

**Final Year Project Proposal**

**Project Title**

**Intelligent Web Spider for Efficient Search Engine**

**submitted by**

**Abdullah Rather (35472)**

**Bilal Ahmed (35708)**

**Muhammad Hassaan (35387)**

**Syed Farid Uddin (35485)**

**Supervisor**

**Sir Israr Ali**

**Coordinator**

**Dr. Mansoor Ebrahim**

**List of Abbreviations and Acronyms**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Abbreviation & Acronym** | **Description** |
|  | WWW | World Wide Web |
|  | SE | Search Engine |
|  | WS | Web Spider |
|  | Int | Intelligent |
|  | Algo | Algorithm |
|  | DB | Database |
|  | IWS | Intelligent Web Spider |

**Table of Contents**

**Description Page #**

[Section – 1 3](#_Toc63687996)

[1.1 Project Identification 3](#_Toc63687997)

[Section – 2 8](#_Toc63687998)

[2.1 Introduction 8](#_Toc63687999)

[2.2 Aim and Objectives 9](#_Toc63688000)

[2.3 Benefits 10](#_Toc63688001)

[2.4 Functional Requirements 11](#_Toc63688002)

[2.5 Non-Functional Requirements 11](#_Toc63688003)

[2.6 Literature Review 12](#_Toc63688004)

[2.7 Research Approach 16](#_Toc63688005)

[Section – 3 20](#_Toc63688006)

[3.1 Annexure–A: Project Schedule / Milestone Chart 20](#_Toc63688007)

[3.2 Functional & Non-Functional Requirements 21](#_Toc63688008)

[3.3 Use Case: 22](#_Toc63688009)

[3.4 Narratives 24](#_Toc63688010)

[3.5 ERD 25](#_Toc63688011)

[3.6 Prototype 27](#_Toc63688012)

[3.7 Annexure–B: Proposed Budget 30](#_Toc63688013)

[3.8 Annexure–C: Business Canvas Model 31](#_Toc63688014)

**Final Year Project Proposal**

# Section – 1

# 1.1 Project Identification

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Project Title: | | | | | | | |
| **Intelligent Web Spider for Efficient Search Engine** | | | | | | | |
| Group Leader (GL): | | | | | | | |
| **1.** | Name: | **Abdullah Rather** | | | | | |
|  | Reg.#: | **021-16-35472** | | | | | |
|  | CGPA: | **2.98** | | | | | |
|  | Mobile #: | **+923341319716** | Email: | | | [**abdullahrather19@gmail.com**](mailto:abdullahrather19@gmail.com) | |
|  | Signature: | My Signature | | | | | |
| **Group Members (GM’s):** | | | | | | | |
| **2.** | Name: | **Bilal Ahmed** | | | | | |
|  | Reg.#: | **021-16-35708** | | | | | |
|  | CGPA: | **2.45** | | | | | |
|  | Mobile #: | **+923312543117** | | | Email: | | [**bilal.taurus92@gmail.com**](mailto:bilal.taurus92@gmail.com) |
|  | Signature: | bilal | | | | | |
| **3.** | Name: | **Muhammad Hassaan** | | | | | |
|  | Reg.#: | **021-16-35387** | | | | | |
|  | CGPA: | **2.64** | | | | | |
|  | Mobile #: | **+923343602355** | | | Email: | | [**hassaan.rasheed125@gmail.com**](mailto:hassaan.rasheed125@gmail.com) |
|  | Signature: | Hassaan Signature | | | | | |
| **4.** | Name: | **Syed Farid Uddin** | | | | | |
|  | Reg.#: | **021-16-35485** | | | | | |
|  | CGPA: | **2.42** | | | | | |
|  | Mobile #: | **+923413599426** | | Email: | | | [**syedzada148@gmail.com**](mailto:syedzada148@gmail.com) |
|  | Signature: | Farid | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| What technology is core to your product? *(Please mark* 🗹 *where applicable)*  |  |  |  | | --- | --- | --- | | [ ] 3D/4D Printing | | [ ] Augmented Reality / Virtual Reality | | [**🗹** ] Big Data, Artificial Intelligence | | [ ] Blockchain | | [ ] Cloud | | [ ] Neurotech | | [ ] Robotics | | [ ] Shared economy | | [ ] The Internet of Things | | [ ] Wearables, Implantables | | [**🗹** ] Others (specify): | Web Crawling | | | | | | |
|  | | | | |
| What is the target market(s) for the products? *(Please mark* 🗹 *where applicable)*  |  |  |  | | --- | --- | --- | | [ ] Automotive, aviation, marine | | [**🗹**] Business, marketing, finance | | [**🗹** ] Defense, security, safety | | [ ] Education and training | | [ ] Environment, water management | | [ ] Entertainment, tourism, sport/recreation | | [ ] Food, livestock, agribusiness | | [ ] Healthcare | | [ ] Infrastructure, housing & transport | | [ ] Mining equipment technology & services | | [ ] Oil, gas, energy | | [ ] Textiles, clothing, footwear | | [ ] Others (specify): |  | | |  | |  | | | | | |
| Other Organizations Involved in the Project: *(Please identify all affiliated organizations collaborating in the project, and describe their role/contribution to the project.)* | | | | |
| **Academic Organizations:** | | | | |
| *#* | *Organization Name* | | | *Role / Contribution* |
|  | Iqra University (Main Campus) | | | Resources & Supervision |
|  | - | | | - |
| **Industrial Organizations:** | | | | |
| *#* | *Organization Name* | | *Role / Contribution* | |
|  | - | | - | |
|  | - | | - | |
| **Funding Organizations:** | | | | |
| *#* | *Organization Name* | | | *Role / Contribution* |
|  | - | | | - |
|  | - | | | - |
| Key Words: *(Please provide a maximum of 5 key words that describe the project)* | | | | |
| * Keywords for Searching * Web Crawler * Efficient Searching * Keyword Storage Optimization * Data Cluster | | | | |
| Research and Development Theme: *(please identify the Research Theme.)* | | | | |
| We shall study different methodologies for this project and select the best one which gives will generate efficient results. We shall study and research different algorithms to identify keywords from the websites, apply linguistic techniques to simplify generated keywords and they will apply generate efficient search results based on the searching from these keywords. We shall also study and implement Object Oriented Technique for development of our project. | | | | |
|  | | | | |
| Project Status: (Please mark 🗹) [**🗹**] New [ ] Modification to previous Project  [ ] Extension of existing project | | | | |
| Project Duration: | | 12 Months | | |
| Proposed Budget: | | Rs. 3,635,000 PKR | | |
|  | | | | |
| The Problem:*(Please describe the problem / opportunity / gap / need that you want to address/solve. Maximum 300 words.)* | | | | |
| Search engine uses keywords to find results from the web. These keywords are collected by Web Crawlers. The duty of a crawler is to crawl through websites automatically, analyze the data written in text format and save this data into manageable hierarchies by applying clustering techniques. This saved data will be accessible to Search Engines to search results based on user’s query and list related web pages.  Now-a-days, the databases for search engines are being populated by using automated tools. These tools search the keywords on each site and maintain a database on the basis of search results. These tools can also be used to analyze statistics and content of any site for searching purpose. But the problem is their cost, these tools are costly and inaccessible for students and small level professionals. Moreover, these tools do not provide the flexibility to define level of depth for searching or specific domain level searching / keyword generation.  So a tool to answer these problem is needed that will provide fast and accurate search results for general users based on their own interest by using some intelligent searching techniques. | | | | |
|  | | | | |
|  | | | | |
| Synopsis:*(A brief description of the idea, in non-technical language, explaining product benefit, target market, basic technology, commercial partners, investors, and potential customers. Maximum 200 words.)* | | | | |
| In today’s modern world, where social media and other information sources are producing humongous data every moment, it is the need to build smart and efficient searching tool that should be able to collect keywords from websites, perform some filtration to purify repetitive / non-required data and then store collected keywords in a database so that searching becomes fast and storage space may be reduced.  This type of tool has huge benefits, some of which are as follows:   * Website navigation to fetch its contents * Collect keywords from any website around the world * Apply linguistic algorithm to filter the keywords * Test links of the website to identify its valid structure/hierarchy. * Means if some website has large number of pages, then it’s a headache to check the correctness of hierarchy of that site. Whereas our tool will make ease for him/her to test the correctness of hierarchy. * While crawling through the website, it can also identify bad links (the links which are invalid or are without target pages) * It may search for copyright violations by filtering keywords. * Clustered based keyword storage for efficient retrieval * It will Maintain ratings of keywords   This type of tool may be used in data intensive markets where data set is really big and searching on such Big Data takes time. Some of the examples may be Business industry, Defense, Electronic Media, Universities and other Government and Private offices.  We would like to develop this tool in Java because java is an efficient language and it will be best suited for such projects. Moreover, we have learned technical expertise in java during our semesters.  We think Universities, Government and Defense Organizations may collaborate with our group in becoming our commercial partners / investors. | | | | |

