Visvesvaraya Technological University

"Jnana Sangama", Belagavi - 590018

WEB TECHNOLOGY LABORATORY WITH MINI PROJECT

REPORT ON

"Wiki Data Scraper"

Submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

ASHISH K AMAR	**1KS17CS013**
CHENNAKESHAVA N T	**1KS17CS019**

Under the guidance of

Mr. Aditya Pai H
Assistant Professor

Mr. Roopesh Kumar BN
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

K. S. Institute of Technology

#14, Raghuvanahalli, Kanakapura Road, Bengaluru - 560109

2020 - 2021

K. S. Institute of Technology

#14, Raghuvanahalli, Kanakapura Road, Bengaluru - 560109

Department of Computer Science & Engineering

CERTIFICATE

Certified that the project work entitled **"Wiki Data Scraper"** is a bonafide work carried out by:

ASHISH K AMAR	**1KS17CS013**
CHENNAKESHAVA N T	**1KS17CS019**

in partial fulfilment for VII semester B.E., **WEB TECHNOLOGY LABORATORY WITH MINI PROJECT (17CSL77)** in the branch of Computer Science and Engineering prescribed by **Visvesvaraya Technological University, Belagavi** during the period of February 2020 to June 2020. It is certified that all the corrections and suggestions indicated for internal assessment have been incorporated in the report deposited in the department library. The Project Report has been approved as it satisfies the academic requirements in report of project work prescribed for the Bachelor of Engineering degree.

.....	.....	.....
Signature of the Guide	**Signature of the HOD**	**Signature of the Principal & CEO**
[Mr. Aditya Pai H]	[Dr. Rekha B. Venkatapur]	[Dr. K.V.A Balaji]

Name of the Student:	**Ashish K Amar**	**Chennakeshava N T**
USN:	**1KS17CS013**	**1KS17CS019**

Signature of the Student:

Signature of the Internal: _____ **Signature of the External:** _____

ACKNOWLEDGEMENT

The successful presentation of the **WEB TECHNOLOGY LABORATORY WITH MINI PROJECT** would be incomplete without the mention of the people who made it possible and whose constant guidance crowned my effort with success.

We take this opportunity to express our sincere gratitude to our college **K.S. Institute of Technology,** Bengaluru for providing the environment to work on our project.

We would like to express our gratitude to our **MANAGEMENT,** K.S. Institute of Technology, Bengaluru, for providing a very good infrastructure and all the kindness forwarded to us in carrying out this project work in college.

We would like to express our gratitude to **Dr. K.V.A Balaji, CEO and Principal,** K.S. Institute of Technology, Bengaluru, for his valuable guidance.

We like to extend our gratitude to **Dr. Rekha.B.Venkatapur**, **Professor and Head**, Department of Computer Science & Engineering, for providing a very good facilities and all the support forwarded to us in carrying out this project work successfully.

We also like to thank our Project Guides', **Mr. Roopesh Kumar B.N., Asst. Professor, and Mr. Aditya Pai H, Asst. Professor, Department of Computer Science & Engineering** for their help and support provided to carry out the project and complete it successfully.

We are also thankful to the teaching and non-teaching staff of Computer Science & Engineering, KSIT for helping us in completing this project work.

ASHISH K AMAR

CHENNAKESHAVA N T

ABSTRACT

In recent years, The Field of Data Science is seeing growth rapidly. Data Science is listed as the Topmost skill of 2020 by many renowned Tech giants. The world of programmers is increasing exponentially. Data is everywhere. Managing this immense data and using it in a constructive way rests in the hands of a Data Scientist.

Data scientists are entirely dependent on Data sets for their work. Data Scientists spends most of their time managing, organizing and analysing data. Many data scientist developers are in need of datasets to work on. A good source of such data is “Wikipedia”.

In this project, we will be using a Web Scraper to scrape data from wikipedia pages and convert them into unstructured datasets. A URL will be taken as input to the scraper, Using the URL the scraper automatically scrapes data, converts them to csv format and makes downloadable copies of data which the data scientist can work upon.

