**Competitor Scraper**

**Introduction**

The Competitor Scraper project, developed under the data science company Distributed Analysis, aims to gather key information from competitor websites through web scraping. This information includes company names, revenues, types of services offered, number of employees, CEO names, and more. By utilizing web scraping techniques with Python and various libraries such as Requests, BeautifulSoup, Scrapy, and Selenium, the project seeks to automate data extraction and analysis for competitive market research.

**Description**

Web scraping involves extracting data from websites to collect valuable insights and information. The Competitor Scraper tool utilizes Python as the primary programming language due to its versatility and rich library ecosystem. The selected libraries, Requests, BeautifulSoup, Scrapy, and Selenium, offer specialized functionalities for fetching web pages, parsing HTML content, and automating web browser interactions, making them well-suited for data mining and web scraping tasks.

### **Planning and Research**

To successfully complete the competitor scraper project, I will follow a detailed planning and research process. First, I will identify the project requirements by thoroughly understanding the key information needed, such as company names, revenues, types of services offered, number of employees, and CEO names. Next, I will use the libraries Requests, BeautifulSoup, Scrapy, and Selenium to scrape data from the specified website. Requests will help fetch HTML/JSON content, BeautifulSoup will parse and navigate the HTML content, Scrapy will handle high-level web crawling, and Selenium will automate browser tasks for data extraction.Once the data is scraped, I will structure and save it into a MongoDB database using Mage. This will facilitate easy data manipulation and analysis. After organizing the data, I will import the MongoDB database into Power BI to create visualizations and dashboards. These visualizations will help in understanding trends, patterns, and insights from the scraped data.Finally, I will validate the viability of the project through comprehensive data analysis and visualization. By meticulously planning and researching each step, I will ensure the project's success and generate valuable insights from the competitor data.

# **Concept of Project Ideas**

The project aims to develop a competitor scraper, a web scraping tool designed to extract key information such as company names, revenues, types of services offered, number of employees, and CEO names from a website named GOV UK Digital marketplace. Additionally, potential solutions including features, benefits, social value, and pricing will be explored. To achieve this, the project will utilize web scraping libraries such as Requests, BeautifulSoup, Scrapy, and Selenium. Requests simplifies the process of making HTTP requests, enabling the fetching of HTML/JSON content from web pages. BeautifulSoup is employed for parsing data from HTML and XML files, providing intuitive ways of navigation and modification. Scrapy, a high-level web crawling and scraping framework, facilitates efficient data extraction. Selenium automates web browsers, including web scraping tasks, to extract valuable information. The data collected will be managed using mage, allowing for filtering and presentation in various formats for analysis. An interface will be designed to enable users to select the desired data types and specify the websites for extraction.

# **Gathering Requirements**

Python was chosen as the programming language due to its versatility and extensive library ecosystem, ideal for web scraping. Visual Studio Code was selected as the code editor for its lightweight yet powerful features, suitable for project management and development across multiple platforms. Extensive research identified Requests, BeautifulSoup, Scrapy, and Selenium as the most suitable libraries for data mining and web scraping tasks. The project will implement algorithms for HTML parsing, data cleaning, storage, and analysis to effectively process the extracted data.

# **Risk Analysis**

The project encountered significant risks during the initial phase, including delays due to illness, extended timeline for requirement identification, and library compatibility issues. These challenges necessitated adjustments to the project timeline and milestones, highlighting the importance of flexibility and proactive risk management. By addressing these risks and documenting the mitigation strategies, a more resilient project plan was developed to overcome hurdles effectively.

# **Initial Evaluation**

The objective of this task is to assess the project's feasibility based on comprehensive research and analysis conducted in previous tasks. The technological stack, attributes for data extraction, and filters to refine data have been identified to ensure clarity and alignment within the team. By the end of this phase, a decision will be made on whether to proceed with the project, considering its viability and potential impact. Should the project be deemed unfeasible, a conclusion will be reached to cease development.

**Implementation**

The Competitor Scraper tool will follow a systematic approach for data extraction and analysis:

* **Data Collection:** Utilizing Requests and Scrapy to fetch HTML content from competitor websites.
* **Data Parsing:** Employing BeautifulSoup and Scrapy to parse the HTML content and extract relevant information such as company details and CEO names.
* **Data Storage:** Storing the extracted data in a structured format, in a database, for further analysis and reporting.
* **Data Analysis:** Performing statistical analysis on the extracted data to identify trends, patterns, and insights relevant to competitive market research.
* **Visualization:** Creating visual representations of the analyzed data using libraries like Matplotlib or Seaborn to generate informative graphs and charts for reporting purposes.

**Database Management with Mage and MongoDB:**

* Mage, a tool used for database management, is leveraged to automate the creation of MongoDB databases.
* MongoDB, a NoSQL database, is utilized to store the scraped data efficiently and flexibly, allowing for easy retrieval and manipulation during the analysis phase.
* The integration of Mage and MongoDB streamlines the database setup process, enabling seamless data storage and retrieval for the project.

**Visualization and Dashboarding with Power BI:**

* Power BI, a powerful business analytics tool, is employed for visualization and dashboarding purposes.
* It allows for the creation of interactive dashboards and reports, providing stakeholders with insightful data visualizations to facilitate decision-making processes.
* By integrating Power BI into the project, the extracted data can be presented in a visually appealing and informative manner, enhancing data analysis and interpretation.

**Conclusion**

The Competitor Scraper project aims to provide valuable insights into competitor data for strategic decision-making in competitive market analysis. Through the utilization of Python and specialized web scraping libraries, the project seeks to automate the process of data extraction, analysis, and visualization. By addressing potential risks and implementing mitigation strategies, the project aims to deliver a robust and effective solution for competitive market research under the Distributed Analysis company.