

# 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 keyword 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.

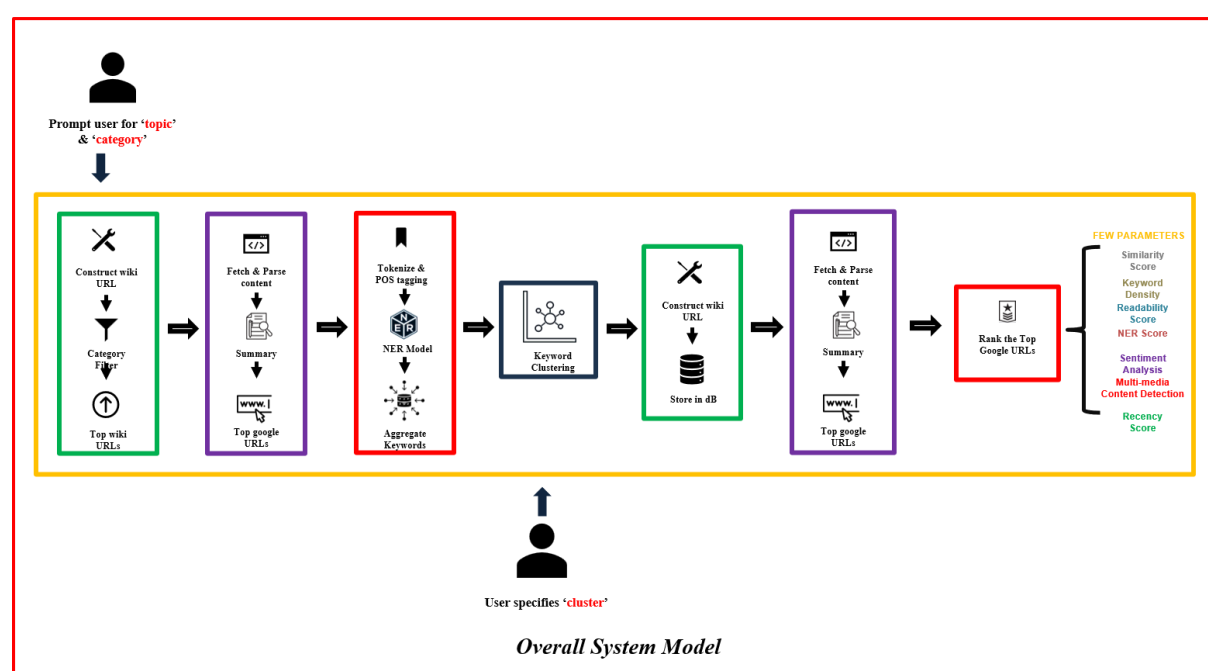
The process begins by prompting the user to specify the topic and category (optional) of interest, which tailors the content extraction and analysis to the user's needs. Based on this input, 5 (default) Wikipedia URLs are constructed to serve as seed URLs, acting as the primary source of high-quality 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.



