Domain-Specific Keyword-Centric Web Crawler

Introduction -

In the current digital age, the overwhelming volume of online information makes it difficult for traditional hyperlink-driven crawlers to quickly locate specific material. This necessitates a domain-specific keyword-centric crawler. Our crawler prioritizes exact key□word searches over random hyperlink traversal, focusing on user-defined keywords to efficiently sort through web pages and exclude irrelevant content. Utilizing advanced parsing and search algorithms, it delivers highly targeted results swiftly, streamlines information retrieval, enhances recommendations, and simplifies data indexing. This innovative approach significantly improves search precision and relevance, making it essential for effective web exploration in specialized domain.

Objective –

To design and implement a domain-specific web crawler to enhance precision in targeted search result by leveraging initial keyword extraction from web content.

Data Set –

In our project, we have chosen Wikipedia as our primary database because it offers comprehensive information on virtually any topic worldwide, making it an invaluable resource for our needs. Wikipedia's vast database is freely accessible, which aligns perfectly with our project's goals of providing open and unrestricted information. Additionally, the availability of the Wikipedia API allows us to efficiently retrieve and integrate data from its extensive repository into our application. Wikipedia is also widely recognized as a reliable source of information due to its community-driven approach to content creation and rigorous editorial standards. This ensures that the information we obtain and present to our users is accurate, up-to-date, and trustworthy.

System Model –

The system architecture diagram outlines the entire workflow for the Crawler Module which is designed to enhance search efficiency and relevance through a structured sequence of steps involving user interaction, content extraction, analysis, and recommendation generation, leveraging both Wikipedia and Google search results.

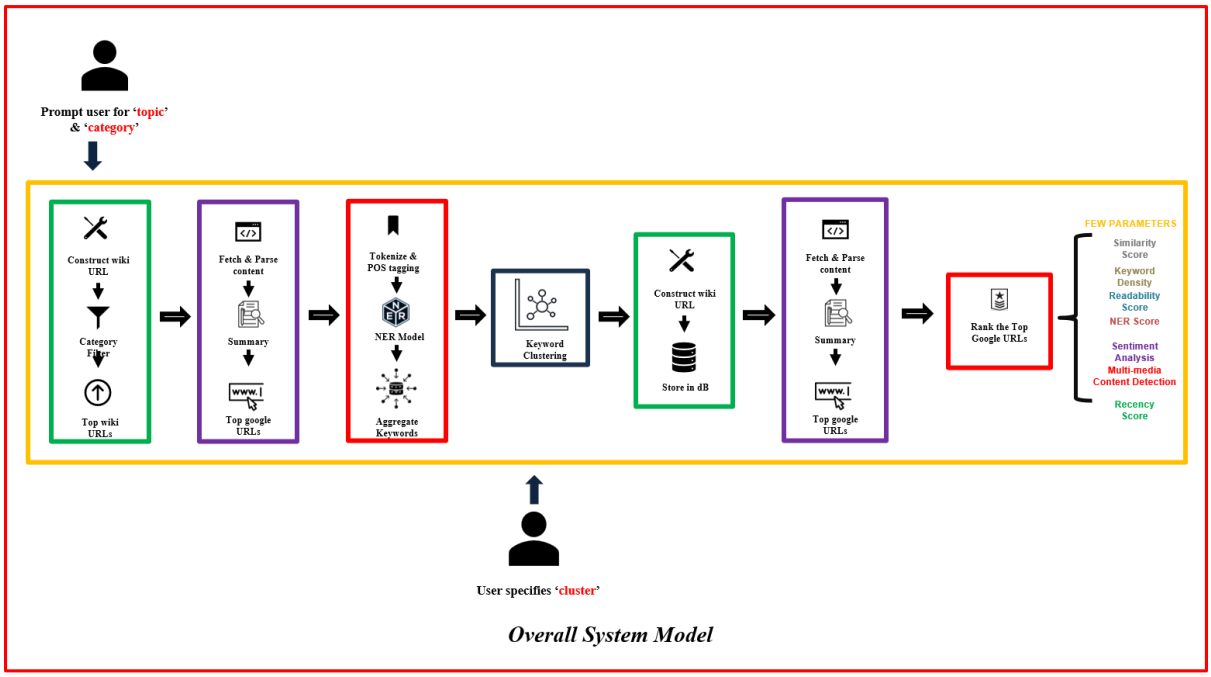
The process begins by prompting the user to specify the topic and category (optional) of interest, which tailors the content extraction and analysis to the user’s needs. Based on this input, 5 (default) Wikipedia URLs are constructed to serve as a seed URLs, acting as the primary source of high-quality information. A category filter is then applied to ensure that the URL falls within the desired scope, narrowing down the results to the most relevant articles. Following this, the system compiles a list of top Wikipedia URLs that match the user’s criteria. The next step involves fetching and parsing the content from the selected URLs, using advanced parsing techniques to extract text and other

