“ Content Aggregation using Web Scraping ”

Mini Project Report

Submitted in partial fulfillment of the requirements of the subject Minor Project

by

Anirudh Bhattacharya
Saumya Shah
Mink Shethia

Supervisor

Prof. Mamta Borle

Department of Computer Engineering

K.J. Somaiya Institute of Engineering and Information Technology

Ayurvihar, Sion Mumbai-400022

2021-22

This is to certify that the project entitled ***“Content Aggregation using Web Scraping”*** *is a bona fide work of Anirudh Bhattacharya, Saumya Shah and Mink Shethia submitted as mini project in the subject of* ***Minor Project*** *in* ***“Computer Engineering”.***

Prof. Mamta Borle

(Project Guide)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. we also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. we understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Anirudh Bhattacharya ____________

Saumya Shah ____________

Mink Shethia ____________

Date:

ACKNOWLEDGEMENT

Before presenting out our seminar work entitled "**Content Aggregation using Web Scraping**" we would like to convey our sincere thanks to many people who guided us throughout the course for this seminar work. First, we would like to express our sincere thanks to our beloved Principal **DR. SURESH K. UKARANDE** for providing various facilities to carry out this report.

We would like to express our sincere thanks to **Prof. Mamta Borle** for her guidance, encouragement, co-operation and suggestions given to us at progressing stages of report.

Finally, we would like to thank our **H.O.D. Dr. SARITA AMBADEKAR** and all teaching, non-teaching staff of the college and friends for their moral support rendered during the course of the report work and for their direct and indirect involvement in the completion of our report work, which made our endeavour fruitful.

Anirudh Bhattacharya

Saumya Shah

Mink Shethia

ABSTRACT

While conducting research, it is often cumbersome to look through an entire article to find all relevant information which can be summarised in just a few lines. Even whilst creating social media posts, one needs to summarise many words in the best way possible.

Thus, our project aims to decrease the word count of articles and thus make it easier to understand for those in a hurry

This system will be most useful to content creators and social media managers. With the help of this web application, they can save hours and quite possibly even days of work.

INDEX

CHAPTER 1
INTRODUCTION

1.1 Introduction:

When performing research, it might be time-consuming to read a whole article in order to uncover all necessary information that can be summarised in a few words. Even while writing social media postings, one must summarise a large number of words in the most effective way feasible.

As a result, our effort attempts to reduce the word count of articles, making them easier to comprehend for individuals in a hurry.

Content authors and social media administrators will benefit the most from this method. They will be able to save hours, if not days, of labor by using this online application.

1.2 Problem Introduction:

Content Aggregator does the following functions:

1. Find important information in the form of articles from reputable sources.
2. Aggregate the information such that it’s available in a concise manner.
3. Present said information to the user on the screen

CHAPTER 2

REQUIREMENT SPECIFICATION

2.1 INTRODUCTION:

To be used efficiently, all computer software needs certain hardware components or the other software resources to be present on a computer. These pre-requisites are known as(computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements.

2.2 HARDWARE REQUIREMENTS:

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware. A hardware requirements list is often accompanied by a hardware compatibility list (HCL), especially in case of operating systems. An HCL lists tested, compatibility and sometimes incompatible hardware devices for a particular operating system or application. The following sub-sections discuss the various aspects of hardware requirements.

HARDWARE REQUIREMENTS FOR PRESENT PROJECT:

PROCESSOR : Intel Pentium dual core or above.

RAM : 2 GB

HARD DISK : 160 GB

2.3 SOFTWARE REQUIREMENTS:

Software Requirements deal with defining software resource requirements and pre-requisites that need to be installed on a computer to provide optimal functioning of an application. These requirements or pre-requisites are generally not included in the software installation package and need to be installed separately before the software is installed.

SOFTWARE REQUIREMENTS FOR OUR PROJECT:

OPERATING SYSTEM : Windows XP and above, Ubuntu v12.04 and above.

