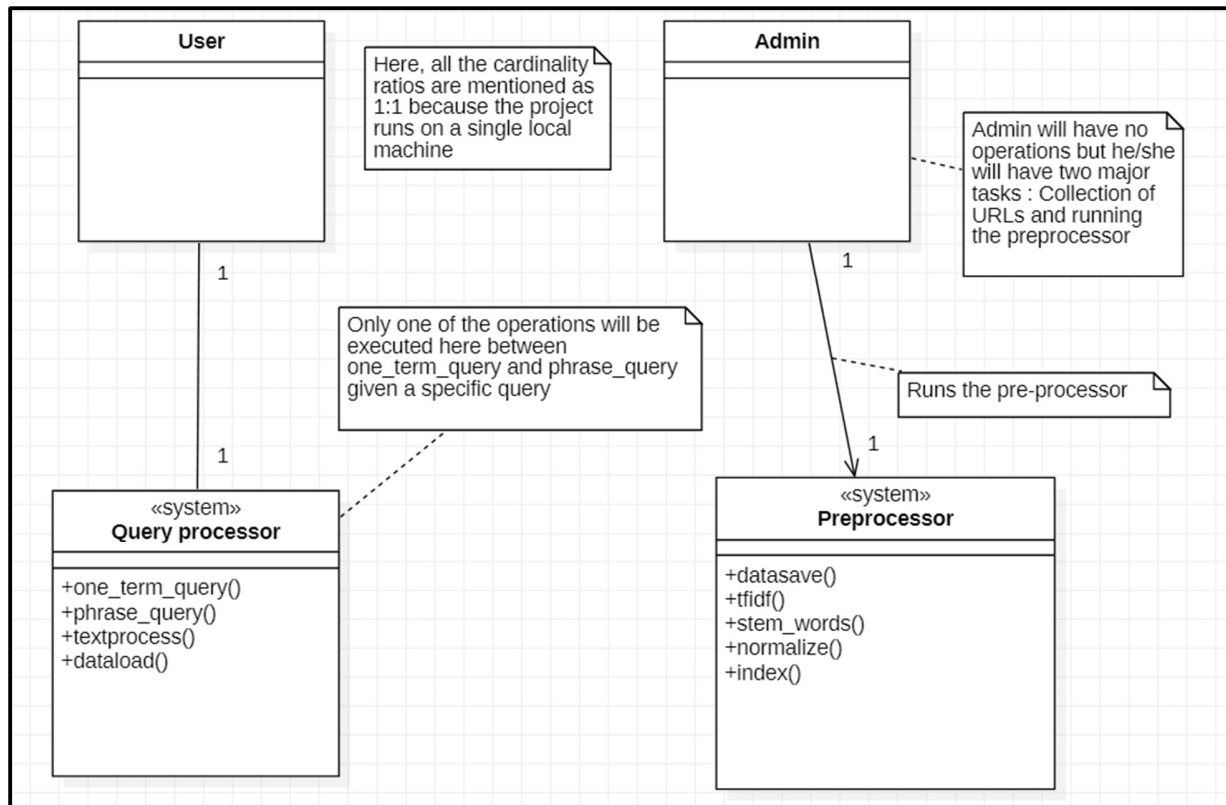
Following is our first sequence diagram where we can see a time-based sequence of events happening between Admin and Preprocessor. Here, FileSystem refers to our local machine where we save the URLs collected by the admin and dictionaries created by Preprocessor.



Following is our second sequence diagram (*main sequence diagram*) where we can see a time-based sequence of events happening between the User and Query-processor.



Following is our basic class diagram where we are primarily showing that admin and pre-processor are independent of User and Query-processor. We also wrote a few basic and important operations happening in Query-processor and Pre-processor.