relevant data from the web pages. A summary of the fetched content is generated by extracting the first three paragraphs from each Wikipedia page.

Simultaneously, the system fetches the top 5 URLs from Google search results for the same topic to expand the pool of data sources. The extracted content is then tokenized and tagged with Part-of-Speech (POS) labels, preparing it for further analysis. A Named Entity Recognition (NER) model is used to identify and classify entities within the text, such as names of people, organizations, and locations. The identified entities and other significant terms are aggregated to form a comprehensive list of keywords. These keywords are then clustered based on their semantic similarity, organizing them into meaningful groups. The user selects the cluster they are most interested in, allowing for personalized content delivery and ensuring relevance to their specific interests. Each cluster represents a different aspect or subtopic related to the main keyword. By analysing these clusters, users can gain a deeper understanding of various facets of their topic.

The system then begins by allowing users to select clusters of keywords. For each keyword in the chosen cluster, one Wikipedia URL is constructed and stored in a SQLite database alongside their respective keywords. These URLs serve as primary sources of information. Next, the system retrieves content from these Wikipedia URLs, parses it using BeautifulSoup to extract text, and generates summaries by condensing the first three paragraphs of each page. Additionally, summaries are used to query Google for further relevant URLs, with the top 5 results fetched and stored. This approach ensures comprehensive coverage of the chosen topics by leveraging both Wikipedia and Google sources effectively.

After fetching the top Google search results based on relevance, the script implements a custom ranking algorithm to further refine the order of URLs. This algorithm evaluates parameters such as recency score, sentiment score, similarity score, keyword density, and Named Entity Recognition (NER) score to prioritize the most pertinent results. Once ranked, the script updates its database with these URLs and presents the top-ranked URLs to the user via the console interface. This ensures that the user receives the most relevant and comprehensive information available from the web on their selected topic clusters.



Overall System Model

Algorithm -

Algorithm – Domain-Specific Keyword-Centric Crawler

Step 1 : Prompt User for Topic

Prompt the user to specify the topic T and category C of interest.