SOFTWARE INSTALLATIONS NEEDED : Django~=3.2.7, requests~=2.26.0, bs4~=0.0.1, beautifulsoup4~=4.10.0, requests-html~=0.10.0

CHAPTER 3

ANALYSIS

3.1 PROPOSED SYSTEM:

This project can be used by content creators and social media managers to make their jobs easier.

This can also be used to summarise news articles for those in a hurry,

For example, one can get a concise report of the Union Budget 2022 instead of reading all details.

Thus, it’ll be very effective in reducing work hours.

3.2 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are:

3.2.1 Economic Feasibility

The project is aimed to be used mainly by Law Enforcement Agencies and Car Manufacturers and will require an integrated webcam or an external camera to be connected with the device on which the system will be running.

3.2.2 Technical Feasibility

The technical feasibility assessment meets with the expected needs of the proposed system. It has evaluated that hardware and software meets the need of the proposed system. The assessment based on the project of online testing consist of an interactive interface between student and teachers reveals the following outline design of system requirements:

->Python 3.0

->Django~=3.2.7

->requests~=2.26.0

->bs4~=0.0.1

->beautifulsoup4~=4.10.0

->requests-html~=0.10.0

To deal with requirements to handle completion of the project we are having strong resource of knowledge over the required technologies among our group members. Furthermore, these technologies are being thought in depth in WT tutorials to overcome any of the difficulties. Also the technologies required are economically and legally feasible for implementation purpose.

3.2.3 Operational Feasibility

Driver Drowsiness Detection system to detect whether a vehicle's operator is physically capable of driving or not. Online pictures are also provided so that practically the users will understand what is the actual product.

3.3 SOFTWARE SPECIFICATION

DJANGO:
What is Django?
Django is a Python framework that makes it easier to create web sites using Python.

Django takes care of the difficult stuff so that you can concentrate on building your web applications.

Django emphasizes reusability of components, also refereed to as DRY (Don't Repeat Yourself), and comes with ready-to-use features like login system, database connection and CRUD operations (Create Read Update Delete).

How does Django Work?

Django follows the MVT design pattern (Model View Template).

Model - The data you want to present, usually data from a database.

View - A request handler that returns the relevant template and content - based on the request from the user.

Template - A text file (like an HTML file) containing the layout of the web page, with logic on how to display the data.

So, What is Going On?

When you have installed Django and created you first Django web application, and the browser requests the URL, this is basically what happens:
Django receives the URL, checks the urls.py file, and calls the view that matches the URL.
The view, located in views.py, checks for relevant models.
The models are imported from the modals.py file.

The view then sends the data to a specified template in the template folder.

The template contains HTML and Django tags, and with the data it returns finished HTML content back to the browser.
Django can do a lot more than this, but this is basically what you will learn in this tutorial, and are the basic steps in a simple web application made with Django.

REQUESTS:
Requests allows you to send HTTP/1.1 requests extremely easily. There’s no need to manually add query strings to your URLs, or to form-encode your POST data. Keep-alive and HTTP connection pooling are 100% automatic, thanks to urllib3.
Django uses request and response objects to pass state through the system.

When a page is requested, Django creates an HttpRequest object that contains metadata about the request. Then Django loads the appropriate view, passing the HttpRequest as the first argument to the view function. Each view is responsible for returning an HttpResponse object.

This document explains the APIs for HttpRequest and HttpResponse objects, which are defined in the django.http module.

BEAUTIFUL SOUP:
Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

This document covers Beautiful Soup version 4.11.0. The examples in this documentation were written for Python 3.8.

You might be looking for the documentation for Beautiful Soup 3. If so, you should know that Beautiful Soup 3 is no longer being developed and that all support for it was dropped on December 31, 2020. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see Porting code to BS4.