## 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:  $Top\ URLs \leftarrow CompileTopURLs(Filter\ ed\ URLs)$

**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 :  $Keywords \leftarrow E\ \cup\ T$

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
Extracting summary from the Wikipedia article...
```

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.

Sanjay, Sanskrit, Agradh, Bruce Lee, Russiagout, IIFA Award, Shahid, Aruti Nayyar, Twinkle, Jasmine, Hay House, Sahitya Akademi Award, Disability, Nitya Bothra Indian, MEDIA, Mudaliar, Raja Mudaliar, Terash, NRI, Souten, Birla Institute, Bollywood Ancestorsquot, Kanchana, Voh Jawani Hai Dewani Box, Malaysia, Software Development Software Testing Product, Lehmann, Return, Thomson Gale, See Jagte Raho, Mughal, Learn, Bedford, Conduct Developers Statistics Cookie, Scandinavia, deet7, Anurag Basu, Movie Reviews Book, Cathedral, Screen Awards, Shah Rukh, Ajay Devgan, Sachin, Pathaan, Hooli, Triptii Dimri, Das, Champs, Gaiety Theatre, Studios, Mani Kaul, Gautam, DTU Quiz Club, Sawal Ram Singh, Movie Kings, Armenia, Rajadhyaksa, Publications Division Ministry, Karnataka, Rangroot Does Well, Shaapit, Sollywood, John, Thaindian News, Raj, Sadhana, Romantics, Copycat, Film World, Kunal, Cuckoo, Charlotte, UAE Banking, Cigarette, Leo, Computation Compiler, Carillet, Hussain, Kajol, Neeraj, Due, Kick, Viswasam, THEN, Iceland, Guinness Book, New Delhi, Padmaavat, Cheddar, World, Ormas, Daily, Wikimedia Commons, Broken Hill, Universe, Farewell My, Ek Tha, Rishab Shetty, Politics, Ranvijay Singh, Past Tumbbad, Dev Anand, Refugees, Forrest Gump, Unit II Case Study Reading Assignment, Digital Logic Software, Sahir Ludhianvi, Methods Strings Arrays, Mithun Chakraborty, Budgets, Search Go Searc, Bharti Enterprises Bharti Enterprises, Mangal Pandey, Nation, Sharmistha, Meena Iyer, Kabir Singh Box, Guinean Singer Mory Kanté, U.S., My Heart, CIS, Paper, BollySpice, Khoya Khoya Chand, DDL3, Development Machine, View Mobile Site Follow, World Works, Oldboy, Currency Exchange Gold, RRR, FIPRESCI, Español Esperanto, Squad, Kabali, Olivia Morris, Spain, Bling, Mala Sinha, JSON Turtle, Lanka Dahan, Marathi Film, Arabia Bahrain Oman Kuwait Qatar, Bollywood Bollywood Prince, Rank Title Gross Language, Mein Band Ho, Barfi, Mall Neighbours, Sociology, Ranju, Eternal, Action Cinema, Himachal Pradesh, Nasik, Ahmedabad, IP, Telegram, DVDs, Secret Superstar, Careers, Queen, Chidanand, National cultures, Dutt, Jo Jeeta Wohi Sikandar, Sumitra Devi, Wajid Khan, Sabah, Symbiosis, Hollywood Cinema, Hum Aapke Hein Koun, Hiralal Sen, Taufik Rahman PT, Dominique Leone, Rajasthan, Frommer, Gorkha, ChatGPT News, Information, Kannada Movies, Basa, Showman, Nepal, Lai Bhaari, Mandarin, Papua, Westquot, Teo, Tumhi Mere Mandir, Dear Cinema, Rajinder, RRR Total, Desire, Sanjay Dutt, Samayam Telugu, Cultural Studies, American, Ek Tha Tiger, Goosebumps, Akhtar, Ano Lectivo, Direction, Khwaja Ahmad Abbas Aan, Naseeb Mein, Favorite Movie Star, Joy Mukherji, Emergency Services, Tutorials Python Tutorial Taking Input, Telugu Hindi, Kotnis Ki Amar Kahani, Bollywood Rides, Bollywood Presents, Rubber Company, America Canada, Zanjeer, Film Industry Award Best Singer, Browser Formats, New York Times, Tamil Movies, Geeta, Film Album Best