# **"Web Scraping Financial Data from Wikipedia: Insights by the social platforms** "

COMPANY NAME :SHARPENED MIND TECHNOLOGIES PRIVATE LIMITED

DOMAIN:DATA ANALYST

MENTOR:MR.MADHAVAN SIVA

DATE:29-09-2024

SUBMITTED BY:

Y.Jeffreyy Paul

**Table of Contents**

1. Abstract

2. Introduction

3. Objective

4. Technologies and Tools Used

5.Libraries

6. Methodology

6.1.Data Collection

6.2.Data Parsing

6.3.Data Collection

7. Code Structure

7.1.Main Function

7.2.Flow of Execution

8. Results

9. Challenges

10. Conclusion

11. References

**1. Abstract**

This project focuses on collecting data from a web page using Python libraries. The project utilizes the BeautifulSoup and Requests libraries to scrape data from Wikipedia, specifically from the "List of social platforms with at least 100 million active users" page. The scraped data is then processed and structured for further use or analysis.

**2. Introduction**

The rise of social media platforms has generated vast amounts of data. This project aims to scrape information from Wikipedia about social media platforms with over 100 million active users. By using web scraping techniques, the project fetches relevant data from the web and organizes it for further analysis.

**3. Objective**

The main objective of this mini project is to:

* Use Python libraries to scrape web data.
* Extract specific information regarding popular social media platforms.
* Process and display the extracted data in a readable format.

**4. Technologies and Tools Used**

* Programming Language: Python

**5.Libraries:**

**1. BeautifulSoup:** For parsing HTML and extracting data from web pages.

**2. Requests:** For making HTTP requests to fetch the webpage contents.

**6. Methodology**

**6.1: Data Collection**

The project begins by fetching the webpage content using the Requests library:

**import requests**

**url="https://en.wikipedia.org/wiki/List\_of\_social\_platforms\_with\_at\_least\_100\_million\_active\_users"**

**response = requests.get(url)**

**6.2: Data Parsing**

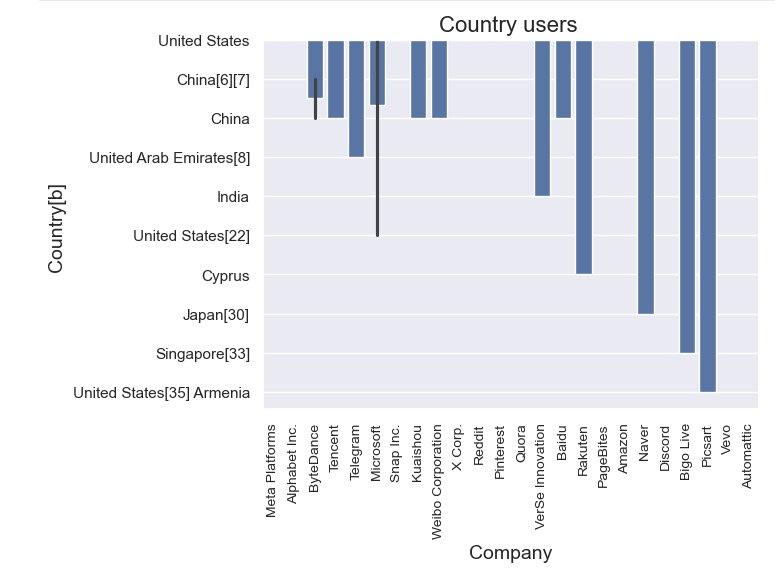
Once the webpage is fetched, the BeautifulSoup library is used to parse the HTML content:

**from bs4 import BeautifulSoup**

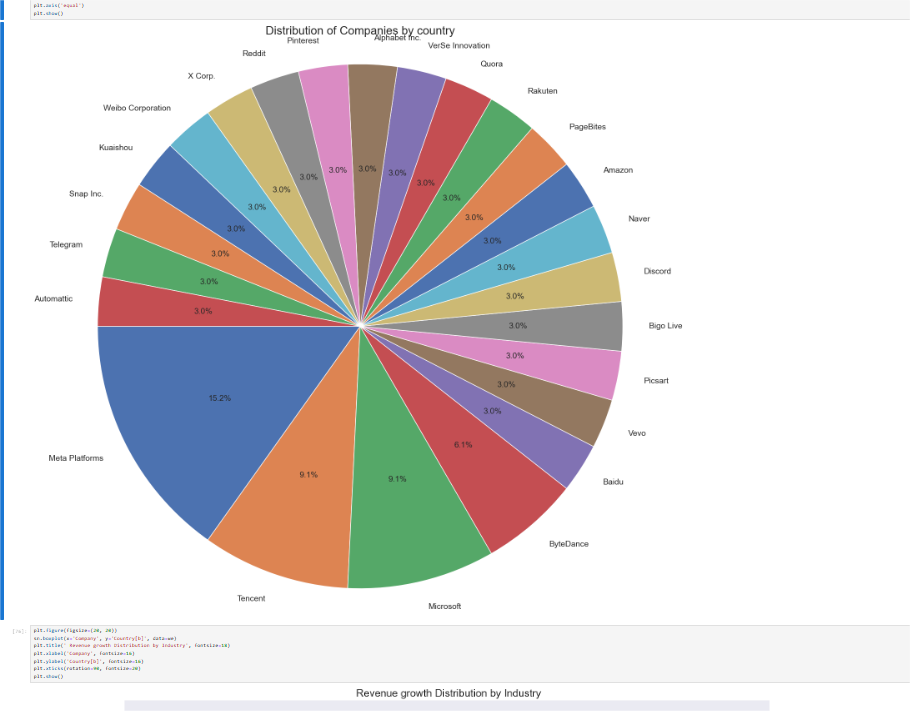
**soup = BeautifulSoup(response.text, "html")**

**6.3 Data visualization**

**1.Bar plot**



**2.Pie chart**



**7. Code Structure**

**7.1.Main Function**

* **Fetching data**: Uses requests.get() to retrieve the HTML of the Wikipedia page.
* **Parsing data**: Utilizes BeautifulSoup to parse the HTML and extract the relevant data.
* **Processing: Structures** the data into a readable format for output.

**7.2.Flow of Execution:**

1. Fetch webpage data using Requests.

2. Parse the data using BeautifulSoup.

3. Process and store the scraped data in a structured format

**8. Results**

The project successfully scrapes data from the Wikipedia page, including the names, companies, and active user counts of various social media platforms. For example, the extracted data for Facebook looks like this:

**{**

**"name": "Facebook",**

**"company": "Meta Platforms",**

**"country": "United States",**

**"launched": "2004",**

**"monthly\_active\_users": "3.070 billion"**

**}**

This data can be used for further analysis or reporting.

**9. Challenges**

Some of the challenges encountered during this project include:

* **Dynamic Content**: Some websites load content dynamically, making it harder to scrape using simple HTTP requests.
* **Handling HTML Structure Changes:** Changes in the HTML structure of the webpage could break the scraping script.
* To address these, careful parsing logic was implemented, and fallback options were considered.

**10. Conclusion**

The project successfully demonstrates the use of Python for web scraping by extracting relevant data about social media platforms from a Wikipedia page. The data was parsed, processed, and structured for analysis. Future enhancements could include automating the script to run periodically and save the data to a database.

**11. References**

BeautifulSoup documentation: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Requests documentation: <https://docs.python-requests.org/en/latest/>

GitHub Repository: <https://github.com/jeffreyypaul053/final>