# Section – 2

# 2.1 Introduction

|  |
| --- |
| **Challenges:** |
| Now-a-days, the databases for search engines are being populated by using automated tools called web crawler. These tools search the keywords on each site and maintain the database on the basis of search results.  These tools can also be used to see the statistics and content of any site for search purpose. But these tools are costly and inaccessible for students and small level professional to use and also these tools don’t provide the complete access and authentication for users, e.g. User can’t restrict the level of depth of search, or any specific domain to search from.  Secondly, the keywords generated by such tools are very large in number from each website and a huge repetition of similar keyword are also generated like normally website are based on single business domain and it is very common that similar words are used on multiple pages. Therefore, the keywords generated on each page are also repeated.  Another big challenge is the storage of generated keywords in a way that is efficient for future searching. Routine storage in database |
| **Motivation and Need:** *(Please describe the motivation and need for this work.* ***Maximum 500 words.****)* |
| Efficient and accurate searching from a huge data set of websites available on the web now a days is a gigantic challenge. Number of universities and various organizations worldwide are doing extensive research to find out ways to make systems that can search the desired information from the widespread information available on the internet in minimum possible time with accurate results. In recent past if one looks behind, the technology giant Google stated its journey just by providing a revolutionary search engine that became so much popular that today it is the top ranked IT organization of the world beating behind many technology giants of that era.  While keeping the importance of efficient searching in mind, we started exploring ways to provide raw material that search engines can utilize to enhance their efficacy and generate better results while maintaining lowest possible foot print and storage size. |

# 2.2 Aim and Objectives

|  |
| --- |
| *(Please describe the measurable objectives of the project and define the expected results. Use results-oriented wording with verbs such as ‘to develop..’, ‘to implement..’, ‘to research..’, ‘to determine..‘, ‘to identify..’* ***The objectives should not be statements and should actually specify in simple words what the project team intends to achieve (something concrete and measurable/ deliverable). Fill only those objectives that are applicable to the proposed project****.)* |
| **Research Objectives:** *(if any)*   * To identify a viable linguistic algorithm that could simplify and transform keywords so that storage could be minimized without compromising searching efficiency of search engine.   **Academic Objectives:** *(if any)*   * To do the research in exploring linguistic algorithm that is best suited for our needs * To explore efficient storage mechanism for keywords in database clusters   **Commercial Objectives:** *(if any)*   * <type here>   **Other Objectives:** *(if any)*   * <type here> |

# 2.3 Benefits

|  |
| --- |
| Expected Outcomes: *(Provide a list of proposed project outputs including publications, databases etc.)*  * Fetch keywords from the contents of a website * Make a list of URLs by crawling the website and make a web of URLs for subsequent keyword fetching * Simplify keywords by applying linguistic algorithm * Store keywords * Maintain frequency of keywords for better searching |
| Key Benefits and Beneficiaries: *(Please identify clearly the benefits and potential customers/beneficiaries of the project.)*  * Controlled crawling option for user * Dataset of keywords generated from a website   **Beneficiaries**   * General Users * Academia * Government Departments |
| Technology Transfer/Diffusion Approach: *(Please describe how the outputs of the project will be transferred to the beneficiaries/customers. Maximum 500 words.)*Fetching desired output from websites is a difficult task. Our solution will allow the user to identify and select a website and tell our crawler to generate keyword from that specific website. Secondly, simplify the generated keywords by applying efficient algorithm so that storage may be reduced without compromising searching efficiency. |

