High Level Design (HLD)

Website Scraper

Revision Number: 1.0

Last date of revision: 11 May 2022

Sunny Sahsi

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 11/04/2022 | 0.1 | Added Introduction and General description | Sunny Sahsi |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Table of Content**

Document Version Control ……………………………………………………………. 2

Abstract ………………………………………………………………………………………… 4

Introduction ………………………………………………………………………………….. 5

1. Why this High-Level Document? ……………………………………………. 5

2. Scope …………………………….………………………………………………………. 5

3. Definition ………………………………………………………………………………. 5

General Description ………………………………………………………………………. 6

1. Product Perspective ……………………………………………………….………. 6

2. Problem Statement ………………………………………………………..………. 6

3. Problem Solution ……………………………………………………………………. 6

4. Proposed Methodology …………………………………………………….……. 6

5. Further Improvements ……………………………………………………………. 6

6. Data Required ……………………………………………………...…………………. 6

7. Tools Used ………………………………………………………………………………. 7

8. Constraints …………………………………………………………………..…………. 7

9. Assumptions ……………………………………………………………………………. 7

Design Details ……………………………………………………………..…………………. 8

1. Process Workflow ………………………………………………………..…………. 8

2. Error Handling ………………………………………………………………..………. 9

Performance …………………………………………………………………………….……. 10

1. Reusability ……………………………………………………………..………………. 10

2. Application compatibility ………………………………………..………………. 10

3. Resources Utilization ………………………………………………………….……. 10

4. Deployment ……………………………………………………………………..…..…. 10

Conclusion ………………………………………………………………………….……..……. 11

Reference ………………………………………………………………………………..………. 12

**Abstract**

Pollution is major problem in 21st century. There is a lot of work going on to reduce the pollution. Air pollution is important issue in all over the world. In India it become worst in winter season in cities like Delhi and Gurugram. Air quality for a human being and all the living species is a vital part of life. Variation of air quality creates a high impact on human life; continuous monitoring and evaluation and investigation of AIR quality are critical. By using various factors air quality is calculated i.e., Air quality Index (AQI) mathematically. But it’s also possible to calculate AQI by using machine learning technique.

**Introduction**

**1. Why this High-Level Design Document?**

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all the design aspects and define them in detail
* Describe the user interface being implemented
* Describe the hardware and software interfaces
* Describe the performance requirements
* Include design features and the architecture of the project

**2. Scope**

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

**3. Definition**

The terms used in the projects are:

* TAS – Twitter Advance Search

**General Description**

**1. Product Perspective**

Website scraping is a huge topic. For scraping useful data then each website takes different codes and functions.

**2. Problem Statement**

To create the scraping-based tool to scrape the data from website then specific code for specific websites.

**3. Problem Solution**

Develop the application to scrape the data then first of all scrap the json text and then categories the data.

**4. Further Improvement**

The project can be extended by more code of blocks or functions to collect data from different websites.

**5. Data Required**

For Twitter – username, number of tweets, TAS query.

For Flipkart – Price Range and Number of pages

**6. Tools Used**

* Python programming language and frameworks such as snscrape, Pandas etc.
* PyCharm and Visual Studio Code is used as IDE.
* GitHub is used as version control system.

**7. Constraints**

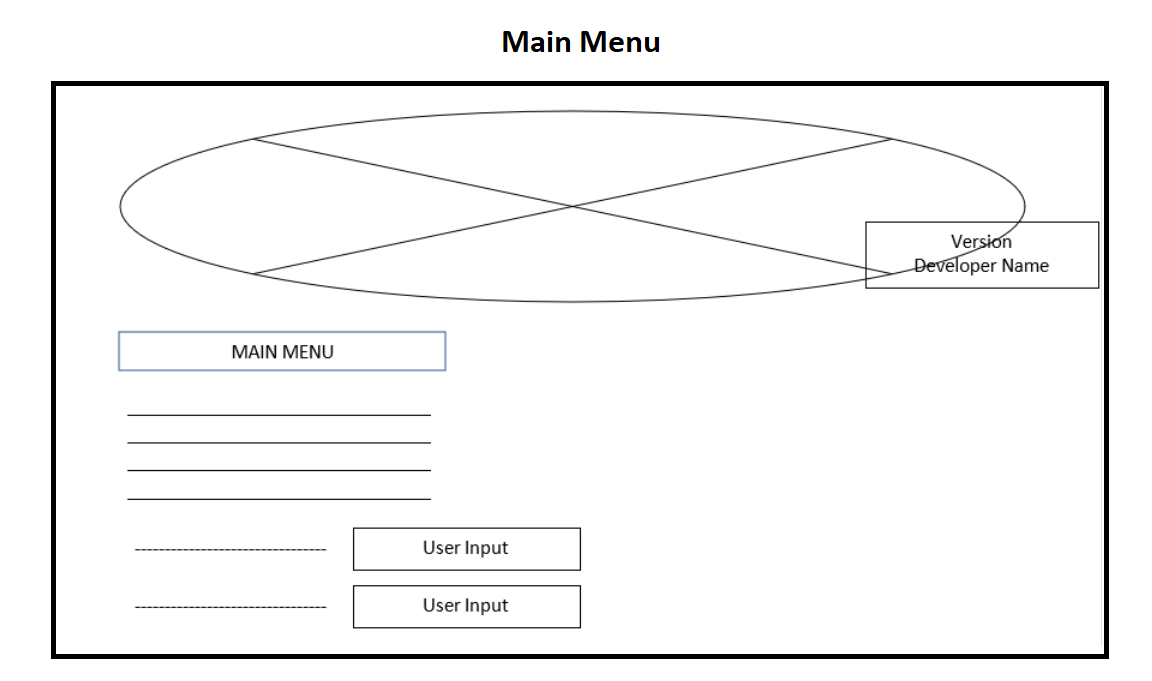
This python tool should be user friendly. Different functions to be created for different websites.

**8. Assumptions**

The main objective of the project is to scrape the useful data and used for analysis and social media reputation.

**Design Details**

**1. Process Workflow**



**Deployment Process**

* git clone <https://github.com/sahsisunny/WebsiteScrapperProjectVIEH>
* cd WebsiteScrapperProjectVIEH
* pip install –r requirements.txt
* python WebScrapper.py

**2. Error Handling**

Initially we got error when scrape the data from Flipkart then use class name the solve the problem.

**Performance**

**1. Reusability**

Data scraped from Twitter and Flipkart is change day to day and we scrape data for updated data.

**2. Application compatibility**

Since we are using python and it is compatible with any platform, we follow Application compatibility.

**3. Resource utilization**

At the initial stage, we were using high space to scrape the data. Once the data is scraped our system only needs at least of 2GB RAM and 1 GB of storage to run the application smoothly. system uses less than 10% of the processing power.

**4. Deployment**

The code is available on GitHub.

**Conclusion**

This project purpose is scraping the useful data. For analysis and Social Media Reputation and for personal used like Choose the Smart phone in short time period.

**References**

1. <https://github.com/JustAnotherArchivist/snscrape>
2. <https://pandas.pydata.org/docs/>
3. <https://twitter.com/search-advanced?lang=en>
4. <https://www.analyticsvidhya.com/blog/2021/10/scraping-flipkart-data-using-python-web-scraping-tutorial/>
5. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
6. <https://docs.python.org/3/tutorial/>