Patenting in Post-Secondary Institutions
Creating a Patent Database Using Web Scraping

Marc Neville marc.neville@canada.ca
Canadian Intellectual Property Office
Mazahir Bhagat mazahir.bhagat93@gmail.com
Canadian Intellectual Property Office
Amira Khadr amira.khadr@canada.ca
Canadian Intellectual Property Office

Background

Canada is a worldwide leader in terms of academic research and has one of the most-educated workforces in the world.[1] A recently released report from the Canadian House of Commons Standing Committee on Industry, Science and Technology entitled *Intellectual Property and Technology Transfer: Promoting Best Practices* provided policy recommendations on topics that included commercializing academic intellectual property (IP).[2] In anticipation of the future demand for analysis around academic patenting activities in Canada, the Canadian Intellectual Property Office's (CIPO) Economic Research and Strategic Analysis team decided to create a repository of patents held by post-secondary institutions (PSIs) and their associated inventors (professors, post-docs, graduate students, etc.). The main challenge with creating this repository is that not all Canadian PSIs follow the same policy around IP protection. Some PSIs have a creator or inventor owned policy, some have a joint ownership policy and sometimes the ownership is determined based on the quantity of resources from the institution that were used in its development. In addition to capturing patents by institutions with institution-owned or joint ownership IP policies, the data included in this repository will also include patents associated with PSIs that have an inventor-owned IP policy.

Objective

The objective of this project is to create an academic patent data repository. This would involve extracting patents by Canadian PSIs as a first step. This would also involve capturing the professors/inventors for which the name of the PSI does not appear in the patent assignee field, due to policies regarding the ownership of IP, using the names previously scraped. The end product will be an easily manageable dataset of patents associated with PSIs that will be used to identify and analyze their patenting activities. This exercise, using web scrapping and data matching techniques, will serve as an opportunity for CIPO to document useful techniques and best practices for data matching using python scripts that will be useful in future projects.

Methodology

This project involves the use of the following three datasets:

Dataset	Description
Scraped dataset	This is a dataset that was scraped from provincial government websites, independent web pages and individual faculty websites to identify inventors

[1] https://www.budget.gc.ca/2017/docs/plan/chap-01-en.html

[2] https://www.ourcommons.ca/DocumentViewer/en/42-1/INDU/news-release/9263749

	associated with PSIs.
Inventor dataset	This is a patent dataset extracted from Derwent Innovation based on inventor names and that were identified in the scraped dataset.
PSI dataset	This is a patent dataset extracted from Derwent Innovation based on Canadian PSIs as assignees.

In order to create an academic patent data repository that includes patents from PSIs, CIPO performed web scraping. For the most part, this data was either published on provincial government websites or on independent web pages. For provinces where such data was not available, CIPO referred to individual faculty websites associated with each PSI. The selenium package in Python was used to perform the web scraping, where the package simulates a user surfing the Internet and automates the process of clicking buttons on websites. The output was then stored in a table with inventor, position and employer data. A Python script was then developed to extract the first name, last name and middle name of the inventor from the Inventor Name field of the scraped dataset. The first names and last names were then used to extract patent data from Derwent Innovation. Meanwhile, a separate search was done to extract patent data based on Canadian PSIs as assignees. This was named our PSI dataset.

The final step of this project was to link the scraped data to create a database of patents by PSIs. This process was carried out using the Python Record Linkage Toolkit. The package performed a fuzzy match in order to link records together by calculating a metric known as the "Jaro-Winkler distance". In this step, an association rule mining algorithm was used to find frequent occurrences of inventors with their respective academic institutions in the PSI dataset. This algorithm was created to account for the *"mobility of professors"*. A metric was then developed to indicate the level of confidence associated with each recorded pair. Finally, in order to improve the accuracy of this record linkage process it was necessary to include the inventor's address and proximity to the PSI.

Follow-on Work

Now that a database has been created consisting of patents in Canadian PSIs, the next steps of the project entail:

1. Validating results.
2. Identifying other criteria that could improve the validation of the result set, including additional datasets.
3. Undertaking an extensive analysis of patenting by PSIs in Canada.
4. Creation of useful indicators based on the analysis described above.
5. Benchmark with other countries.