Keywords*: Scraper, URL, CSV, Wikipedia, Data Science, Analysis, Unstructured.*

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Overview of the project

The field of data science has been listed as the top most skill in 2020. The world of programmers is increasing and there are many data scientist developers who are in need of datasets to work on. The Wikipedia pages contain loads of data from all over the world and our application facilitates the need for fetching that data for programmers so that it will be easy to work with. A brief of how it works is that it uses python for back-end and HTML for front-end. There will be a database to collect the input queries provided by the user. The user has to enter a URL and if it contains any form of data collection, that will be scraped and converted to a csv file which can be used by the developer to work with. The generated output along with the URL will also be stored in the database.

1.1.1 Web Application

This Component is the Interactive module which users use in order to provide the input to the web scraper and get the scraped results generated by the web scraper. The web application is dependent on the scraper for its operations.

1.1.2 Web Scraper

Scraper is a tool that uses the user input URL and processes it to fetch content from the URL website and scrapes the data to Web Application. The Scraper is dependent on the web application for its operation such as fetching the input url and to provide the users the desired output.

1.2 Purpose of this Project

Most common problem for data scientists is collecting data, building datasets. Since Wikipedia has collaborative data from the public it would be a lot of help for the data scientists and programmers if they had a way to collect data from here. This is where our application comes in, it creates an unstructured database from the page and gives it to the programmer for further modifications.This is an application to build datasets from Wikipedia pages.

Chapter 2

DATA FLOW DIAGRAM

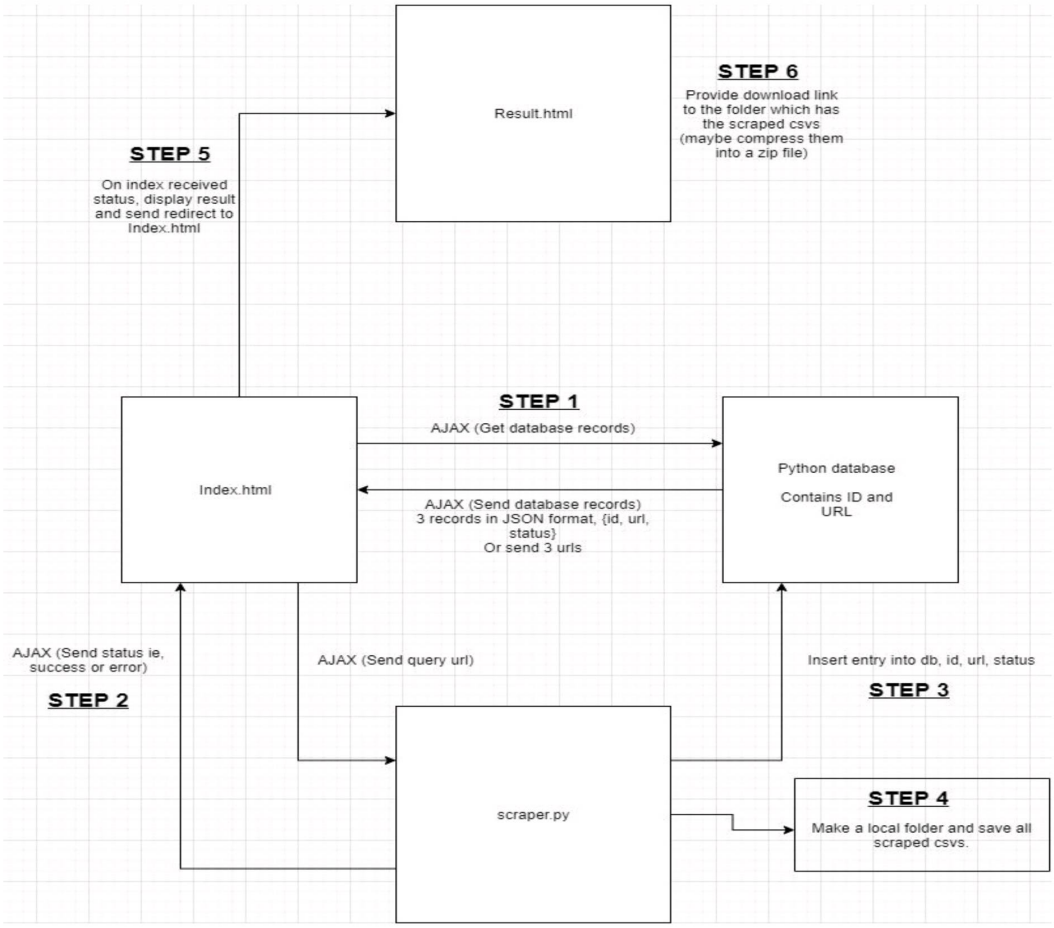

fig:- 2.1: Work-Flow Diagram

The above figure shows the data flow diagram for the wiki data scraper. The whole project is divided into 4 basic modules.