Step 2 : Construct Wikipedia URL

Construct a Wikipedia URL : $URL \leftarrow https://en.wikipedia.org/wiki/+T$

Filter by category : $URL \leftarrow FilterByCategory(T,C)$

Compile top URLs: $Top\ URLs \leftarrow CompileTopURLs(Filtered\ URLs)$

Step 3 : Fetch, Parse, and Summarize Content

Fetch and Parse : $Content \leftarrow FetchAndParse(Top\ URLs)$

Generate Summary : $Summary \leftarrow GenerateSummary(Content)$

Step 4 : Fetch and Rank Google URLs

Construct query : $Google\ Query \leftarrow ConstructQuery(Summary)$

Fetch top URLs : $Google\ URLs \leftarrow FetchTopGoogleURLs(Google\ Query)$

Rank URLs : Ranked URLs $\leftarrow RankingAlgorithm(Google\ URLs)$

Step 5 : Tokenize, POS Tag, and NER

Tokenize Content : $Tokenized\ Content \leftarrow Tokenize\ (Content)$

Apply NER : Entities $\leftarrow NERModel(Tokenized\ Content)$

Step 6 : Keyword Aggregation and Clustering

Aggregate Keywords : $Keywords \leftarrow E\ UT$

Cluster Keywords : Clusters $\leftarrow Cluster(Keywords)$

Present clusters and analyze selected cluster : $Analyze(c_i)$

Step 7 : Store, Fetch, and Display URLs

Construct URLs for keywords : $URL_j \leftarrow https://en.wikipedia.org/wiki/+k_j$

Store URLs in database : $DB\ Insert : (URL_j\ , k_j)$

Fetch and parse content : $Content_j \leftarrow FetchAndParse(URL_j\ )$

Generate Summary : $Summary_j \leftarrow ExtractSummary(Content_j\ )$

Fetch and rank Google URLs: $Google\ Query_j \leftarrow ConstructQuery\ (Summary_j\ )$

Update database with ranked URLs : $DB\ Update: UpdateGoogleURLs(URL_j\ ,\ Ranked\ URLs_j)$

Display top URLs : Display :$PrintTopURLs\ (\ .\ Ranked\ URLs_j\ )$

Implementation & Results –

The process begins by prompting the user to specify a topic and category of interest. This input is essential for constructing relevant URLs for both Wikipedia and Google sources. The steps are as follows:

1. **User Input**: The user specifies a topic and category of interest. This guides the construction of relevant URLs.
2. **Wikipedia URLs**: We create a tailored Wikipedia URL based on the user's topic and category preferences, ensuring focused content extraction and analysis.
3. **Category Filter**: A category filter is applied to ensure the results fall within the desired scope, narrowing down to the most relevant articles.
4. **Number of URLs**: By default, we use the top 5 Wikipedia URLs. This number balances breadth and depth, providing a diverse yet manageable set of high-quality, topic-specific information directly from Wikipedia, our primary source for content extraction.

This structured approach ensures that we retrieve comprehensive and relevant information efficiently.

Example 1 –
User Input (Domain) : - Bollywood

```
Enter a topic: bollywood
Enter a category (optional):
Top 5 URLs related to 'bollywood' (any category):
1. https://en.wikipedia.org/wiki/Hindi_cinema
2. https://en.wikipedia.org/wiki/Lists_of_Hindi_films
3. https://en.wikipedia.org/wiki/Bollywood_Hungama
4. https://en.wikipedia.org/wiki/List_of_highest-grossing_Indian_films
5. https://en.wikipedia.org/wiki/List_of_Hindi_horror_films
```

Fig. 1 - Top 5 Primary Wikipedia URLs acting as Seed URLs for Bollywood Domain

Each Wikipedia URL is complemented by 5 Google URLs retrieved from related search results. This approach broadens our data sources beyond Wikipedia, enriching our dataset with diverse perspectives and ensuring comprehensive coverage of the chosen topic.

```
Extracting summary from the Wikipedia article...
Generating Google search query...
Fetching additional URLs from Google...
Top 5 URLs related to the Wikipedia article:
1. https://en.wikipedia.org/wiki/Hindi_cinema
2. https://www.quora.com/What-is-the-history-of-Hindi-cinema
3. https://www.slideshare.net/slideshow/bollywood-1216177/1216177
4. https://www.slideshare.net/slideshow/100yrs-of-indian-movie/25715171
5. https://artdepartmental.com/resources/filmmaking-glossary/
Extracting summary from the Wikipedia article...
Generating Google search query...
Fetching additional URLs from Google...
Top 5 URLs related to the Wikipedia article:
1. https://en.wikipedia.org/wiki/Hindi_cinema
2. https://www.slideshare.net/slideshow/bollywood-1216177/1216177
3. https://www.quora.com/What-is-Bollywood-Is-it-a-representation-of-Indian-culture-If-yes-
4. https://www.khaleejtimes.com/bollywood/ultimate-guide-to-bollywood
5. https://ucf.digital.flvc.org/islandora/object/ucf%3A45055/datastream/OBJ/view/NOT_REALLY
Extracting summary from the Wikipedia article...
Generating Google search query...
Fetching additional URLs from Google...
Top 4 URLs related to the Wikipedia article:
1. https://en.wikipedia.org/wiki/Bollywood_Hungama
2. https://zh.wikipedia.org/zh-tw/Bollywood_Hungama
3. https://jhmovie.fandom.com/wiki/Bollywood_Hungama
4. https://dbpedia.org/page/Bollywood_Hungama
```

