DATA ENGINEERING -1 REPORT

DHARSHAN DHANASHEKAR (11038649)

* I have done the web scrapping and have used three different websites like:
* <https://www.olympiandatabase.com/index.php?id=44917&L=1%27>
* https://www.espn.com/olympics/summer/2004/medals/\_/view/overall
* https://www.espn.com/olympics/summer/2004/medals/\_/view/athletes
* <https://www.olympedia.org/sports>
* In the beginning, I wanted to scrap Athlete name and using the link , we could get the bio of the athlete , for this I have used BeautifulSoup and Scrapy but wasn’t able to scrape dynamic website, I have added some of the snapshots . I was finding difficulty with that.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In this I used BeautifulSoup , and used this url but by default it showed like “Show top 25” and after that even if I change it to “Show all” , there wasn’t any change .

* For this my group mate Arun had suggested me to use Selenium for the scrapping which is generally used for dynamic website which I was facing previously.
* After that I had found the Athlete data in ESPN website and I followed the scrapping process.
* In the Olympedia website, I scrapped the sports data , in this we can find Abbreviation , discipline, sport, season , olympic status . I have removed Olympic status.
* In ESPN website, I have extracted the Athlete data and the Country data , where in Athlete data consist of Athlete name, Gold, Silver , Bronze, Total medals. In this I have generated athlete id and added it a column.
* Whereas in Country data consist of country name , gold, silver, bronze , total , year , I have generated country id and added it to a column .
* Also in the Olympian website , I have added scrapped the table containing the year , continent , rank , gold , silver, Bronze, total medals.

These were the task that was being performed by me .

TEAM MEMBER SCORE

* ARUN KUMAR - He was being very helpful throughout the project and also assisted me with the scrapping part using Selenium . I also wanted to give special credits to him who assisted our team in overall project. ( Grade = A)
* TERRY POONACHA – He assisted me with scrapping the flag url in the country data. And also did the data ingestion part. (Grade = A)
* SHRUTHI PARADESHI – She had some knowledge about the database, which was very helpful for our project.(Grade = A)