Engineer, Bollywood Madness, Style, Jigna, Global Events, Dabangg, بالاجلريه, South India, Europe Luxembourg, Phuket, Jaipur Nahargarh Fort, Glamour, Keralaakumudi, Lage Raho Munna Bhai, Gujarati, Creoles Hinglish, BHEL, Sameer Turki Sergei, U.S., Film Certification, 一比一原版, Career Beginnings Childhood, Karisma Kapoor, Popular Choice Awards Best, Pawan Kumar Chaturvedi, North, Critical Study, Bridgestone, McFarland, Male New Musical Sensation, Border, Katyar Kaljat Ghusali, Maharashtra Tourism Development Corporation, Mumbai Call Girls, Jumbo , Indian Popular Cinema, Bharat Sanchar Nigam Limited, Actor Ishaan, Labels Issue, Arveen Shaheel, Jagdish Faryadi, Deepika Padukone, Contested, Covid, Indian People, World Bank, Data Structure, Everyone DSA, Legal Aspects, Sarkar, Luxurious Call Girls Gwahati, Female Special, Express News, Ally McBeal, Abhishek Bachchan, Sound Sound, Artikel3 Artikel3 Caribbean Cuisine, Jaigarh Fort, Bibekamand, Kabayc, School, Wimal Dissanayake, Kelly Poon, Niharika, Borough Market, Pacific Exchange Rate Service, Answers, South Iceland, Examples Python, Than Zero, DNAIndia, Sprskohrvatski, Firdaus Ashraf, عرب, Duke University, Parvam, Golang, Mahesh Babu, Jyotirindra Moitra, Suhaag, Cookie 黎明, Year Release Languages Worldwide Collection, Dadasaheb Phalke Award, Filmfare Awards East, Let, Magyar Bahasa Melayu Nederlands, Matsuoaka, Americas, Anjali, MCU, Diplomat, New Zealand Bollywood, Orion Pictures, Classic Bollywood, Naushad, James, Patrick Dempsey, census, Singh Chaddha, Uzbekistan, Heroine, IPTV, Request, Comedy Films, Vietnam, Media culture, German, Aamir Khan, Zeenat Aman, Republic, Happy New Year, Hutchinson, Roopa, People, Ronda, راندا, Tamil Main, Haunted, Dying, Dinesh, Madhubala, Rakesh Roshan,

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

*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
<a href="https://en.wikipedia.org/wiki/Sajid-...">https://en.wikipedia.org/wiki/Sajid-...</a>	<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	1.0
<a href="https://en.wikipedia.org/wiki/Sajid-...">https://en.wikipedia.org/wiki/Sajid-...</a>	<a href="https://simple.wikipedia.org/wiki/...">https://simple.wikipedia.org/wiki/...</a>	0.69809990977983
<a href="https://en.wikipedia.org/wiki/Sajid-...">https://en.wikipedia.org/wiki/Sajid-...</a>	<a href="https://kids.kiddle.co/...">https://kids.kiddle.co/...</a>	0.902590880493145
<a href="https://en.wikipedia.org/wiki/Sajid-...">https://en.wikipedia.org/wiki/Sajid-...</a>	<a href="https://dbpedia.org/page/...">https://dbpedia.org/page/...</a>	0.859426757553036
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	0.867654483500175
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://kids.kiddle.co/Salim_Khan">https://kids.kiddle.co/Salim_Khan</a>	0.764644566766447
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://...">https://...</a>	0.046413650135794
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	0.892007406065478
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://musicbrainz.org/artist/...">https://musicbrainz.org/artist/...</a>	0.344286947575122
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://kids.kiddle.co/Aamir_Khan">https://kids.kiddle.co/Aamir_Khan</a>	0.718859598079368
