LinkedIn Web Scraping - Migration

Ersin Sönmez
Data Science Non-Thesis Master Program - (ID: 0079025)

1) Abstract

In this study, the analysis of a sociological issue has been studied to be considered within the scope of the course. As a subject, an analysis of a private sector/vertical in Turkey (Defence Industry) and white-collar workers (profiles who have migrated abroad) has been made. The data to be analyzed was captured with various scripts on LinkedIn and presented by anonymizing. For information about the project and pipeline, "Project_Presentation.pdf" can be viewed. For detailed information, the GitHub address of the project can be examined. (https://github.com/ersinsonmez/linkedin_scraping)

2) Scope and Motivation

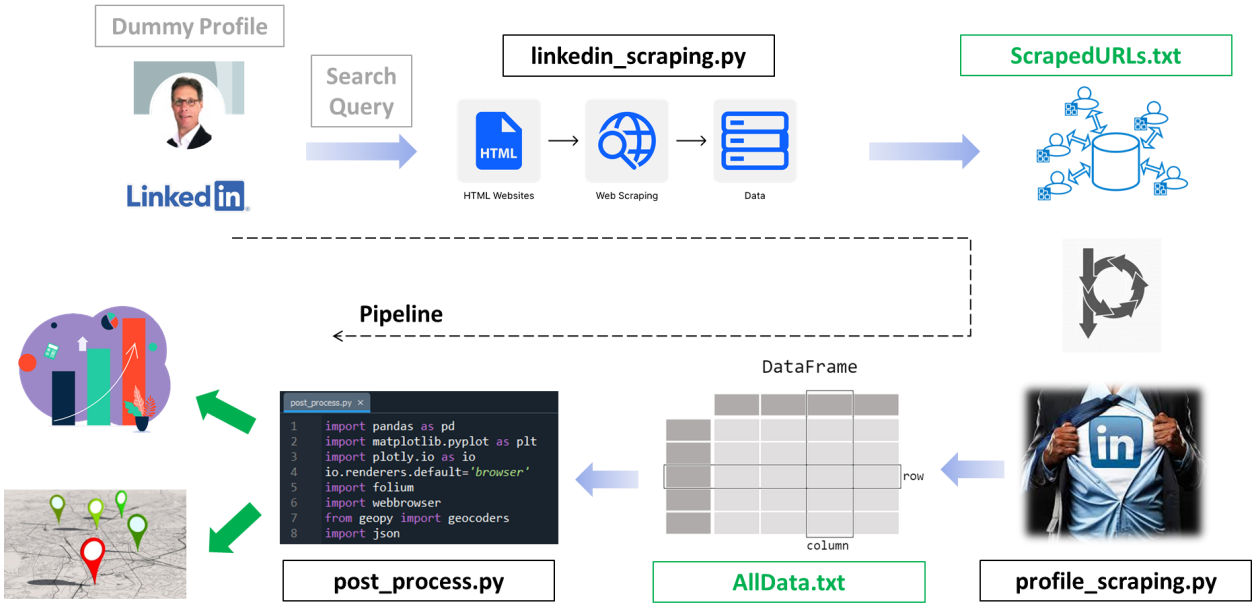
The data-based analysis of the migration movement, which has increased considerably in recent years, is the main purpose. This study will provide preliminary information on whether the news that is reflected in the media or that we see a lot on the internet and social media is true. It will be tried to access the data of the profiles that meet the migration criteria through "LinkedIn", the most effective site of the professional business network. The basis of motivation for this project is to validate the data-driven analysis of a sociological issue.

- ➢ An Example from the News: Savunma Sanayiinde Liyakat ve Beyin Göçü
- ➢ An Example from the News: Savunma Sanayisinde Beyin Göçü Engellenebilir mi?

Important Note / Disclaimer: Persons and institutions within the scope of the project have been examined for academic purposes and to be evaluated entirely within the scope of the course, personal data has been anonymized for this purpose and cannot be used for other purposes.

3) Structure and Pipeline

The structure is built on web scraping via LinkedIn. The process starts with a dummy profile (why API is not used in the following sections). The project is run with 3 scripts. The main purpose here is to pull data at discrete times, to process it in parallel and to post-process it. For this purpose, instead of a single script, several consecutive scripts have been created. The outputs in the report and presentation -highlighted in green- include data and/or visualizations at each stage. An interactive map at the end of the stream opens migration data around the world in a .html viewer.



In "Project_Presentation.pdf" and in scripts with .ipynb extension, what is obtained at which stage and coding is carried out by using which libraries are explained in detail. In this report, the important parts of the subject will be discussed rather than the technical flow.