*Note: All the above UML diagrams will be updated accordingly as we might plan to change a few functionalities to improve our project for the final phase.*

## MOCK-UP SCREENSHOTS:

Admin will collect all the relevant URLs for the pre-processor and save it.  
Following is a screenshot of how the file looks :

Genre	Wiki Page	Plot
unknown	<a href="https://en.wikipedia.org/wiki/Kansas_Saloon_Smashers">https://en.wikipedia.org/wiki/Kansas_Saloon_Smashers</a>	A bartender is working at a saloon, se
unknown	<a href="https://en.wikipedia.org/wiki/Love_by_the_Light_of_the_Moon">https://en.wikipedia.org/wiki/Love_by_the_Light_of_the_Moon</a>	The moon, painted with a smiling face
unknown	<a href="https://en.wikipedia.org/wiki/The_Martyred_Presidents">https://en.wikipedia.org/wiki/The_Martyred_Presidents</a>	The film, just over a minute long, is co
unknown	<a href="https://en.wikipedia.org/wiki/Terrible_Teddy_the_Grizzly_King">https://en.wikipedia.org/wiki/Terrible_Teddy_the_Grizzly_King</a>	Lasting just 61 seconds and consisti
unknown	<a href="https://en.wikipedia.org/wiki/Jack_and_the_Beanstalk_(1902_film)">https://en.wikipedia.org/wiki/Jack_and_the_Beanstalk_(1902_film)</a>	The earliest known adaptation of the
unknown	<a href="https://en.wikipedia.org/wiki/Alice_in_Wonderland_(1903_film)">https://en.wikipedia.org/wiki/Alice_in_Wonderland_(1903_film)</a>	Alice follows a large white rabbit dow
western	<a href="https://en.wikipedia.org/wiki/The_Great_Train_Robbery_(1903_film)">https://en.wikipedia.org/wiki/The_Great_Train_Robbery_(1903_film)</a>	She enters a kitchen, in which there is
comedy	<a href="https://en.wikipedia.org/wiki/The_Suburbanite">https://en.wikipedia.org/wiki/The_Suburbanite</a>	The film opens with two bandits head
unknown	<a href="https://en.wikipedia.org/wiki/The_Little_Train_Robbery">https://en.wikipedia.org/wiki/The_Little_Train_Robbery</a>	Men
unknown	<a href="https://en.wikipedia.org/wiki/The_Night_Before_Christmas_(1905_film)">https://en.wikipedia.org/wiki/The_Night_Before_Christmas_(1905_film)</a>	The re
short	<a href="https://en.wikipedia.org/wiki/Dream_of_a_Rabbit_Friend_(1906_film)">https://en.wikipedia.org/wiki/Dream_of_a_Rabbit_Friend_(1906_film)</a>	Scenes are introduced using titles or
short action/crime western	<a href="https://en.wikipedia.org/wiki/From_Leadville_to_Aspen:_A_Hold-Up_in_the_Rockies">https://en.wikipedia.org/wiki/From_Leadville_to_Aspen:_A_Hold-Up_in_the_Rockies</a>	The Rabbit Friend gorges on Welsh ra
short film	<a href="https://en.wikipedia.org/wiki/Kathleen_Mavourneen_(1906_film)">https://en.wikipedia.org/wiki/Kathleen_Mavourneen_(1906_film)</a>	The film features a train traveling thro
biographical	<a href="https://en.wikipedia.org/wiki/Daniel_Boone_(1907_film)">https://en.wikipedia.org/wiki/Daniel_Boone_(1907_film)</a>	Irish villager Kathleen is a tenant of C
comedy	<a href="https://en.wikipedia.org/wiki/How_Brown_Saw_the_Baseball_Game">https://en.wikipedia.org/wiki/How_Brown_Saw_the_Baseball_Game</a>	Boone's daughter befriends an Indian
comedy	<a href="https://en.wikipedia.org/wiki/Laughing_Gas_film#1907_Film">https://en.wikipedia.org/wiki/Laughing_Gas_film#1907_Film</a>	Before heading out to a baseball gam
drama	<a href="https://en.wikipedia.org/wiki/The_Adventures_of_Dollie">https://en.wikipedia.org/wiki/The_Adventures_of_Dollie</a>	The plot is that of a black woman goi
drama	<a href="https://en.wikipedia.org/wiki/The_Black_Viper">https://en.wikipedia.org/wiki/The_Black_Viper</a>	On a beautiful summer day a father a
		A thug accosts a girl as she leaves he