# 2.4 Functional Requirements

|  |
| --- |
| **List of Functional Requirements :** |
| * System should have a user management mechanism that will allow access to the user to use the system * The crawling depth should be a positive integer, and can’t be infinity. * System should fetch keywords from the contents of URLs provided by the user. * System should transform keywords generated into simplified words * System should keep user posted about the results * Generate frequency of keywords * Store keywords in respective database clusters |
|  |
|  |

# 2.5 Non-Functional Requirements

|  |
| --- |
| * Only authorized persons/administrators should be able to run this system * A reliable internet to fetch keywords form website as the performance is dependent on bandwidth * Software should be available and workable as and when required to the user * System should be flexible to accommodate future releases / enhancements * System should be cost effective |

# 2.6 Literature Review

|  |
| --- |
| **Literature Review:** *(Detailed summary of what all has been done internationally in the proposed area quoting references and bibliography.* ***Maximum 1500 words.****)* |
| Literature review is conducted before the development of the proposed system ‘Intelligent Web Spider for Efficient Search Engine’. Many crawling software were in the domain of this survey. Some of them are listed here with their brief description. **Trellian SiteSpider**  Trellian SiteSpider is a powerful, yet simple to use program which is loaded with sophisticated search, web crawling and site mapping functions. SiteSpider can be used simply as a web browser, which works cooperatively with Internet Explorer to allow importing and utilization of our favorite’s directory. Trellian SiteSpider can help to extract valuable data from any website.  bluechillies **Icegiant Data Extractor** MP3 Extractor is a state-of-the-art internet MP3 download manager. It’s easy to use design and multithreaded crawling engine allows the user to quickly download MP3 urls from a website. MP3 Extractor ties directly into Microsoft's Internet Explorer as a toolbar for easy access while browsing. After download is finished, the acquired urls are displayed on a page for easy viewing.  MP3Extractor-92810 **ETagFix by ISAPILabs** ETagFix is an ISAPI plugin for Internet Information Server (IIS) that changes the etag produced by IIS so that the etag does not change every time out computer reboots. This can result in web crawlers and user clients wanting to reach our web page when they don't need to.  ETagFix-26090 **StreamBull by J.P.C. Development Services**  StreamBull is a fast and easy-to-use multi-threaded adult picture and video downloader. It reads the necessary internet addresses from a database residing on an internet server. The user selects his favorite categories like teens, science, research etc. (32 x 2 in all) and gets images accordingly. All images have been pre-screened to avoid multiples or illegal content, and to avoid the local downloader having to spend time on filtering or searching.  StreamBull-6160 **Inspyder Site Map Creator** Inspyder Sitemap Creator is a web-crawling XML sitemap file generator. It's easy to use, yet powerful enough to handle large sites. Inspyder Sitemap Creator can help us generate an accurate sitemap and help us keep it up to date.  SitemapCreator-medium **Comparison** All of these mentioned software tools are for public usage. Anyone can buy and use them, but their usage is limited to MP3 extraction, site map creation, image searching etc. If a person want to scan any specific web page to get to know the keywords at some page, then these tools might not be suitable. **Summary** So our web spider will provide the functionality of scanning web page, to extract keywords on them. These keywords will not only be helpful to define the content type of the webpage but also helpful to categorize the webpage on the basis of keywords. Moreover, storing these keywords in an efficient form will help reduce storage and improve search engine efficiency providing fast keyword selection on runtime. |

# 2.7 Research Approach

|  |
| --- |
| Development / Research Methodology: *(Please describe the technical details and justification of your development and research plan. The block diagrams, system flow charts, high level algorithm details etc. have to be provided in this section.* ***Maximum 3000 words.****)* |
| The research and development plan for our proposed project “Intelligent Web Spider for Efficient Search Engine” is based on study of various available systems and comparing the features of each system with our envisaged functional and nonfunctional requirements.  **Development Plan**  Our development plan is divided in two phases, which are as follows:  • Database population from keywords (with Spider/Crawler)  • Web application, to search from keyword database.  The responsibility of first module will be to move like a spider from one site to other, and to find the keywords of each webpage. These keywords will be then be separated from a list of words like a, an, the, that, is, are, etc. And then the keywords will go through a linguistic algorithm that will stemmed these keywords by using Stemming Algorithm. Means the Management will be converted into Manage, Meetings into Meet and Cats into Cat etc. And these stemmed words will be stored in the database. This makes storage size reduced as when billions of keywords will be stemmed this will save ample space. While searching from site to site, this software will consider the privileges set for users on current site, means it will also examine the robots.txt (i.e. A file usually placed on restricted domains, and this file describes the privileges to use the content of that site), this software will also do the cycle handling while moving from site to site and will also counts number of times a certain website is cited, means it will also do the statistical analysis. And on the basis of these analyzed statistics, the system will manage found keywords into manageable clusters/categories.  Second module will be web based and the responsibility of this module will be to search from the database maintained by the first module. It will search by taking value/input from user. And before search from database, this input will be stemmed by using the same Stemming algorithm i.e. used by first module.  **Block Diagram of the System**    Above picture shows a high level block diagram of web crawler. It shows that the crawler has two major processors i.e. Scheduler and Multi-Threaded Downloader. The job of downloader is to initiate and send HTTP request to server (or Internet) for downloading content of Web pages. Then the URLs and Text/metadata is separated from the downloaded content. This separated text/metadata is to be stored on Storage location (i.e. database). And the list of URLs will be passed to Scheduler for recursively download.  **Flow Chart**    Above is the high level application flow chart depicting starting point, processing nodes and ending phase of the project. |