They are:

1. Index – Which takes in the URL as an input from the user.

2. Scraper – Scrapes the required information.

3. Database – Which stores the URL and the status of the scraping operation.

4. Result – Which provides a download link to the scraped file.

Chapter 3

REQUIREMENT SPECIFICATION

A Requirement Specification is basically an organization's understanding of a customer or potential client's system requirements and dependencies at a particular point prior to any actual design or development work. The information gathered gives a brief description of the services that the system should provide and also the constraints under which the system should operate. Generally, the Software Requirement completely describes what the proposed software should do without describing how the software will do it. It's a two-way insurance policy that assures that both the client and the organization understand the other's requirements from that perspective at a given point in time.

3.1 Software Requirements

We use the following software requirements:

- Operating system : Windows 7 and above/macOS/ Linux
- Coding language : PHP, Python, JavaScript, HTML, CSS and Bootstrap
- Editor and Tools : VS Code, Xampp
- Database : MySQL (Database)

3.2 Hardware Requirements

- Processor: Intel i3 2nd generation and above
- RAM: 4GB
- Hard disk: 50GB
- Input devices: keyboard, mouse
- Internet Connection

Chapter 4

TESTING

4.1 Testing Plan

First inspection is performed followed by white box testing which is applied by the programmer. The programmer may plan it either as a unit test or an integrated test.

4.2 Testing Strategy & Methods

If it is white box testing then all statements are checked whether they are logically correct.
Two methods are used: Boundary value checking. Equivalence class partitioning

4.3 Test Cases

Test Case 1: Testing scraper module.

Expected response: Downloadable Datasets
Error message should be displayed if entered URL could be used to fetch datasets
System response: "Error: Scrape Another URL"

Test Case 2: Testing Web Application.

Expected Response;- Launch of web portal
Error message will be automatically shown by the web browser

Chapter 5

SNAPSHOTS

1. Index Page:

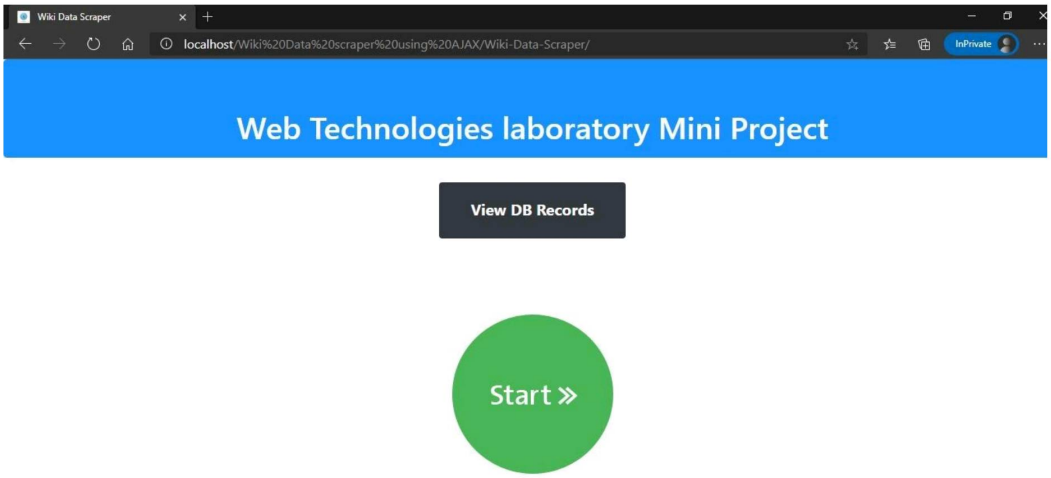

fig:- 5.1: Index Page

- The above figure shows the Index page of the web application.
- This is the page that loads when the web app is initialized.
- This page sends a parameter to the main page.