CHAPTER 4

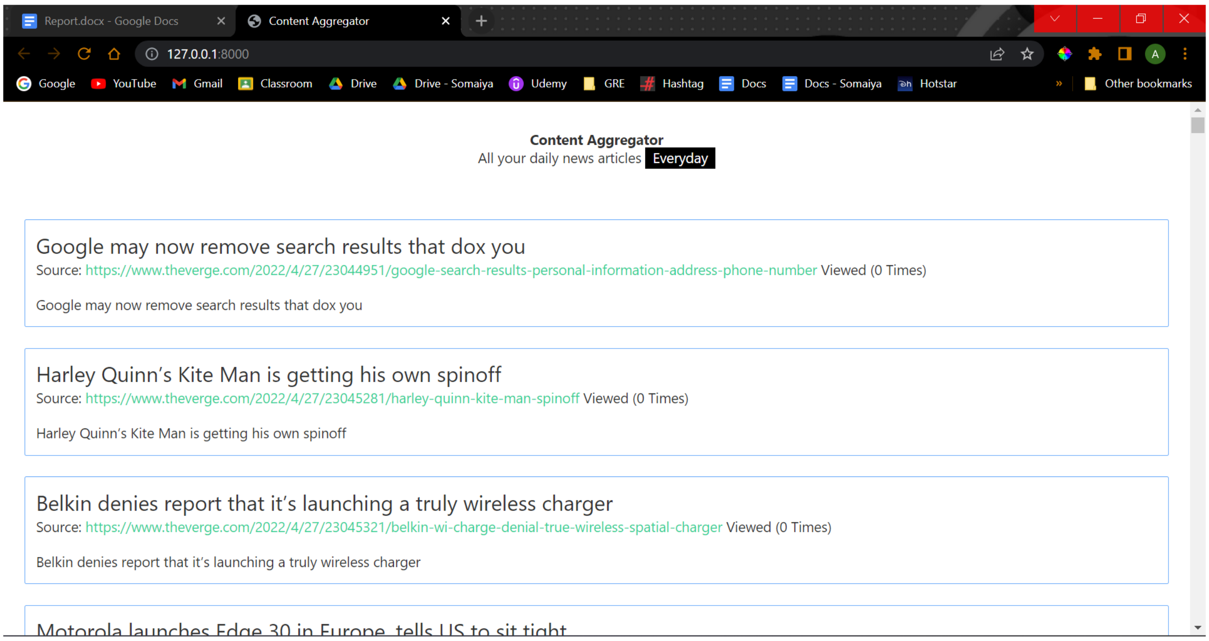
IMPLEMENTATION

4.1 Introduction:

We plan to Implement our system in a slow and gradual manner. We fill first test the detection of the eyes then the detection of drowsiness in the eyes.

The implementation stage involves careful planning, investigation of the existing system and it’s constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

GUI Screenshots



Content Aggregator
All your daily news articles Everyday

Google may now remove search results that dox you
Source: https://www.theverge.com/2022/4/27/23044951/google-search-results-personal-information-address-phone-number Viewed (0 Times)