|  |  |  |  |
| --- | --- | --- | --- |
| Key Milestones and Deliverables: *(Please list and describe the principal milestones and associated deliverables of the project. The timing of milestones is also to be shown in the Gantt chart in Annexure-A.* ***Quarterly deliverables are preferred****.)* | | | |
|  | | | |
| **S. No.** | **Duration** | **Milestone** | **Deliverable** |
| 1. | Week 01 – 04 | Requirement Analysis | SRS Document |
| 2. | Week 05 – 08 | Learning Technologies | Learn Java, C++, PHP, HTML, CSS |
| 3. | Week 09 – 12 | Linguistic Algo | Identify, Understand and Shortlist linguistic algorithm best suited for this project |
| 4. | Week 13 – 16 | Technical Design | Technical Design Document |
| 5. | Week 17 – 20 | Application Development | Crawler Application |
| 6. | Week 25 – 24 |
| 7. | Week 25 – 28 |
| 8. | Week 29 – 32 | Web Application |
| 9. | Week 37 – 36 |
| 10. | Week 37 – 40 |
| 11. | Week 41 – 45 | Application Testing | Testing and Bugs Fixing |
| 12. | Week 46 – 50 | User Documentation | User Guide |
| 13. | Week 51 – 52 | Application Installer |
| 14. | Week 51 – 52 | Demo | Application Demonstration |
| (Please add more rows if required.) | | | |

# Section – 3

# 3.1 Annexure–A: Project Schedule / Milestone Chart

***(Project schedule using MS-Project (or similar tools) with all tasks, deliverables, milestones, cost estimates, payment schedules clearly indicated are preferred.)***

**MS Project GantChart .mpp file Double click icon to open**





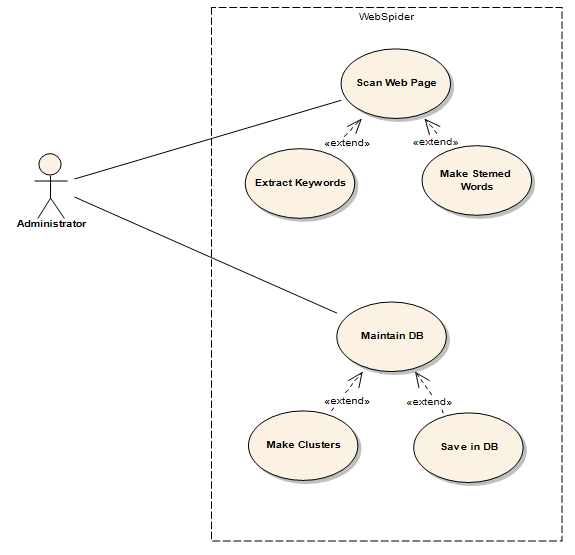
# 3.2 Functional & Non-Functional Requirements

**Functional requirements define** what the system does or must not do, **non-functional requirements** specify how the system should do it. **Non-functional requirements** do not affect the basic **functionality** of the system (hence the name, **non-functional requirements**).

* 1. **List of Functional Requirements:**
  + System should have a user management mechanism that will allow access to the user to use the system
  + The crawling depth should be a positive integer, and can’t be infinity.
  + System should fetch keywords from the contents of URLs provided by the user.
  + System should transform keywords generated into simplified words
  + System should keep user posted about the results
  + Generate frequency of keywords
  + Store keywords in respective database clusters
  1. **List of Non-Functional Requirements:**
  + Only authorized persons/administrators should be able to run this system
  + A reliable internet to fetch keywords form website as the performance is dependent on bandwidth
  + Software should be available and workable as and when required to the user
  + System should be flexible to accommodate future releases / enhancements
  + System should be cost effective

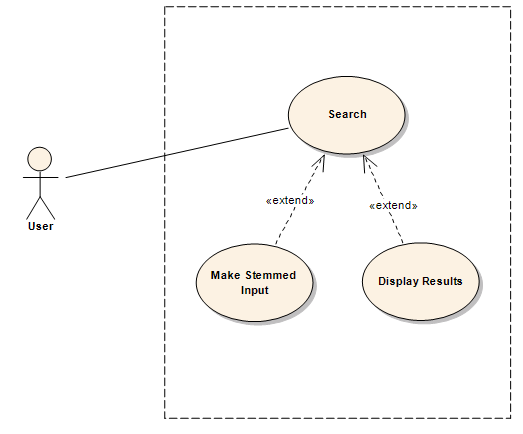
# 3.3 Use Case:

* 1. **Administrative Use Case**



|  |  |  |
| --- | --- | --- |
| **Use Case Name:** | Web Spider | |
| **ID:** | UC-LHD-01 | |
| **Actors Involved:** | Administrator | |
| **Brief Description** | Administrator initiates web spider to start crawling the web pages and generate keywords.  Then after stemming these keywords, store keywords in respective clustered database | |
| **Pre-Conditions** | System is available for use | |
| **Post-Conditions** | Keywords stored in database | |
| **Normal Flow of Events:** | **Actor Action** | **System Response** |
| 1. Administrator starts the application by clicking a button. 2. Application starts crawling web pages 3. Keywords are granted and stored in respective clustered database | 1. Application starts crawling web pages 2. Keywords are granted and stored in respective clustered database. |