2. Main Page:

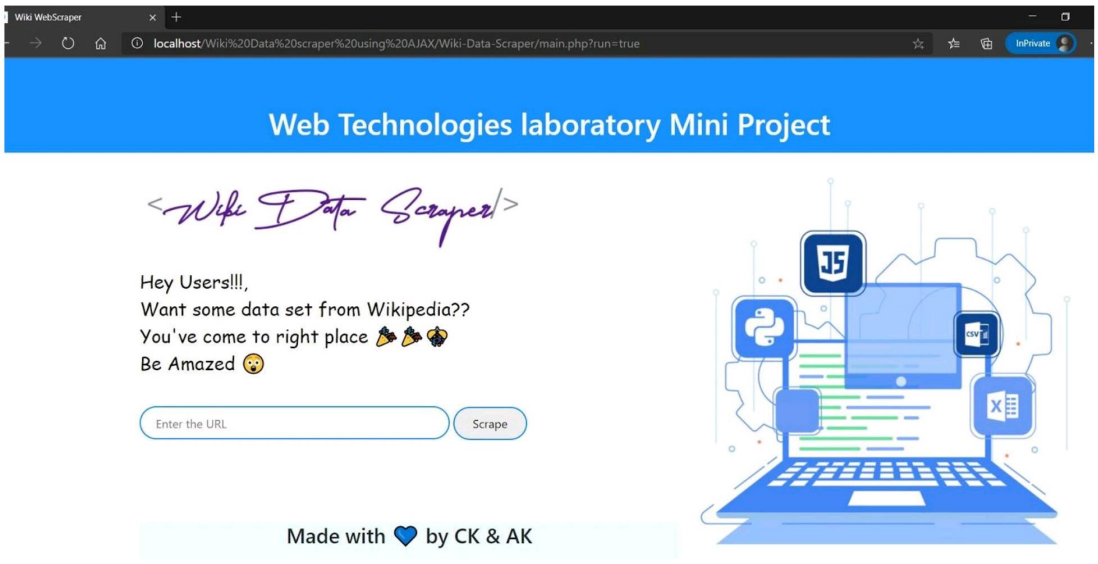

fig:- 5.2: Main Page

- The above figure shows the main page of the web application.
- This is the page that loads when it receives the URL parameter from the Index page.
- This page executes a shell script [1] which invokes a python script [2] to refresh the database records.
- This page also accepts the URL input from the user and passes it to the scraper.
- The scraper sends a status to this page after all it's operations and this page redirects to the result page on receiving the status.
- The status could be either positive or negative.
- Positive or success if the scraper was able to scrape some data. Negative or error if the scraper was not able to capture some data from the given URL.