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://pantheon.world/profile/perso...">https://pantheon.world/profile/perso...</a>	0.653468149109513
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	0.878424524655296
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://zh.wikipedia.org/zh-cn/...">https://zh.wikipedia.org/zh-cn/...</a>	0.71767766955259
<a href="https://en.wikipedia.org/wiki/...">https://en.wikipedia.org/wiki/...</a>	<a href="https://www.britannica.com/...">https://www.britannica.com/...</a>	0.788280789758231

*Fig 6 - Ranked URLs with combined scores*

<p>Top 10 URLs based on ranking:</p> <p>Rank 1: Keyword: Wajid Khan Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Sajid-Wajid">https://en.wikipedia.org/wiki/Sajid-Wajid</a> Google URL: <a href="https://en.wikipedia.org/wiki/Sajid%E2%80%93Wajid">https://en.wikipedia.org/wiki/Sajid%E2%80%93Wajid</a> Combined Score: 1.0</p> <p>Rank 2: Keyword: Saif Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Saif">https://en.wikipedia.org/wiki/Saif</a> Google URL: <a href="https://en.wikipedia.org/wiki/Saif">https://en.wikipedia.org/wiki/Saif</a> Combined Score: 0.9298797835104914</p> <p>Rank 3: Keyword: Wajid Khan Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Sajid-Wajid">https://en.wikipedia.org/wiki/Sajid-Wajid</a> Google URL: <a href="https://kids.kiddle.co/Sajid%E2%80%93Wajid">https://kids.kiddle.co/Sajid%E2%80%93Wajid</a> Combined Score: 0.9025908804931453</p> <p>Rank 4: Keyword: Mohammed Rafi Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Mohammed_Rafi">https://en.wikipedia.org/wiki/Mohammed_Rafi</a> Google URL: <a href="https://en.wikipedia.org/wiki/Mohammed_Rafi">https://en.wikipedia.org/wiki/Mohammed_Rafi</a> Combined Score: 0.8930830485201887</p>	<p>Rank 6: Keyword: Mehboob Khan Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Mehboob_Khan">https://en.wikipedia.org/wiki/Mehboob_Khan</a> Google URL: <a href="https://en.wikipedia.org/wiki/Mehboob_Khan">https://en.wikipedia.org/wiki/Mehboob_Khan</a> Combined Score: 0.8877481186754779</p> <p>Rank 7: Keyword: Mohammad Rafi Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Mohammed_Rafi">https://en.wikipedia.org/wiki/Mohammed_Rafi</a> Google URL: <a href="https://en.wikipedia.org/wiki/Mohammed_Rafi">https://en.wikipedia.org/wiki/Mohammed_Rafi</a> Combined Score: 0.8838970988947825</p> <p>Rank 8: Keyword: Salman Khan Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Salman_Khan">https://en.wikipedia.org/wiki/Salman_Khan</a> Google URL: <a href="https://en.wikipedia.org/wiki/Salman_Khan">https://en.wikipedia.org/wiki/Salman_Khan</a> Combined Score: 0.8784245246552964</p> <p>Rank 9: Keyword: Nusrat Fateh Ali Khan Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Nusrat_Fateh_Ali_Khan">https://en.wikipedia.org/wiki/Nusrat_Fateh_Ali_Khan</a> Google URL: <a href="https://en.wikipedia.org/wiki/Nusrat_Fateh_Ali_Khan">https://en.wikipedia.org/wiki/Nusrat_Fateh_Ali_Khan</a> Combined Score: 0.8753499692112339</p> <p>Rank 10: Keyword: Salim Khan Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Salim_Khan">https://en.wikipedia.org/wiki/Salim_Khan</a> Google URL: <a href="https://en.wikipedia.org/wiki/Salim_Khan">https://en.wikipedia.org/wiki/Salim_Khan</a></p>
--	---

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

## Example 2 –

User Input (Domain) : - Cricket