Fig. 2 - Top 5 Google URLs extracted for each Seed URLs

Integrating Named Entity Recognition (NER) into our pipeline enhances content summarization by identifying and categorizing entities like people, organizations, and locations from sources like Wikipedia and Google. This advanced technique improves data analysis accuracy, enriching our understanding of extracted information for more insightful reporting. The identified entities and significant terms are aggregated to form a comprehensive list of keywords.

Fig 3 – 4285 Keywords Extracted from Google

Keywords are divided into 30 clusters by default to enhance the organization and accessibility of information. This clustering process is based on semantic similarity, grouping related keywords into coherent clusters. Users can select the cluster that aligns most closely with their interests, ensuring personalized content delivery and relevance. Each cluster represents a distinct aspect or subtopic related to the main keyword, allowing users to explore different facets comprehensively. This approach not only facilitates deeper insights but also streamlines information retrieval, making complex topics more accessible and understandable.

Tables (30)

Name	Type	Schema
Cluster_awards_academy_filmfare		CREATE TABLE Cluster_awards_academy_filmfare (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_best_film_romance		CREATE TABLE Cluster_best_film_romance (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_bharti_universe_lanka		CREATE TABLE Cluster_bharti_universe_lanka (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_bollywood_hungama_type		CREATE TABLE Cluster_bollywood_hungama_type (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_bollywood_national_film		CREATE TABLE Cluster_bollywood_national_film (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_box_kantara_office		CREATE TABLE Cluster_box_kantara_office (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_cinema_bengali_quiz		CREATE TABLE Cluster_cinema_bengali_quiz (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_express_chennai_kabul		CREATE TABLE Cluster_express_chennai_kabul (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_film_city_industry		CREATE TABLE Cluster_film_city_industry (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_films_tamil_list		CREATE TABLE Cluster_films_tamil_list (Keyword TEXT)

Fig 4 – Creation of 30 clusters & storing them in database

The system allows users to select keyword clusters, each associated with a primary Wikipedia URL stored in an SQLite database. Using BeautifulSoup, it extracts text and generates summaries by condensing the initial three paragraphs of each page. Additionally, the system queries Google for further relevant URLs, extracting and storing the top 5 results. This dual-source approach ensures comprehensive coverage of chosen topics, enhancing the depth and breadth of information retrieval.

Aamir Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Salman Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Aamir	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Star Aamir Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Sajid Nadiadwala	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Mohammed Rafi	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Sameer Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Shahrukh Khan	https://en.wikipedia.org/wiki/...	https://simple.wikipedia.org/wiki/...
Shabina Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Nusrat Fateh Ali Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Ali Abbas Zafar Padmavat	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Kader Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Firoz Nadiadwala	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Sajid Khan	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Mehboob	https://en.wikipedia.org/wiki/Mahboob	https://en.wikipedia.org/wiki/Mahbo...

*Fig 5 - **36 keywords followed with Wiki and Google URLs***

After fetching top Google search results, a custom ranking algorithm refines URLs based on parameters like recency, sentiment, similarity, keyword density, and NER scores. These scores are normalized to a scale of 0 to 1 for insightful results. The script updates its database with ranked URLs and presents the most relevant ones to the user, ensuring comprehensive web information on selected topic clusters.