* 1. **User Use Case**

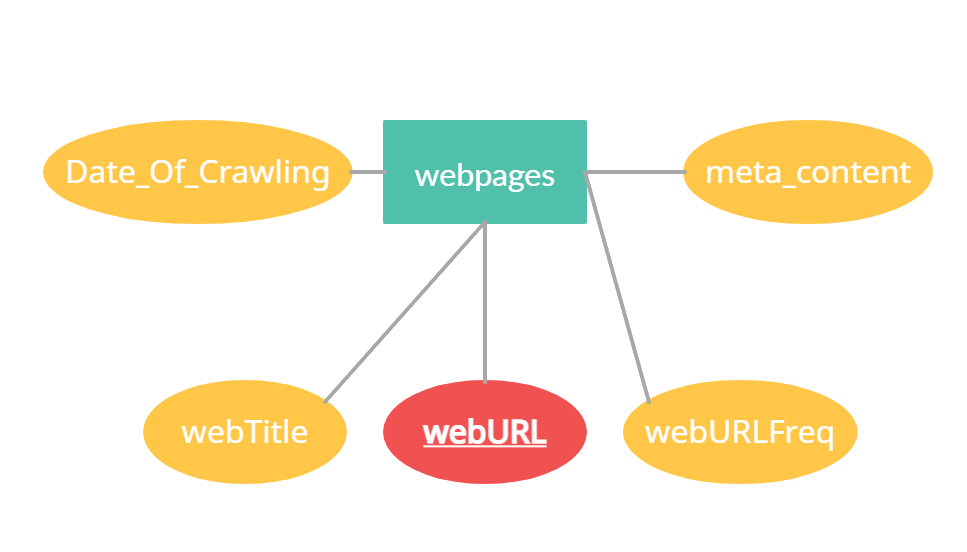
****

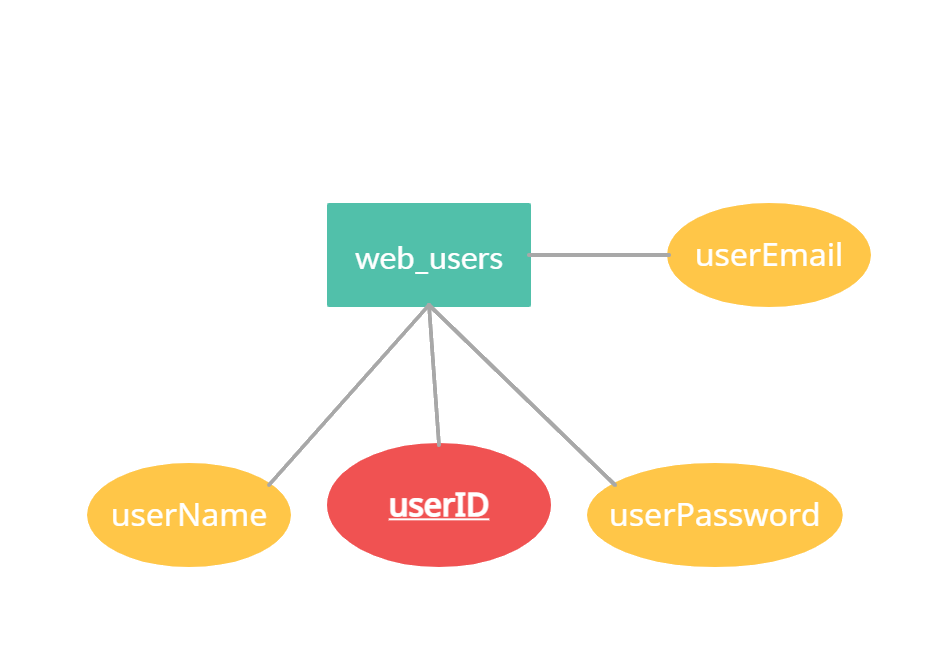
|  |  |  |
| --- | --- | --- |
| **Use Case Name:** | Search from keywords | |
| **ID:** | UC-CNT-002 | |
| **Actors Involved:** | User | |
| **Brief Description** | Convert user input search criteria into stemmed input and apply searching from generated keywords | |
| **Pre-Conditions** | Use Case ID: UC-LHD-01 | |
| **Post-Conditions** | Results are displayed. | |
| **Normal Flow of Events:** | **Actor Action** | **System Response** |
| 1. User provide search string 2. Press Search button | 1. System will convert input into stemmed input 2. Apply search from keyword database 3. Display results |