```

Enter a topic: cricket
Enter a category (optional):
Top 5 URLs related to 'cricket' (any category):
1. https://en.wikipedia.org/wiki/Cricket
2. https://en.wikipedia.org/wiki/West\_Indies\_cricket\_team
3. https://en.wikipedia.org/wiki/International\_Cricket\_Council
4. https://en.wikipedia.org/wiki/Daren\_Sammy\_Cricket\_Ground
5. https://en.wikipedia.org/wiki/Pakistan\_national\_cricket\_team

```

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

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



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

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

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

*Fig 11 - Ranked URLs with combined scores*



<p>Rank 6: Keyword: Test Championship Wikipedia URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Google URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Combined Score: 0.8699734195734458</p> <p>Rank 7: Keyword: ICC Test Wikipedia URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Google URL: <a href="https://en.wikipedia.org/wiki/Test_cricket">https://en.wikipedia.org/wiki/Test_cricket</a> Combined Score: 0.8179459578398574</p> <p>Rank 8: Keyword: World Test Championship Wikipedia URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Google URL: <a href="https://en.wikipedia.org/wiki/Test_cricket">https://en.wikipedia.org/wiki/Test_cricket</a> Combined Score: 0.8169998695507957</p> <p>Rank 9: Keyword: Test Championship Wikipedia URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Google URL: <a href="https://www.topendsports.com/events/cricket/test-championship/index.htm">https://www.topendsports.com/events/cricket/test-championship/index.htm</a> Combined Score: 0.8160603730282804</p> <p>Rank 10: Keyword: Important Cricket Matches Played Wikipedia URL: <a href="https://en.wikipedia.org/wiki/First-class_cricket">https://en.wikipedia.org/wiki/First-class_cricket</a> Google URL: <a href="https://en.wikipedia.org/wiki/First-class_cricket">https://en.wikipedia.org/wiki/First-class_cricket</a> Combined Score: 0.8057405209950171</p>	<p>Top 10 URLs based on ranking:</p> <p>Rank 1: Keyword: First Test Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Test-driven_development">https://en.wikipedia.org/wiki/Test-driven_development</a> Google URL: <a href="https://en.wikipedia.org/wiki/Test-driven_development">https://en.wikipedia.org/wiki/Test-driven_development</a> Combined Score: 1.0</p> <p>Rank 2: Keyword: First Test Wikipedia URL: <a href="https://en.wikipedia.org/wiki/Test-driven_development">https://en.wikipedia.org/wiki/Test-driven_development</a> Google URL: <a href="https://stackoverflow.com/questions/1572669/in-test-driven-development-do-you-write">https://stackoverflow.com/questions/1572669/in-test-driven-development-do-you-write</a> Combined Score: 0.9136955406318659</p> <p>Rank 3: Keyword: World Test Championship Wikipedia URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Google URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Combined Score: 0.9041334318988196</p> <p>Rank 4: Keyword: World Test Championship Wikipedia URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Google URL: <a href="https://www.topendsports.com/events/cricket/test-championship/index.htm">https://www.topendsports.com/events/cricket/test-championship/index.htm</a> Combined Score: 0.9011892926327828</p> <p>Rank 5: Keyword: ICC Test Wikipedia URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Google URL: <a href="https://en.wikipedia.org/wiki/ICC_World_Test_Championship">https://en.wikipedia.org/wiki/ICC_World_Test_Championship</a> Combined Score: 0.8820983470251993</p>
--	--

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

### Example 3 – User Input (Domain) : - IPL

```

Enter a topic: IPL
Enter a category (optional):
Top 5 URLs related to 'IPL' (any category):
1. https://en.wikipedia.org/wiki/Indian\_Premier\_League
2. https://en.wikipedia.org/wiki/2024\_Indian\_Premier\_League
3. https://en.wikipedia.org/wiki/Delhi\_Capitals
4. https://en.wikipedia.org/wiki/Kolkata\_Knight\_Riders
5. https://en.wikipedia.org/wiki/Chennai\_Super\_Kings

```

Output – Top 5 ranked results displayed to the user

```

Top URLs based on ranking (with unique Google URLs):
Rank 1:
Keyword: KKR Kolkata Knight Riders
Wikipedia URL: https://en.wikipedia.org/wiki/Kolkata\_Knight\_Riders
Combined Score: 1.0

Rank 2:
Keyword: KKR Kolkata Knight Riders
Wikipedia URL: https://en.wikipedia.org/wiki/Kolkata\_Knight\_Riders
Combined Score: 0.9364998858455528

Rank 3:
Keyword: Knight Rider
Wikipedia URL: https://en.wikipedia.org/wiki/Knight\_Rider
Combined Score: 0.896371112212667

Rank 4:
Keyword: KKR Kolkata Knight Riders
Wikipedia URL: https://en.wikipedia.org/wiki/Kolkata\_Knight\_Riders
Combined Score: 0.8826980251880616

Rank 5:
Keyword: Abu Dhabi Knight Riders
Wikipedia URL: https://en.wikipedia.org/wiki/Abu\_Dhabi\_Knight\_Riders
Combined Score: 0.8625490986495513

```