3. Result Page:

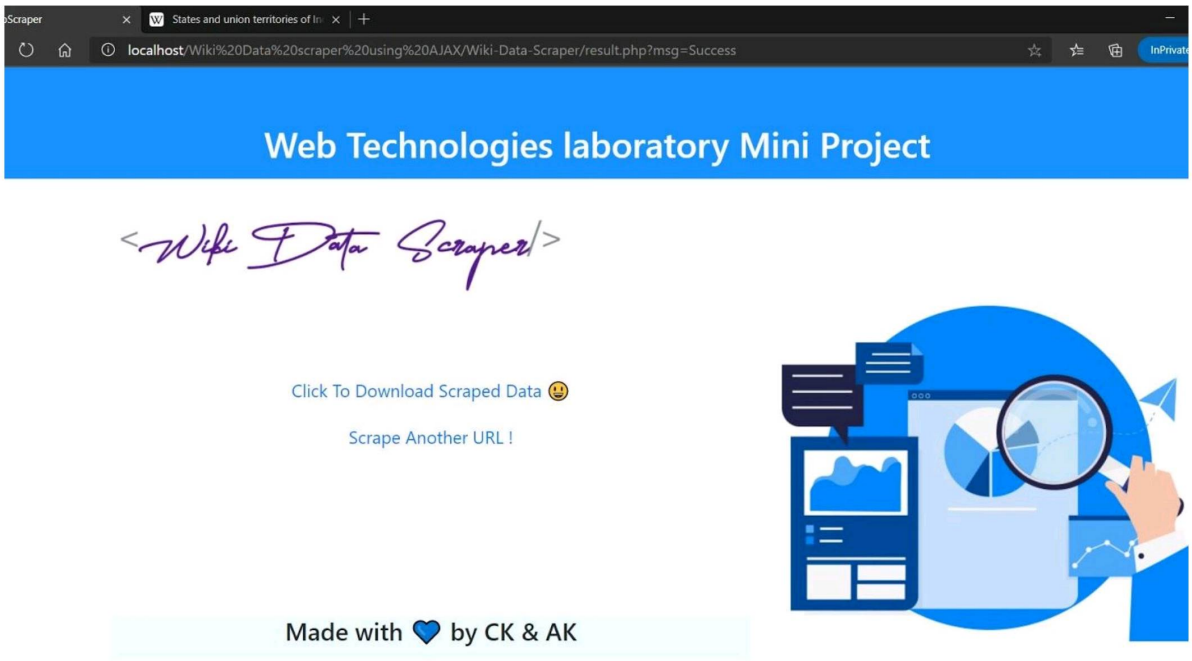

fig:- 5.3: Result Page

- The above figure shows the result page of the web application.
- This is the page that loads after the scraping operation is done.
- If the status is success, it provides a download link to the scraped data.
- If the status is an error, it provides a message and a redirect link to the index page.

4. Records Page:

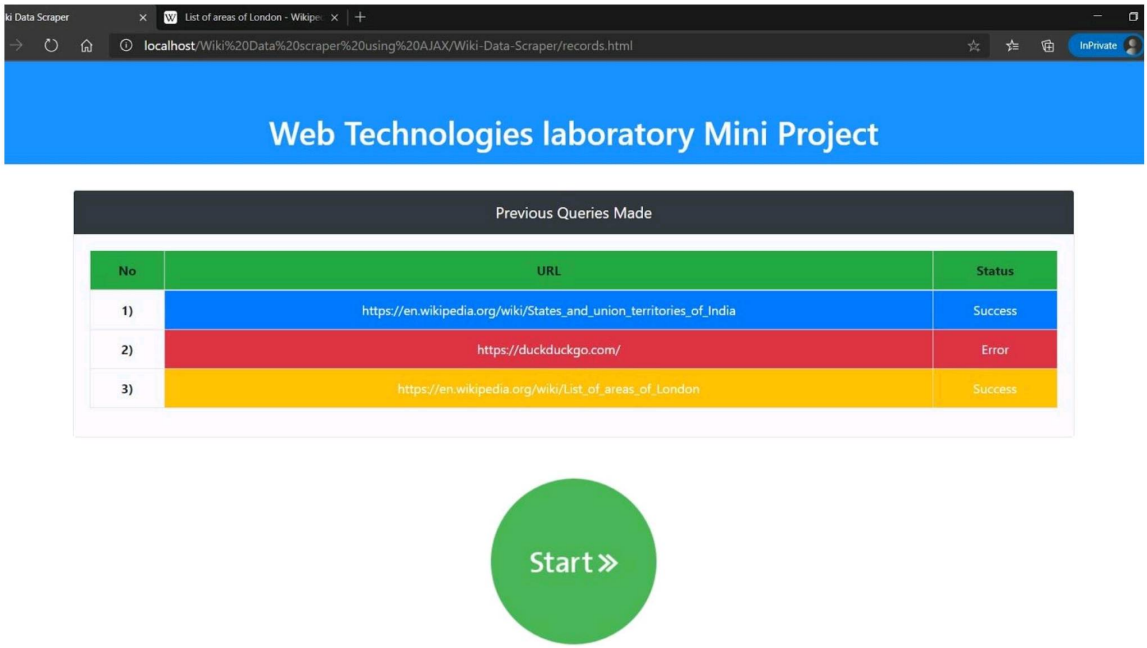

fig:- 5.4: Records Page

- The above figure shows the records page of the web application.
- This is the page that loads when the user clicks the “view db” button on the Index page.
- This page takes in the data from the database which is modified by the python program and uses AJAX, to fetch the records.
- The records are stored in json format.
- Everytime the main page is loaded, the database records are refreshed and a new json file is formed.