Wiki_URL	Google_URL	Combined_Score
Filter	Filter	Filter
https://en.wikipedia.org/wiki/Sajid–...	https://en.wikipedia.org/wiki/...	1.0
https://en.wikipedia.org/wiki/Sajid–...	https://simple.wikipedia.org/wiki/...	0.698099990977983
https://en.wikipedia.org/wiki/Sajid–...	https://kids.kiddle.co/...	0.902590880493145
https://en.wikipedia.org/wiki/Sajid–...	https://dbpedia.org/page/...	0.859426757553036
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...	0.867654483500175
https://en.wikipedia.org/wiki/...	https://kids.kiddle.co/Salim_Khan	0.764644566766447
https://en.wikipedia.org/wiki/...	https://...	0.046413650135794
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...	0.892007406065478
https://en.wikipedia.org/wiki/...	https://musicbrainz.org/artist/...	0.344286947575122
https://en.wikipedia.org/wiki/...	https://kids.kiddle.co/Aamir_Khan	0.718859598079368
https://en.wikipedia.org/wiki/...	https://pantheon.world/profile/perso...	0.653468149109513
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...	0.878424524655296
https://en.wikipedia.org/wiki/...	https://zh.wikipedia.org/zh-cn/...	0.71767766955259
https://en.wikipedia.org/wiki/...	https://www.britannica.com/...	0.788280789758231