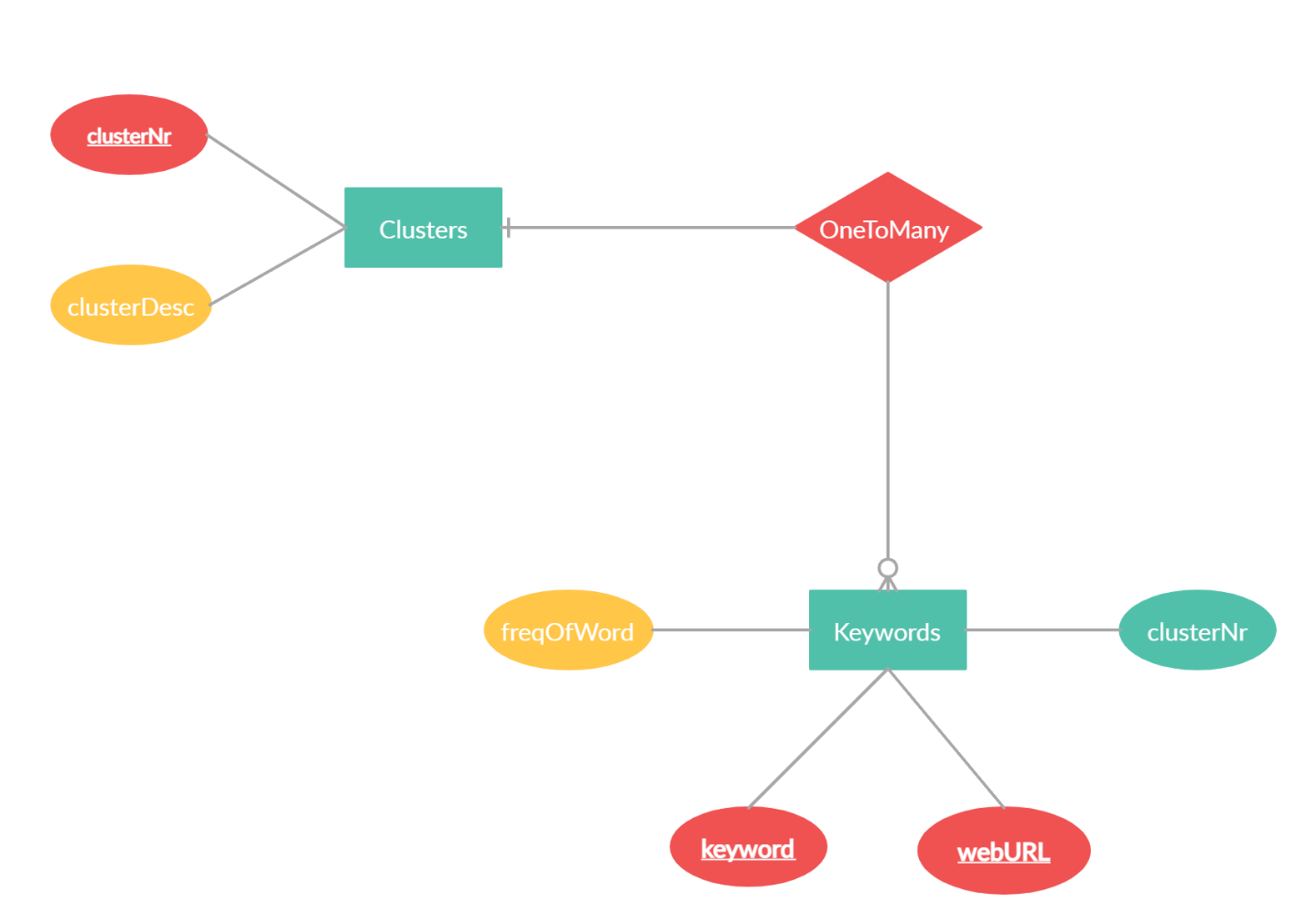
# 3.4 Narratives

|  |  |  |
| --- | --- | --- |
| Actor | Goal | Brief |
| Administrator | Start Web Spider | Administrator can trigger the application to start crawling and start generating stemmed keywords. Store them in clustered database. |
| User | Search from clustered keywords | User can apply search criteria and seek results from the system based on keywords stored in the database. |