Pre-processor will  
scrape these URLs

Admin will run the pre-processor and pre-processor will scrape all the URLs  
and give us an output file which looks like follows:

```
(l) positions.json > ...
{"scandal": {"https://en.wikipedia.org/wiki/Scandal_Sheet_(1952_film)": [0, 331, 403], "https://en.wikipedia.org/wiki/Simon_Birch": [150]}, "sheet": {"https://en.wikipedia.org/wiki/Scandal_Sheet_(1952_film)": [1, 332, 404], "https://en.wikipedia.org/wiki/Whiplash_(2014_film)": [255, 263, 500]}, "american": {"https://en.wikipedia.org/wiki/Scandal_Sheet_(1952_film)": [2], "https://en.wikipedia.org/wiki/Three_Stripes_in_the_Sun": [3, 17], "https://en.wikipedia.org/wiki/The_Candy_Snatchers": [2], "https://en.wikipedia.org/wiki/Run_(1991_film)": [1], "https://en.wikipedia.org/wiki/Hemingway%27s_Adventures_of_a_Young_Man": [4], "https://en.wikipedia.org/wiki/Bungalow_13": [1], "https://en.wikipedia.org/wiki/Nine_(2009_live-action_film)": [164], "https://en.wikipedia.org/wiki/Le_Divorce": [109], "https://en.wikipedia.org/wiki/Higher_Than_a_Kite": [8], "https://en.wikipedia.org/wiki/Desert_Sands": [2], "https://en.wikipedia.org/wiki/The_Road_(2009_film)": [1, 34], "https://en.wikipedia.org/wiki/Toy_Story_3": [2, 2349], "https://en.wikipedia.org/wiki/Monty_Python%27s_The_Meaning_of_Life": [296, 352, 861, 1337], "https://en.wikipedia.org/wiki/Shark!_(film)": [7], "https://en.wikipedia.org/wiki/Avunu_(film)": [39], "https://en.wikipedia.org/wiki/They_Call_It_Sin": [2], "https://en.wikipedia.org/wiki/Hell_Baby": [2], "https://en.wikipedia.org/wiki/The_Incident_(1967_film)": [1, 175], "https://en.wikipedia.org/wiki/Simon_Birch": [2, 641], "https://en.wikipedia.org/wiki/21_(2008_film)": [0, 590, 603, 615, 635, 657, 679, 685, 702, 729], "https://en.wikipedia.org/wiki/The_Frozen_Ground": [2], "https://en.wikipedia.org/wiki/Billy_Jack": [2, 48, 214, 544, 625, 662], "https://en.wikipedia.org/wiki/RoboCop_3": [1], "https://en.wikipedia.org/wiki/Guilty_as_Sin": [2], "https://en.wikipedia.org/wiki/The_Long_Night_(1947_film)": [2], "https://en.wikipedia.org/wiki/Fighting_(2009_film)": [1, 306], "https://en.wikipedia.org/wiki/The_Killing_Fields_(film)": [17, 186, 424, 505, 1025], "https://en.wikipedia.org/wiki/The_True_Story_of_Eskimo_Nell": [492], "https://en.wikipedia.org/wiki/Jesse_James_vs._the_Daltons": [4], "https://en.wikipedia.org/wiki/Anthony_Adverse": [2], "https://en.wikipedia.org/wiki/The_Boondock_Saints_II:_All_Saints_Day": [4], "https://en.wikipedia.org/wiki/The_Kid_from_Kokomo": [2], "https://en.wikipedia.org/wiki/Whiplash_(2014_film)": [1, 101, 1194], "https://en.wikipedia.org/wiki/Rosewood_(film)": [1, 259, 367, 493, 668], "https://en.wikipedia.org/wiki/Kalangarai_Vilakkam": [20], "https://en.wikipedia.org/wiki/Blade_(film)": [1], "https://en.wikipedia.org/wiki/Driven_(2001_film)": [1], "https://en.wikipedia.org/wiki/Loose_Cannons_(1990_film)": [2], "https://en.wikipedia.org/wiki/Moscow_on_the_Hudson": [2, 46, 263, 301], "https://en.wikipedia.org/wiki/Lost_in_Translation_(film)": [17, 30, 262, 285, 1622, 1794, 1840, 2494, 3343, 3544], "https://en.wikipedia.org/wiki/Deadly_Eyes": [438], "https://en.wikipedia.org/wiki/Hot_Young_Bloods": [283], "https://en.wikipedia.org/wiki/Tarnished": [1], "https://en.wikipedia.org/wiki/Come_Spy_with_Me_(film)": [2, 106], "https://en.
```



