Web Scraping and Data Analytics for an AI-as-a-Service Platform

Eng Kay Tze, Kong Mun Jun
BSc of CompSc (Hons) in Data Science
Supervisor: Dr. Lim Siew Mooi (Sandy)

TARC
TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE
BEYOND EDUCATION

Industry Partner
CE.A.I. Centre of AI Innovation
MYFINB.COM

1. Abstract

Some information from the universities in Malaysia is fragmented and available in many places. Consequently, the public is facing difficulties knowing the universities' strengths. Therefore, this project proposed to develop an AI-as-a-Service platform with a centralized back-end database to overcome the problem. The potential students can find their desired courses from the right universities through this platform. Among many other functions, the university researchers can convert their areas of expertise into prototypes, curriculum, industry-friendly collaboration models and develop new research areas with AI-enabled engines. The proposed platform will store the details of the educators, his/her specialisation and achievement, then create a personalised AI expert system. The scope of the project is within Malaysia.

2. Problem Statement

1. Some university and ministry data are outdated, not standardized, and even fragmented.

2. Researchers face difficulties finding the universities and ministry information for their research and projects.

3. Government needs to have a strong connection with the ministries and universities to ensure the quality and coverage of the learning scope.

3. Objectives

1. To crawl and convert various required information from the university, government agency and ministry websites into csv formats.

2. To construct a centralized database to store the appropriate information obtained from item 1.

3. To provide convenient, efficient and accurate analytical results to the targeted users.

4. Design and Methodology

The evolutionary Prototype Model is used for constructing the proposed University-Industry Partnerships (UIP) Centralized Database Platform. This model allows the project to be continuously built and refined until the system's best condition is developed.

Target Market for this project are listed below:

i. University researchers
ii. Academic communities
iii. Centres of excellence of universities
iv. Government agencies and ministries
v. SMEs and corporations in 15 industry Groups
vi. Investors
vii. Undergraduates and Postgraduate students

5. Construction and Testing

We used Google Colab for data sorting and data cleaning with Python language. While Visual Studio Code is used to construct the database user interface with HTML5, CSS and JavaScript.

The databases for the project comprise universities, research cases, government agencies and ministries in Malaysia. We used BeautifulSoup, Selenium and ParseHub for scrapping the required information from the websites, then storing the data into Microsoft Excel and csv format.

User-Interface were created in a simplistic style by showing the module's features. White Box testing and Black Box testing were applied to refine the proposed website modules.

6. Contribution

The prototype system brings convenience to the targeted users. The dashboard and data analysis modules allow the users to make better decisions. Furthermore, this platform enables university researchers to commercialize their research projects and to be adopted by the industry.

7. Conclusion

In a nutshell, this project's initial aims are achieved. The prototype has clear goals and objectives. The UI design is user friendly and simple. This prototype provides data visualization with insights to improve the data-driven decision-making process. Although this prototype system brings numerous advantages for the target users, it may also consist some limitations which can be improved in the future.