Google may now remove search results that dox you

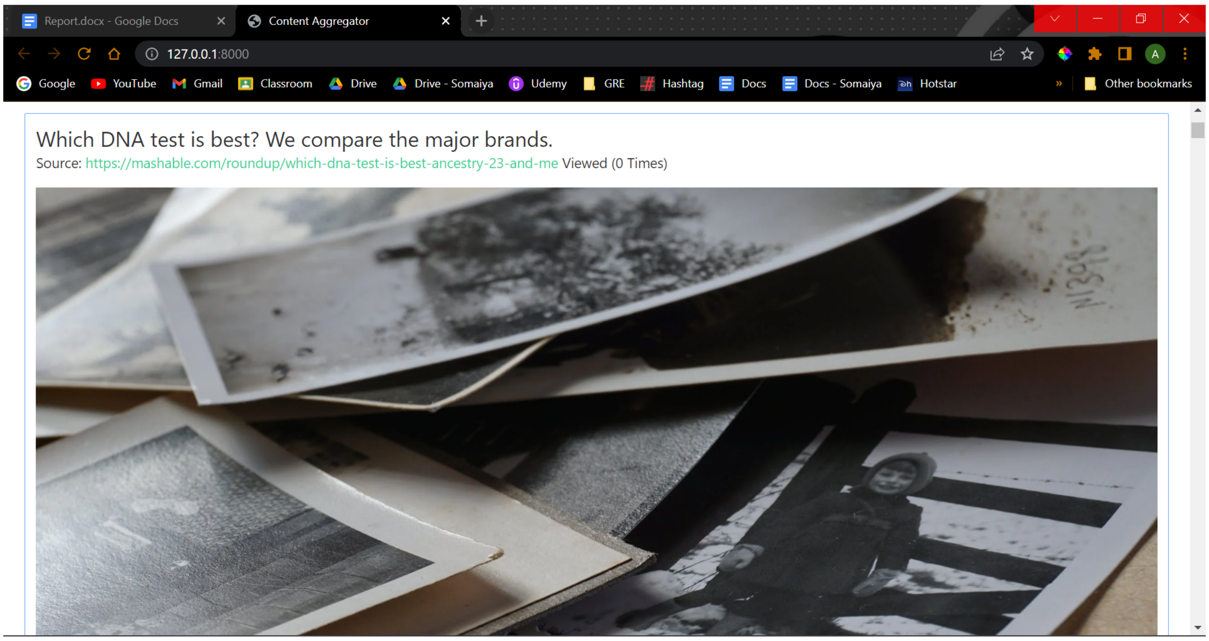
Harley Quinn’s Kite Man is getting his own spinoff
Source: https://www.theverge.com/2022/4/27/23045281/harley-quinn-kite-man-spinoff Viewed (0 Times)

Harley Quinn’s Kite Man is getting his own spinoff

Belkin denies report that it’s launching a truly wireless charger
Source: https://www.theverge.com/2022/4/27/23045321/belkin-wi-charge-denial-true-wireless-spatial-charger Viewed (0 Times)

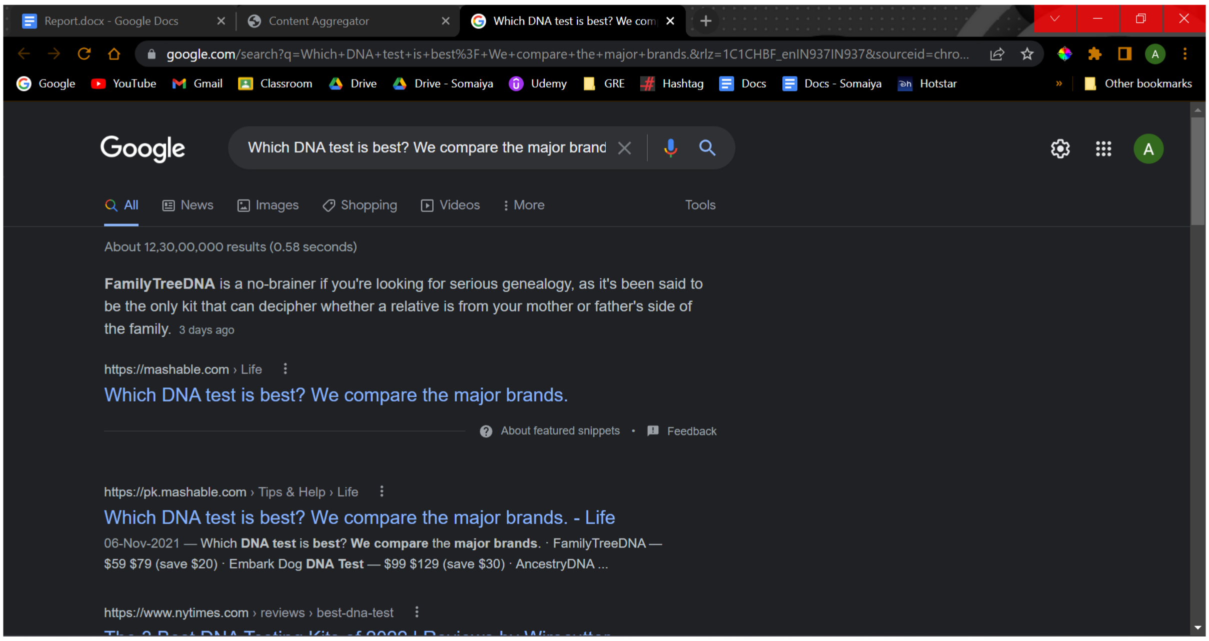
Belkin denies report that it’s launching a truly wireless charger