Following is the screenshot of document which will have TF-IDF scores of terms scraped in respective URLs in the form {TF, IDF, TF-IDF Scores} :

```

"sheet": {"https://en.wikipedia.org/wiki/Scandal_Sheet_(1952_film)": [3, 1.6989/00043360185, 3.5654793291780806], "https://en.wikipedia.org/wiki/Whiplash_(2014_film)": [3, 1.69897000433660185, 3.5654793291780806]}, {"american": {"https://en.wikipedia.org/wiki/Scandal_Sheet_(1952_film)": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Three_Stripes_in_the_Sun": [2, 0.3098039199714863, 0.524543633626141], "https://en.wikipedia.org/wiki/The_Candy_Snatchers": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Run_(1991_film)": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Hemingway%27s_Adventures_of_a_Young_Man": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Bungalow_13": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Nine_(2009_live-action_film)": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Le_Divorce": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Higher_Than_a_Kite": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Desert_Sands": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/The_Road_(2009_film)": [2, 0.3098039199714863, 0.524543633626141], "https://en.wikipedia.org/wiki/Toy_Story_3": [2, 0.3098039199714863, 0.524543633626141], "https://en.wikipedia.org/wiki/Monty_Python%27s_The_Meaning_of_Life": [4, 0.3098039199714863, 0.7392833472807957], "https://en.wikipedia.org/wiki/Shark!_(film)": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Avunu_(film)": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/They_Call_It_Sin": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/Hell_Baby": [1, 0.3098039199714863, 0.3098039199714863], "https://en.wikipedia.org/wiki/The_Incident_(1967_film)": [2, 0.3098039199714863, 0.524543633626141], "https://en.wikipedia.org/wiki/Simon_Birch": [2, 0.3098039199714863, 0.524543633626141], "https://en.wikipedia.org/wiki/21_(2008_film)":

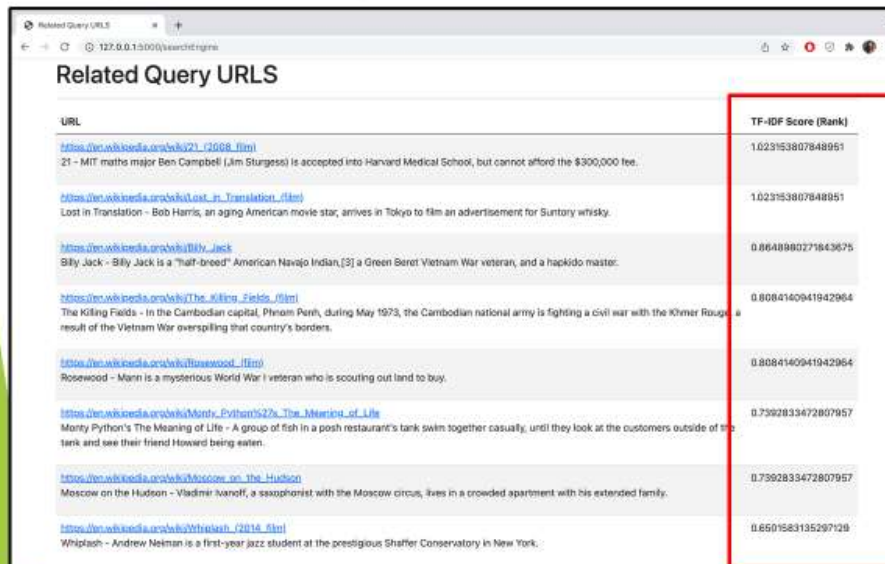
```

### Sample-1 : Single word search



Here, we are searching for only a single word. So our search engine will return only those URLs which is in relevance to this word after stemming the word and matching it with the positions document saved by the pre-processor.