Fig 6 - Ranked URLs with combined scores

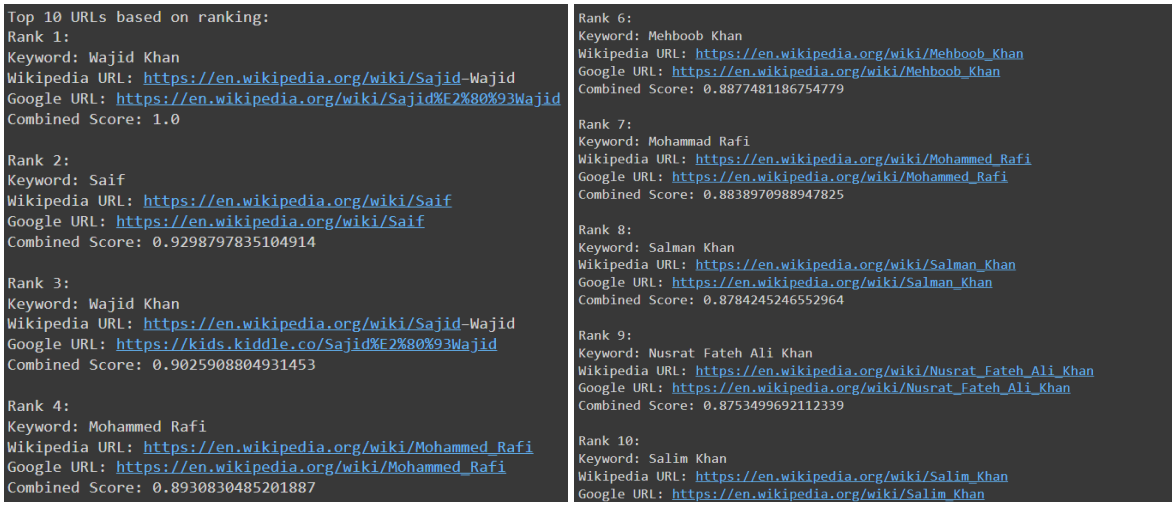


Fig 7 – Presenting top 10 ranked search results to the user

Example 2 –
User Input (Domain) : - Cricket

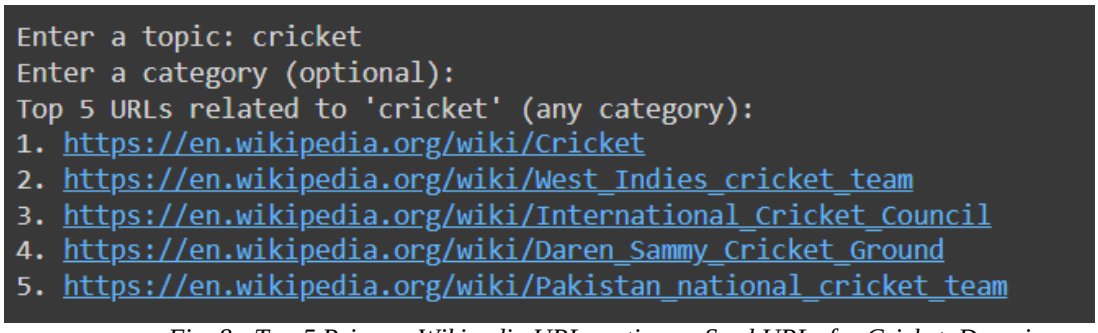


Fig. 8 - Top 5 Primary Wikipedia URLs acting as Seed URLs for Cricket Domain

Name	Type	Schema
Keyword	TEXT	"Keyword" TEXT
Cluster_men_cricket_islands		CREATE TABLE Cluster_men_cricket_islands (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_national_super_venues		CREATE TABLE Cluster_national_super_venues (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_new_zealand_guinea		CREATE TABLE Cluster_new_zealand_guinea (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_position_round_host		CREATE TABLE Cluster_position_round_host (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_rankings_team_player		CREATE TABLE Cluster_rankings_team_player (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_scorecard_asia_pakistan		CREATE TABLE Cluster_scorecard_asia_pakistan (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_series_schedule_report		CREATE TABLE Cluster_series_schedule_report (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_south_africa_kock		CREATE TABLE Cluster_south_africa_kock (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_sports_dubai_hidden		CREATE TABLE Cluster_sports_dubai_hidden (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_stadium_aziz_abu		CREATE TABLE Cluster_stadium_aziz_abu (Keyword TEXT)
Keyword	TEXT	"Keyword" TEXT
Cluster_sylhet_stadium_strikers		CREATE TABLE Cluster_sylhet_stadium_strikers (Keyword TEXT)

Fig 9 – Creation of 30 clusters & storing them in database

Keyword	URLs	GoogleURLs
Filter	Filter	Filter
Rupganj Test	https://en.wikipedia.org/wiki/...	https://www.reddit.com/r/Cricket/...
Important Matches	https://en.wikipedia.org/wiki/First-...	https://en.wikipedia.org/wiki/First-...
First Test	https://en.wikipedia.org/wiki/Test-...	https://en.wikipedia.org/wiki/Test-...
Test Status	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/First-...
Test Championship	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Bangladesh Test	https://en.wikipedia.org/wiki/...	https://www.reddit.com/r/Cricket/...
ODI Matches	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
ICC Test	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
World Test Championship	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Matches Played	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Mock Test	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
Matches	https://en.wikipedia.org/wiki/Match	https://en.wikipedia.org/wiki/Match...
Important Cricket Matches Played	https://en.wikipedia.org/wiki/First-...	https://en.wikipedia.org/wiki/First-...
Test	https://en.wikipedia.org/wiki/Test	https://quizlet.com/484316933/alber...
Test Matches	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/First-...