4) Analysis of Collected Data and Results

The collected data, the operations performed in the stages and its reflection on the results can be briefly summarized as follows:

➢ **ScrapedURLs.txt**

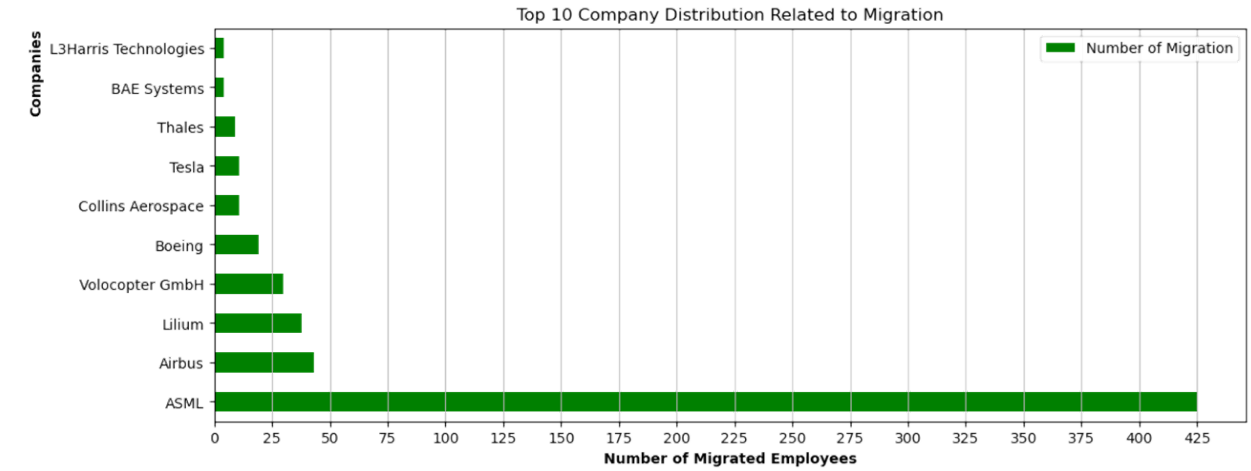
It is the row based data of the links of the profiles that are brought to us in accordance with the search result. (Anonymized under GDPR/KVKK)

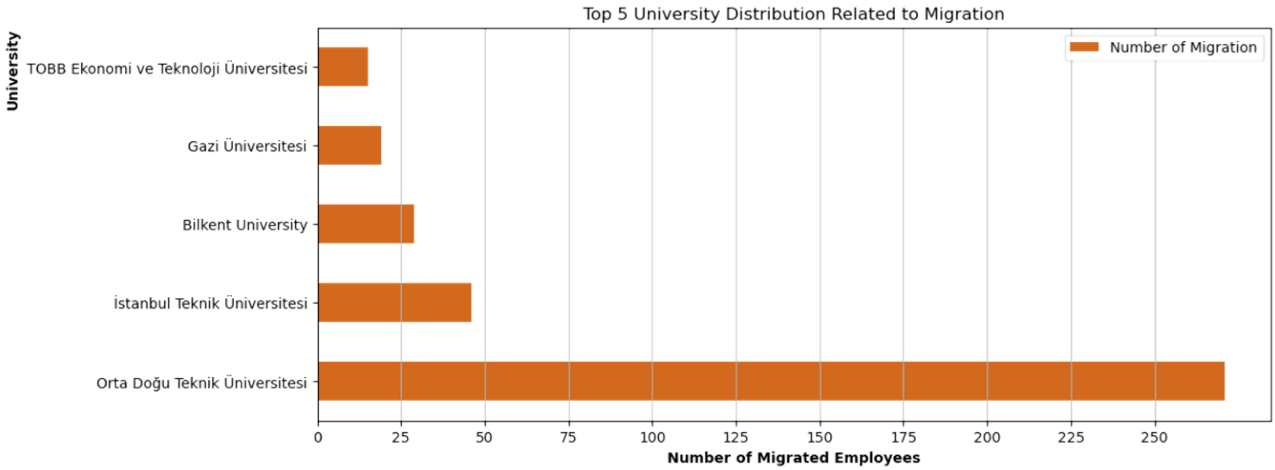
➢ AllData.txt

It is the data set in which the information from the short summaries of the profiles is compiled. It collects name, last job and university, title, location information. It is designed not to list missing information. (Anonymized under GDPR/KVKK)