## Sample-1 : Single word search Result



URL	TF-IDF Score (Rank)
<a href="https://en.wikipedia.org/wiki/21_(2008_film)">https://en.wikipedia.org/wiki/21_(2008_film)</a> 21 - MIT maths major Ben Campbell (Jim Sturgess) is accepted into Harvard Medical School, but cannot afford the \$300,000 fee.	1.023153807848951
<a href="https://en.wikipedia.org/wiki/Lost_in_Translation_(film)">https://en.wikipedia.org/wiki/Lost_in_Translation_(film)</a> Lost in Translation - Bob Harris, an aging American movie star, arrives in Tokyo to film an advertisement for Suntory whisky.	1.023153807848951
<a href="https://en.wikipedia.org/wiki/Billy_Jack">https://en.wikipedia.org/wiki/Billy_Jack</a> Billy Jack - Billy Jack is a "half-breed" American Navajo Indian,[3] a Green Beret Vietnam War veteran, and a hapkido master.	0.8648880271843675
<a href="https://en.wikipedia.org/wiki/The_Killing_Fields_(film)">https://en.wikipedia.org/wiki/The_Killing_Fields_(film)</a> The Killing Fields - In the Cambodian capital, Phnom Penh, during May 1973, the Cambodian national army is fighting a civil war with the Khmer Rouge, a result of the Vietnam War overspilling that country's borders.	0.8084140941942964
<a href="https://en.wikipedia.org/wiki/Rosewood_(film)">https://en.wikipedia.org/wiki/Rosewood_(film)</a> Rosewood - Mann is a mysterious World War I veteran who is scouting out land to buy.	0.8084140941942964
<a href="https://en.wikipedia.org/wiki/Monty_Python%27s_The_Meaning_of_Life">https://en.wikipedia.org/wiki/Monty_Python%27s_The_Meaning_of_Life</a> Monty Python's The Meaning of Life - A group of fish in a posh restaurant's tank swim together casually, until they look at the customers outside of the tank and see their friend Howard being eaten.	0.7392833472807957
<a href="https://en.wikipedia.org/wiki/Moscow_on_the_Hudson">https://en.wikipedia.org/wiki/Moscow_on_the_Hudson</a> Moscow on the Hudson - Vladimir Ivanoff, a saxophonist with the Moscow circus, lives in a crowded apartment with his extended family.	0.7392833472807957
<a href="https://en.wikipedia.org/wiki/Whiplash_(2014_film)">https://en.wikipedia.org/wiki/Whiplash_(2014_film)</a> Whiplash - Andrew Neiman is a first-year jazz student at the prestigious Shaffer Conservatory in New York.	0.6501583135297128

Results are displayed in decreasing order of TF-IDF Scores

## Sample-2 : Multiple words search



Query:

Here, we are searching for multiple words. So our search engine will return those URLs which will match with documents saved by pre-processor having both the words together in the same documents stemming the words.

## Sample-2 : Multiple words search Result

Related Query URLs	
URL	TF-IDF Score (Rank)
<a href="https://en.wikipedia.org/wiki/Avunu_(film)">https://en.wikipedia.org/wiki/Avunu_(film)</a> Avunu - Long time lovers Mohini(Poorna)and Harsha(Harshvardhan Rane) get married and move into a new community where Harsha had bought a house.	3.1972245773362196
<a href="https://en.wikipedia.org/wiki/Hell_Baby">https://en.wikipedia.org/wiki/Hell_Baby</a> Hell Baby - An expectant couple move into a haunted fixer-upper in New Orleans.	2.906907902412793
<a href="https://en.wikipedia.org/wiki/Deadly_Eyes">https://en.wikipedia.org/wiki/Deadly_Eyes</a> Deadly Eyes - The plot and subplots center on the movie's leading man, Paul Harris, a divorced high school teacher and basketball coach and his interactions during a killer rat infestation with a health department inspector, Kelly Leonard, a high school cheerleader, Trudy White, his friend a professor and rat expert, Dr.	2.09861228866811
<a href="https://en.wikipedia.org/wiki/Monty_Python%27s_The_Meaning_of_Life">https://en.wikipedia.org/wiki/Monty_Python%27s_The_Meaning_of_Life</a> Monty Python's The Meaning of Life - A group of fish in a posh restaurant's tank swim together casually, until they look at the customers outside of the tank and see their friend Howard being eaten.	1.8611995140092674
<a href="https://en.wikipedia.org/wiki/Blade_(film)">https://en.wikipedia.org/wiki/Blade_(film)</a> Blade - In 1967, a pregnant woman is attacked by a vampire while giving birth.	1.6931471805599454
<a href="https://en.wikipedia.org/wiki/Automatons_(film)">https://en.wikipedia.org/wiki/Automatons_(film)</a> Automatons - Somewhere in the distant future, The Girl is living alone in a bunker.	1.6931471805599454
<a href="https://en.wikipedia.org/wiki/RoboCop_3">https://en.wikipedia.org/wiki/RoboCop_3</a> RoboCop 3 - OCP is on the verge of bankruptcy after a series of failed business plans and drop of stocks, and are now struggling with their plans to create the new Delta City.	1.0

Results are displayed in decreasing order of TF-IDF Scores