VIRTUAL COOK

(Project Proposal)

➢ Project Code

<Project code assigned by the Project Office>

➢ Project Advisor

Raja Fahad Bin Asif
(Web Developer, CEO of Techndevs)

➢ Project Manager

Mr.Fahad Maqbool

➢ Submission Date

September 23, 2019

➢ Project Team

S#	Roll Number	Name	Email	
1	Mansha Qarib	BCSF16E008	mansha.qarib777@gmail.com	**Team Member**
2	M. Furqan	BCSF16E018	chmfurqan@hotmail.com	**Team Member**
3	Zafar Abbas	BCSF16E006	rajazafarabbas1997@gmail.com	**Team Member**

Table of Contents

1. Abstract

A web crawler may be a piece of code that travels the web and collects knowledge from numerous sites, conjointly referred to as web scraping. Some web crawlers area unit autonomous and need no directions once started. This project can specialise in a user driven web crawler wherever user input can direct wherever the crawler goes and the way the collected data is analyzed.

Web scraping replaces the requirement for manual data entry and a lot of simply reveals trends among data collected. It can also aggregate info from multiple sources into one central location. whereas this application provides 3 specific samples of web crawling/scraping, it may be simply altered to higher suit further markets and/or desires.

2. Background and Justification

- **Background**

As the Internet and the web continue to expand, it can become difficult to access webpages without knowing beforehand the address of the page. This is where search engines come in. Search engines use a process called web crawling, which is an algorithm designed to scan or crawl through a collection of websites, which is indexed and searched . This algorithm has three main components; a webpage is fetched, it is parsed to extract all linked URLs

- **Justification**

A web crawler is created to find and gather complete or partial content from public websites, and therefore the data is provided to you in an simply manageable format. the data is hold on stored program or database, integrated with an in-house system or tailored to the other target. There are multiple ways in which to access the data you gathered. It is as straightforward as receiving a regular e-mail message with a .csv file or setting up search pages or an online app. you'll also add functionality to type the content, like pulling data from a selected timeframe, by certain keywords or no matter you wish.

3. Project Methodology

➢ **Design**

- **Design the User Interface**

We will design a user interface that is easy to use. First we design the user signup and signin page. After that we will design admin dashboard and after that we will design frontend.

- **Design the Database**

We will design an entity-relationship schema (ER diagram) for our crawler database. The ER diagram will help us design efficient and stable database. We will design different tables for media and text. So that we will store text and media in separate tables.

- **Design Data Scraping Techniques**

We will design algorithms to periodically scrape different websites and collect data for our database.

➢ Implementation

- **Build User Interface**

First we build the user signup and signin page. After that we will build admin dashboard and after that we will build frontend.

- **Build Database**

Based on our ER diagram, we will use My SQL to build our crawler database.

- **Develop Crawler**

Based on our design, we will build web scrapper. We store data in database from where user will download data. Following are the functionalities.

- Removing a URL from the URL list.
- Determining the IP address of its host name.

- Downloading of the related documents.
- Extracting any links available in the documents.

➢ **Testing**

- **Test Web Crawler**

In this we will test a web crawler. We will insert urls and check whether it is working fine. It is extracting data from url and storing it is database or not.

4. Project Scope

Web is growing at a very high scale everyday. For the crawlers to achieve high coverage and great performance, it needs to give very high throughput. This led to the creation of a large number of engineering based problems. To solve these problems, the comppanies need to employ a huge number of systems that may count to thousand and almost a dozen of high speed network links.

5. High level Project Plan

This section should contain a high level project plan with important milestones and their submission dates. It should contain:

1. High level activities for the proposed system
2. Time allocated to each activity
3. Resources that will be assigned to complete each activity

(Project plan should be developed in MS project) >

6. References

- C. C. Aggarwal, P. S. Yu. Intelligent , and F. Al-Garawi crawling on the World Wide Web. In WWW10, Hong Kong, May 2001.

- B. Amento, W. Hill and L. Terveen . Does “authority” mean quality? Predicting expert quality ratings of web documents. In Proc. 23rd Annual Intl. ACM SIGIR Conf.

on Research and Development in Information Retrieval,
2000.

- Arasu, J. Cho, A. Paepcke, H. Garcia-Molina and S. Raghavan. Searching the Web. ACM Transactions on Internet Technology, 1(1), 2001.