Sabin Hashmi
Junior Data Scientist, KOPKOPI
10 June 2021

Brief Report

Duration : 3 Months

I hereby submit the final report of the work progress that I have done during the job period of three months at KopKopi, Gdansk - Poland.

The initial phase of the work includes converting the Green Beans competitors' data given in the form of pdf files and compiling it as a single excel file for further analysis. Some of the files have been converted using Python Scripts, some are converted using online web-portals which allows to grab the texts from the images/pdf files and convert them to excel or word files, the rest of the files were added manually to avoid mistakes. The files have been added to the drive.

After the completion of the first task, I moved to work on collecting the cafe details spread across Poland. Collecting all the data was challenging, the easiest way to fetch the data was through Google Maps API, but it was dropped due to the following reasons; 1. For each ping, the API will be charged a minimal amount, but to try out the fetching script was not a good choice to go with. 2. It always gets stopped with collecting 20 results. As an alternative, I used another website (http://www.yelp.com) to scrape the data of cafes in Poland. The country was split into different grids with major cities and scrapped over 1000 cafes across the country and the data converted into an excel file and added to the drive.

The next duty was to design and suggest a stable database for the organization's website. On 12th April, presented the final presentation on the database structure and other specifications to the team. Suggested two main possible web databases 1. Database for the already existing WordPress extension and 2. Google Cloud Database services. It was yet to decide which one to go with since the migration from one database to another is a long process.

Later as the final task, the team collected the competitors' data for roasted bean suppliers. This was most challenging due to the non-uniform structure of data fetched from

websites. After collecting the data, it underwent more preprocessing to make it a standard uniform table using Python Scripts automation. The final table has been added to google drive and shared with the team. This re-engineered table can be easily used for cross-comparison and analytics with the Green Beans data table collected earlier the month.

With all this, I conclude that the duties assigned to me are complete and submitted to the team. All the samples and works are shared on google drive to the team.

I take this opportunity to show my gratitude to the firm and the team. It was my pleasure to be a part of the firm.

-Sabin Hashmi