Multilingual Crawler for Indian Origin Academicians in Foreign Universities

Problem Statement:

There is no common platform in existence where the compiled data of Indian scientists and academicians working around the world in various universities/research laboratories under different disciplines of engineering and science is present. Along with this, the academic websites of universities of many nations are in their native language, especially BRICS nations. So it poses a problem to crawl and translate them in the English language. This data is required by Govt. of India to facilitate academicians, researchers as well as students (looking for internships, higher studies, research projects, etc. in foreign countries) with whereabouts of Indian scientists, professors to seek their help.

Proposed Solution:

A Multilingual crawler would help in extracting the URLs which might have faculty member names in their pages and would fetch all the details of the respective faculty members like their names, email, department, etc. It will find proper nouns, mostly names on those pages by applying NER and, in the case of BRICS nations, translate these names to the English language. After that, we can manually extract the other parts of the dataset required using the final set of names received through translation.

After the translation of the foreign websites, we may fetch their data and analyze the required details of the faculty members, and then we would render them in a graphical user interface where the users can interact with the data.

Technologies:

So, for the above-explained problem statement following technologies would be used:

➢ UNL (Universal Networking Language)

➢ NER (Named Entity Recognition)

➢ Machine Learning

➢ Neural Networks

➢ Translation APIs

Motivation:

This project would help many students who are interested in doing foreign internships or research projects and the institute managements who want to invite the Indian professors working in foreign universities to take guest lectures in their institutes. After the implementation of the project, we would be able to fetch the proper details of newly joined Indian faculties in foreign universities, and the data would remain up to date, dynamically.

This project has much significance, and thereby, it motivated the team to take this project as a minor. We would try to contribute as much as we can in this project.

Team Id: 25

Team Members:

Vishwajeet Singh (18103001)   **Piyush Girdhar (18103083)**

Rahul Garg (18103068)   **Anmol Salhvi (18103090)**

Team Mentor:

Prof. Manish Kamboj