Multilingual Crawler for Indian Origin Academicians in Foreign Universities

Introduction:

The purpose of this document is to build an automated system for the Government where the database of the Indian Origin Academicians extracted from the BRICS nations has to be updated according to all the changes made in their personal and professional contact details later on so that the actual database should remain consistent with the new data.

Intended Audience:

This project is a prototype for the Data Revisiting System and is restricted within the college premises. The project is being implemented under the guidance of Prof. Manish Kamboj. This project is useful for the Indian Origin Academicians database to stay consistent with the actual information present on their respective faculty' websites.

Project Scope:

The purpose of this Data Revisitor is to add the updated changes to the new CSV file corresponding to the given faculties by applying Natural Language Processing and extracting the labeled data according to the information required and compare it with the original data we have in our database so that it always remains consistent with the new data updated on their respective websites.

Moreover, to crawl all the Indian Faculties Information, we first need the actual seed URLs to start crawling and then apply the respective algorithms to filter out the required data. It would also be extracting all the Seed Urls for different universities, colleges, and other higher education institutes across the entire country.

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 for 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. For all the faculty information extraction, first, we also need to scrap all the seed URLs for different universities, colleges, and higher education institutes across the entire country. 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 would render them in a graphical user interface where the users can interact with the data.

After we extracted the required data, there poses a need for the database to remain up-to-date whenever the user hits on the website for searching for any particular faculty(mostly its updated contact information). This could be done by extracting all the necessary data corresponding to a particular faculty training with the Machine Learning model using NLP and text annotation tools. This would help in comparing the new and the old data up to which percentage the change is there and whether replacement is needed or not.

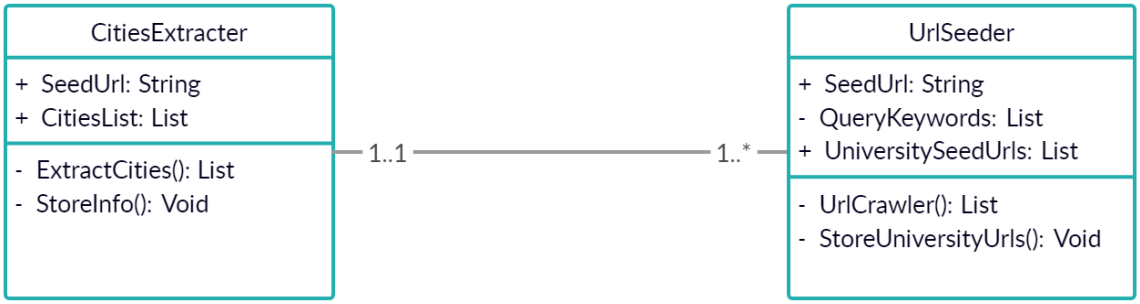
Technologies:

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

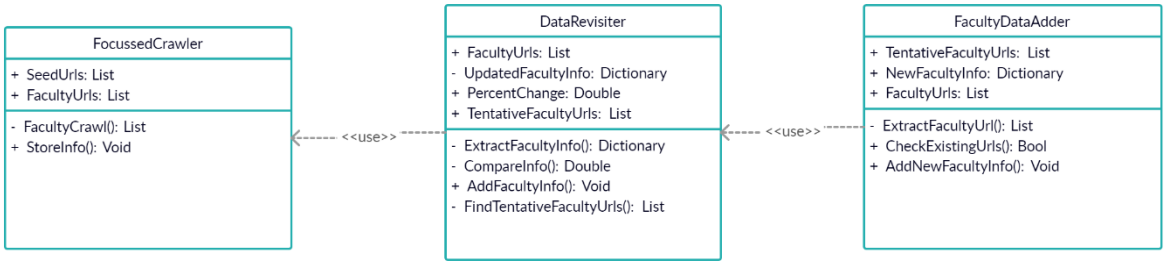
- Natural Language Processing
- NER (Named Entity Recognition)
- Machine Learning
- Neural Networks
- Scrapy
- Annotation Tools(TagTog)

Class Diagrams:

The below diagram shows the interactions among different functions while extracting the Seed URLs for all the universities and colleges present.



This second below diagram shows how the data is being used for revisiting the existing faculty URLs once again and making sure that they are consistent with the new updated information on the respective faculty websites.



Motivation:

This project would be helping a lot of students who are interested in doing foreign internships or research projects under them, or if some institutes want to invite the Indian professors working in foreign universities to take guest lectures in their institutes. After implementing the project, it would be dynamically able to fetch the correct details of newly joined Indian faculties in foreign universities, and the data would remain up to date.

This project has much significance for students, and thereby, it motivated us to take this project as our minor project, and we would try to contribute as much as we can in this project.

Team Id: 25

Team Members:

- **Vishwajeet Singh (18103001)**
- **Rahul Garg (18103068)**
- **Piyush Girdhar (18103083)**
- **Anmol Salhvi (18103090)**

Team Mentor:

Prof. Manish Kamboj