5. Database:

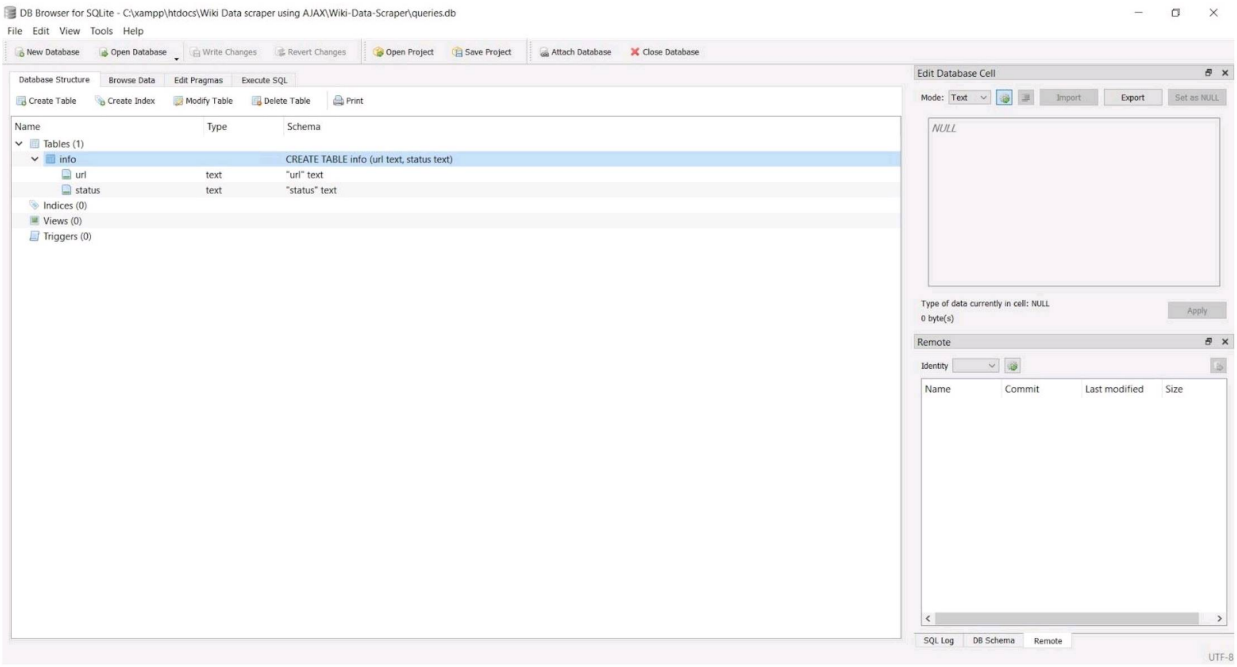

fig:- 5.5: Database Snapshot

- The above figure shows a snapshot of the database.
- The name of the database is “queries.db”.
- The database has a single table called “Info”.
- This table stores the URL and status.
- Whenever a call to the db is made, an Id is generated and the URL and status are in a list.
- The ID and the generated list are stored in Key-Pair values which make it easy for us to manipulate the data and use it for our operations like displaying, sorting etc...

Chapter 6

CONCLUSION

The proposed Wiki Scraper creates an online platform for Data Scientists to scrape data much easier and comfortable compared to the existing rigid availability of fixed data sets. It reduces the number of human resources that spend time on making data sets. This System makes the work faster and more efficient. The advantages of Scraper is having flexible Data sets in moments notice. Any person without any prior knowledge can create unstructured Data sets instantly.

Some of the benefits of this application are:

- Easy to use.
- No cost involved.
- Easy understanding of data sets because of the simplicity

Chapter 7

FUTURE ENHANCEMENT

- It is not possible to develop a system that meets all user requirements, as user requirements keep changing as the system is being used.
- There are many places where we can improve. The following are the features that we can implement in near future :
  - Increase the efficiency of the scraper
  - Add Multi Language Scraping option
  - Add more wide range of download format options
  - An admin management can be added to look over the entire system.

REFERENCES

1) https://stackoverflow.com/questions/6235785/run-a-shell-script-with-an-html-button

2) https://stackoverflow.com/questions/1125637/running-python-code-from-a-server#1125645

3) https://duckduckgo.com/?q=execute+python+code+from+server&ia=web

4) https://stackoverflow.com/questions/1901093/calling-python-from-javascript

5) https://duckduckgo.com/?q=how+can+i+execute+a+python+file+using+javascript&ia=web

6) https://www.wikihow.com/Make-a-File-Downloadable-from-Your-Website