# 3.5 ERD

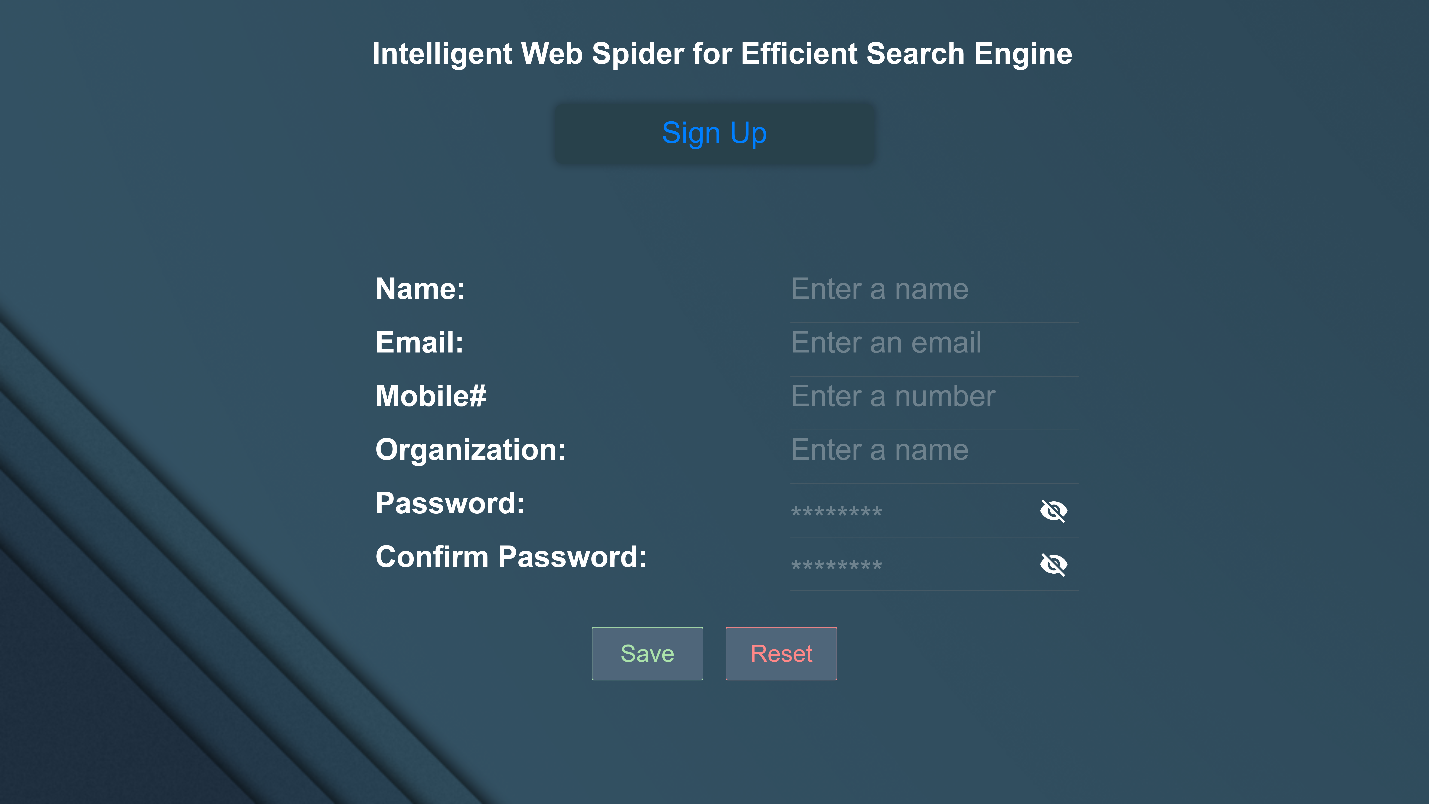




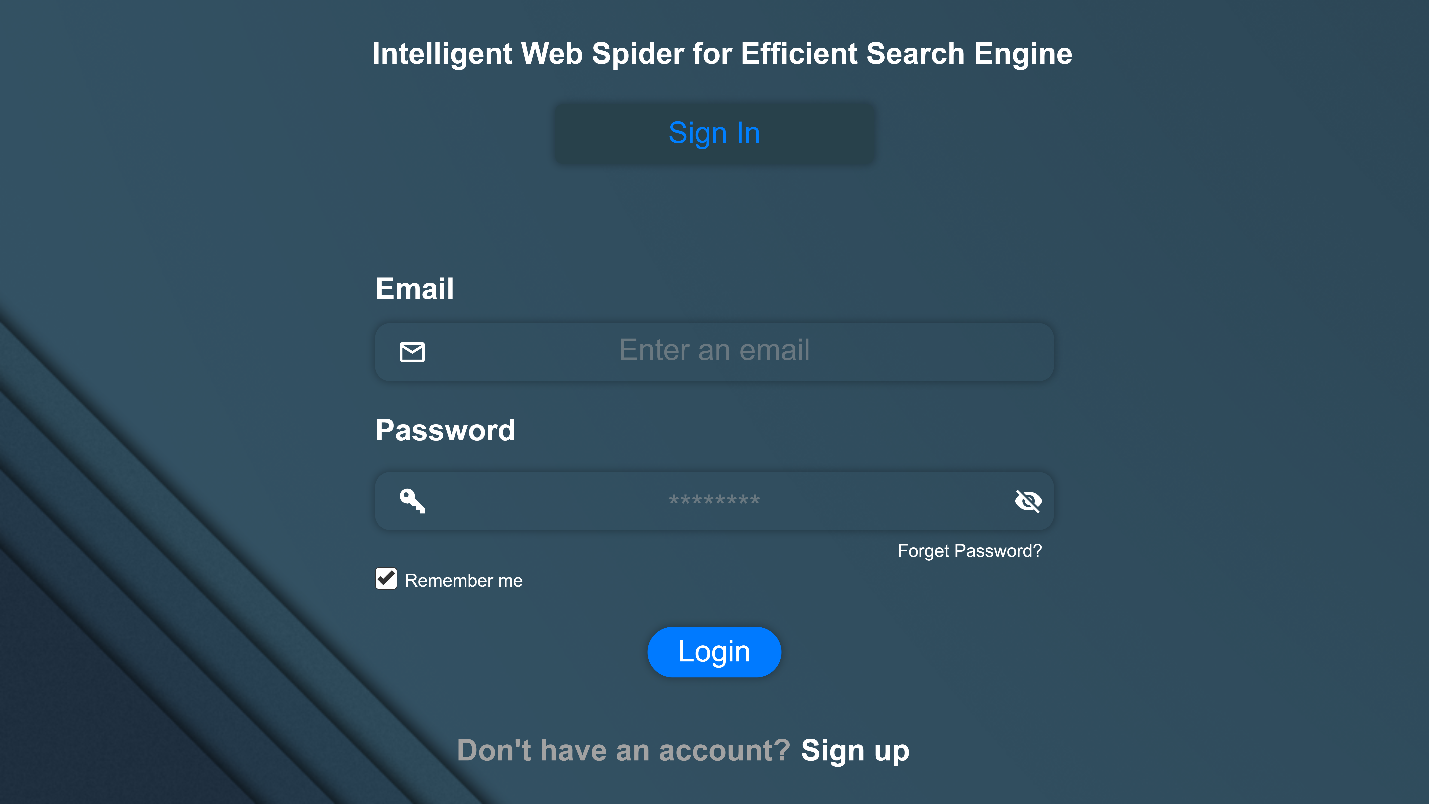


# 3.6 Prototype

* 1. **Sign up Screen**



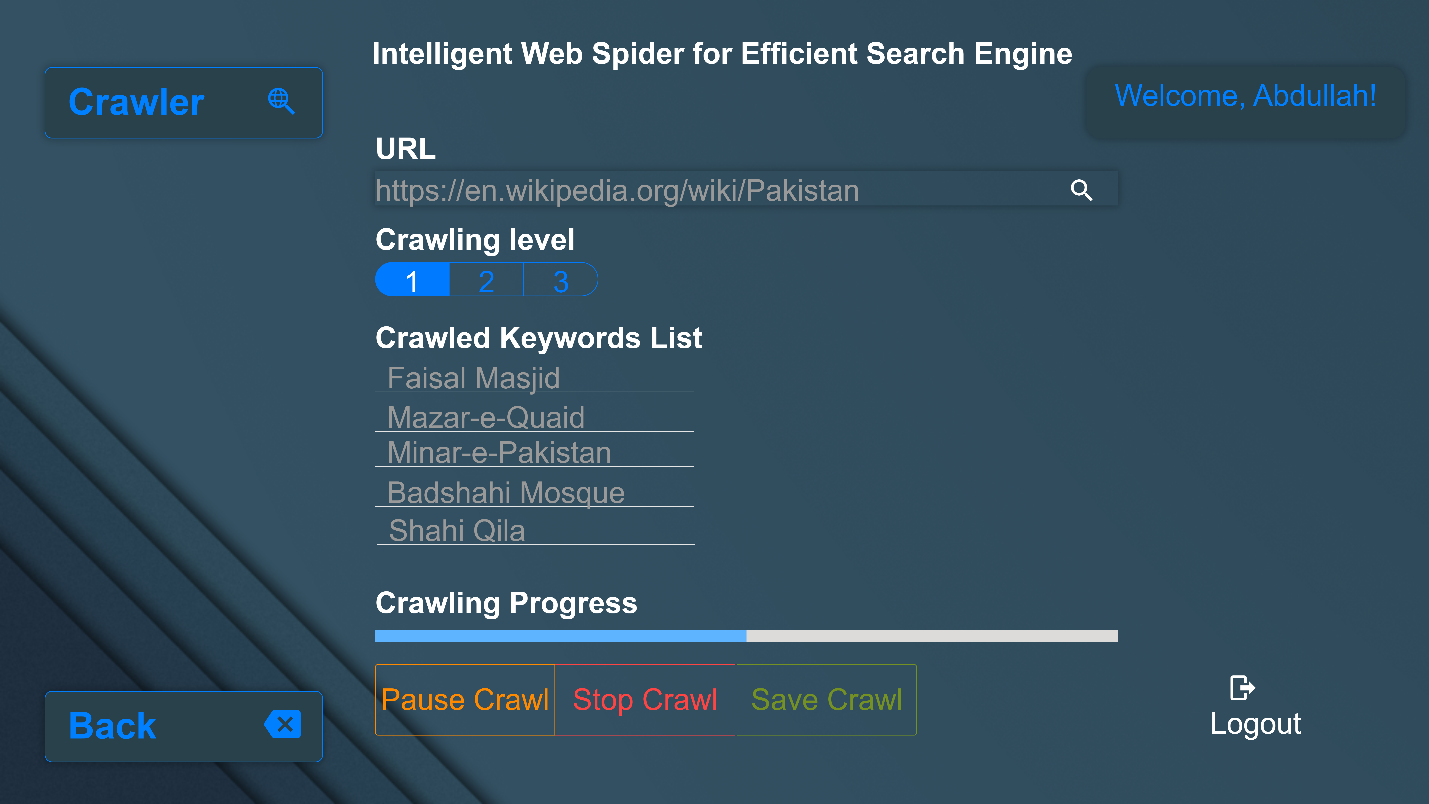
* 1. **Sign in Screen**



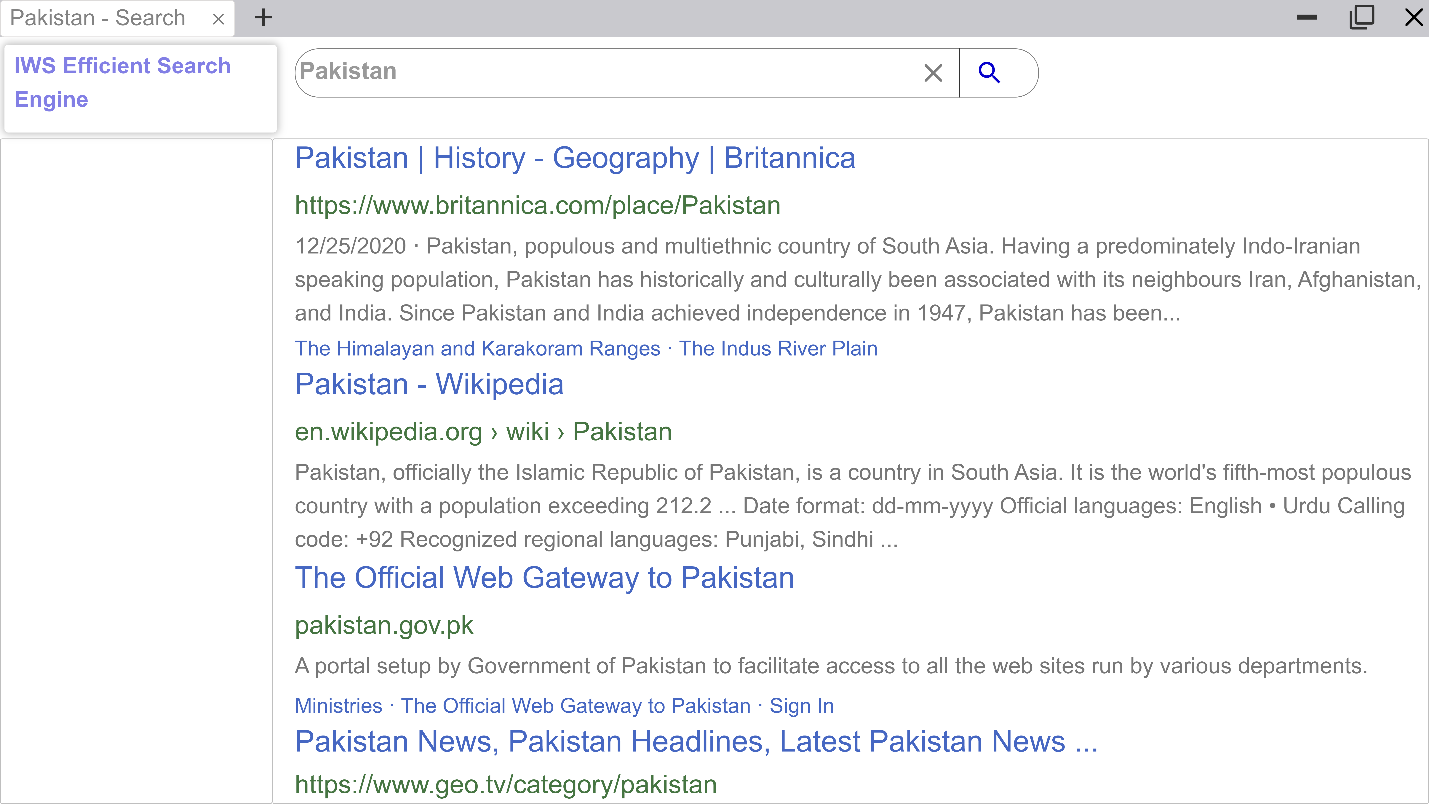
* 1. **Dashboard Screen**



* 1. **Crawler Screen**



* 1. **Search Engine Screen**



# 3.7 Annexure–B: Proposed Budget

Please use the embedded Excel Worksheet for providing budget details.

Double click the icon to open the worksheet.



# 3.8 Annexure–C: Business Canvas Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Problem***:   * Usually, Crawlers generate keywords in huge numbers and they do not provide the flexibility to define level of depth for searching or specific domain level searching / keyword generation * Efficient storage optimization of generated keywords is a big challenge * Crawlers are costly tools to be used for research purposes and generally inaccessible | ***Solution***     * Collect keywords from any website around the world * Apply linguistic algorithm to filter / refine and shrink keywords * Clustered based keyword storage for efficient storage and fast retrieval | ***Unique Value Proposition***  Generate keywords from websites and apply linguistics algorithm to purify them and simplify their storage to conserve space for quick and efficient searching. | **Unfair Advantage**  As our project covers end to end cycle used by search engines including its own crawler, keyword generation, their storage in our defined data clusters and a web based search engine component that search from our own generated keywords. This whole coverage gives us an additional advantage over other solutions | **Customer Segment**   * Defense * Security * Safety * Business * Marketing * Finance |
| ***Key Metrics***    Fast searching from keywords generated from our crawler and stored after applying linguistic algorithm that simplifies those keywords. | ***Channels***   * Facebook * WhatsApp * YouTube * In-person Marketing |
| ***Cost Structure***   * Technical / HR = 2.64 M * Support Staff = 0.34 M * Equipment = 0.38 M * Miscellaneous = 0.1 M * Contingency = 0.05 M * Institutional Overheads = 0.12 M | | | ***Revenue Stream***   * License Fee * Support and Services Cost * Customization Charges * Digital Ads | |