Fig 10 - ***15 keywords followed with Wiki and Google URLs***

Wiki_URL	Google_URL	Combined_Score
Filter	Filter	Filter
https://en.wikipedia.org/wiki/...	https://shop.ecb.co.uk/	0.410249902643148
https://en.wikipedia.org/wiki/First-...	https://en.wikipedia.org/wiki/First-...	0.708903459570936
https://en.wikipedia.org/wiki/First-...	https://www.quora.com/Has-there-...	0.0
https://en.wikipedia.org/wiki/First-...	https://en.wikipedia.org/wiki/...	0.706185518138634
https://en.wikipedia.org/wiki/Test-...	https://en.wikipedia.org/wiki/Test-...	1.0
https://en.wikipedia.org/wiki/Test-...	https://circleci.com/blog/test-driven-...	0.66536889736707
https://en.wikipedia.org/wiki/Test-...	https://www.browserstack.com/guid...	0.647102299924974
https://en.wikipedia.org/wiki/Test-...	https://stackoverflow.com/questions...	0.913695540631866
https://en.wikipedia.org/wiki/Test-...	https://agiletechnicalexcellence.com/...	0.797749178278843
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/First-...	0.707251897456714
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...	0.693510470108282
https://en.wikipedia.org/wiki/...	https://www.quora.com/Has-there-...	0.00656569500036633
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...	0.869973419573446
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...	0.803814566698715
https://en.wikipedia.org/wiki/...	https://www.topendsports.com/...	0.81606037302828
https://en.wikipedia.org/wiki/...	https://westernsportscentre.com.au/...	0.752712771524511
https://en.wikipedia.org/wiki/...	https://shop.ecb.co.uk/	0.410249902643148
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...	0.641896275242023
https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/First-...	0.62792170049126
https://en.wikipedia.org/wiki/...	https://www.quora.com/What-type-...	0.0129235419609152

Fig 11 - Ranked URLs with combined scores

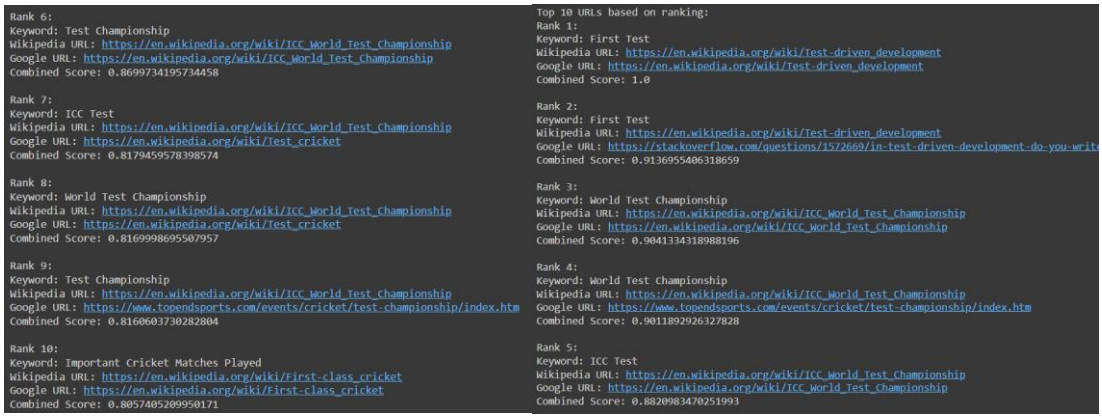
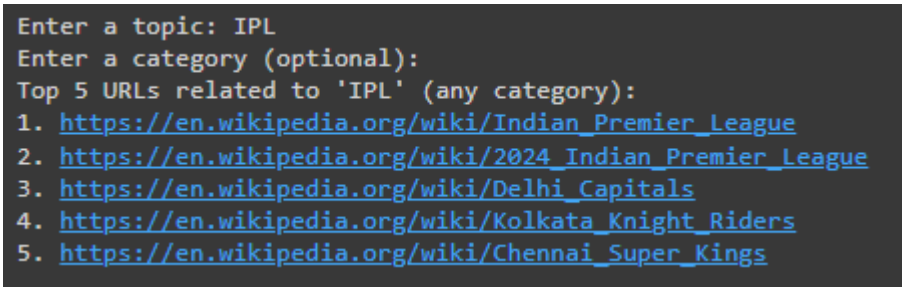


Fig 12 - Presenting top 10 ranked search results to the user

Example 3 –
User Input (Domain) : - IPL



Output – Top 5 ranked results displayed to the user