Motorola launches Edge 30 in Europe, tells US to sit tight



Which DNA test is best? We compare the major brands.

Source: https://mashable.com/roundup/which-dna-test-is-best-ancestry-23-and-me Viewed (0 Times)


Report.docx - Google Docs
Content Aggregator
Which DNA test is best? We com
google.com/search?q=Which+DNA+test+is+best%3F+We+compare+the+major+brands.&rlz=1C1CHBF_enIN937IN937&sourceid=chro...
Google
YouTube
Gmail
Classroom
Drive
Drive - Somaiya
Udemy
GRE
Hashtag
Docs
Docs - Somaiya
Hotstar
Other bookmarks
Google
Which DNA test is best? We compare the major brand
All
News
Images
Shopping
Videos
More
Tools

About 12,30,00,000 results (0.58 seconds)

FamilyTreeDNA is a no-brainer if you're looking for serious genealogy, as it's been said to be the only kit that can decipher whether a relative is from your mother or father's side of the family. 3 days ago

https://mashable.com › Life

Which DNA test is best? We compare the major brands.

About featured snippets • Feedback

https://pk.mashable.com › Tips & Help › Life

Which DNA test is best? We compare the major brands. - Life

06-Nov-2021 — Which **DNA test** is **best**? **We compare** the **major brands**. · FamilyTreeDNA — $59 $79 (save $20) · Embark Dog **DNA Test** — $99 $129 (save $30) · AncestryDNA ...

https://www.nytimes.com › reviews › best-dna-test

Functioning:

The web scraping library - beautiful soup - finds headlines from reliable news sources. The parser - requests - saves them to the database and Django renders the frontend (HTML, CSS) so that it is accessible.

To run the project, open terminal in the project folder and run the following commands:

python manage.py makemigrations

python manage.py migrate

```
python manage.py runserver
```

The first two commands update the database and the last command starts the server. The Django backend then works as the site is rendered.

CHAPTER 5

TESTING

5.1 INTRODUCTION TO SYSTEM TESTING:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

5.2 TYPES OF TESTING:

1. **Unit testing:**

   Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

2.Integration testing:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

3.Functional test:

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:
Valid Input : identified classes of valid input must be accepted.
Invalid Input : identified classes of invalid input must be rejected.
Functions : identified functions must be exercised.
Output : identified classes of application outputs must be exercised.
Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

4. System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

5. **White Box Testing:**

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

6. **Black Box Testing:**

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

7. **Unit Testing:**

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

8. **Integration Testing:**

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

9. **Acceptance Testing:**

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

5.3 Testing of Project

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

● All field entries must work properly.
● Pages must be activated from the identified link.
● The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Test Results:

All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER 6

CONCLUSION

Content providers and social media administrators can benefit from this initiative by making their tasks simpler.

For those in a hurry, this may also be used to summarise news stories.

Instead of reading all the information, one can acquire a succinct summary on the Union Budget 2022.

As a result, it will be extremely successful in reducing work hours.

CHAPTER

REFERENCES

We have referred the following websites:

1) Web Scraping And Data Acquisition Using Google Scholar
2) Data Analysis by Web Scraping using Python
3) Requests Documentation
4) Beautiful soup documentation
5) Django documentation