➢ Migration --> Companies/Universities

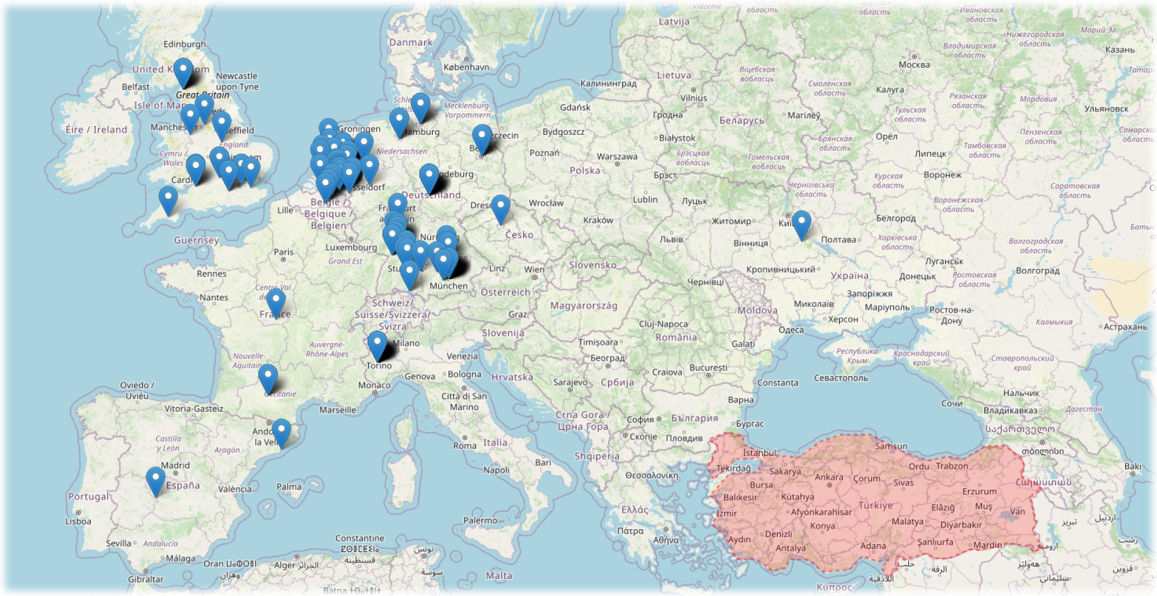
The Dutch company ASML received the most immigration, by a clear margin. The claim made in the sample news (in the "Scope and Motivation" section) was also clearly verified in this way. On the basis of universities, METU ranks first. Then ITU and other universities are listed. The reason for the dominant result here is related to the fact that the defense industry companies are mostly located in Ankara and accept many graduates from Ankara universities (especially technical universities such as METU). It can be observed that this trend continues in the continuation of the ranking. (Bilkent, Gazi, TOBB etc.)





➢ Migration --> Interactive Map

Geospatial analysis was performed by reflecting the location information sharing on latitude and longitude. City, Region or Country data is evaluated by a service (geopy - geocoders) and positioned on the map (folium and json). Within the scope of this project, in order for a profile to be considered as immigrated, it is considered that they have previously worked in companies in Turkey and then transferred to a company specified in the search criteria. Profiles that comply with this rule as a migration criterion but specify a location in Turkey were not considered as migration and were separated from the data. Some images of the outputs are given below. For a detailed and interactive review, the "migration_map.html" presented in the sources can be viewed by opening it with a browser.



5) Challenges and Recommendations

This project could have been handled using the LinkedIn API, but the challenge was that it required a verified company profile and the information it could provide was limited. As there was a time criterion within the scope of the project and the duration of the course, the API was excluded from the evaluation, and the data was tried to be extracted with web scraping.

Except for the basic biography, other elements of the profile such as previous company experiences, education details could not be scraped directly. When the experiments were carried out, the relevant parents in the source of the page did not flow with the same systematic. It has been observed that the classes that define the sections are replaced with unique IDs so that they cannot be scraped.

As a result of the site's blocking of the account in 125 profiles per day and beyond, the process was repeated with different accounts at certain time intervals and the data was collected and brought together at discrete times. Waits were made with certain periods (static times or random) on each page, so that the site does not detect the automation tool.

Rule-based arrangements were made in scenarios where users entered the university information incorrectly, unnecessary details in the name and title were deleted, very few data were visually checked and removed from the list when necessary. While the script was running, information messages about the status were printed to the user.

The maximum transaction quota of the service, which provides transition from location to latitude and longitude information, has been reached. In order to minimize the time elapsed between the data sent to the service and the received response, the map of the decrypted result was attached to the source codes. Multi-source use may be recommended for cases where the analysis of 500 or more results is examined.

6) Conclusion

When the sectoral situation is examined, it has been observed that the data set in question represents the population, and it is obvious that the brain migration from domestic defense industry companies to foreign related companies (Defence Industry - Top 100) abroad.