Date: 6th August 2019

Co-op Reflective Report

Financial Variable Extraction from PDF Documents using Graph-Based Techniques

Professional Master's Program (Big Data or Visual Computing)

Co-op Job Title: Data Scientist
Company Name: Statistics Canada
Semester & Year: Summer Term (III) – 2019
Name & Email ID: Anurag Bejju – abejju@sfu.ca

This report is based on the project undertaken during the coop program at *Statistics Canada* located in *Ottawa - Ontario, Canada.*

1. Statistics Canada

1.1. An Overview of the Company

Statistics Canada (French: Statistique Canada) was established in the year 1971 by the Canadian government to produce statistics that help Canadian public understand its population, resources, economy, society, and culture. It is headquartered in Ottawa and has regional branches in Vancouver, Toronto, Halifax, Montreal, Winnipeg, Edmonton and Regina [1]. StatCan is considered to be the best statistical organization in the world by The Economist which produces statistics for all the provinces as well as the federal government. In addition to conducting about 350 active surveys on virtually all aspects of Canadian life, Statistics Canada undertakes a country-wide census every five years on the first and sixth year of each decade.

1.2. An Overview of Data Science Accelerator Team

Statistics Canada is undertaking a significant transformation and leading efforts to be more responsive to the data needs of the society. Their modernization efforts are responding to issues raised by Canadians, obtained through extensive coast-to-coast consultations held this past year. They have started an ambitious journey to provide timely, relevant and quality information that Canadians expect and deserve from their national statistical agency with the strong privacy and confidentiality protections.

In *Spring 2018*, STATCAN decided to establish the data science accelerator [DSA] program to build their data science capacity catered to solve concrete problems. From its inception, DSA has pioneered new ways to implement applied research and state-of-the-art techniques to develop, automate & operationalize data processing capabilities across the entire organization. Having worked on more than 40+ use cases, they have been a driving factor in reaching the organization's modernization goals.

2. Project and Co-op Overview

Over the last 4 months, I have worked on a single main project as well as an additional supplementary project. The objective of my main project was to *correctly identify a page from a financial report document and extract some key variables from a specific table on it*. Since the PDF Structure of a financial report can drastically change from one company to another and the labeling of the intended variables can differ, it becomes a very challenging task to create a standardized extraction process that can generalize well for all types of PDF formats.

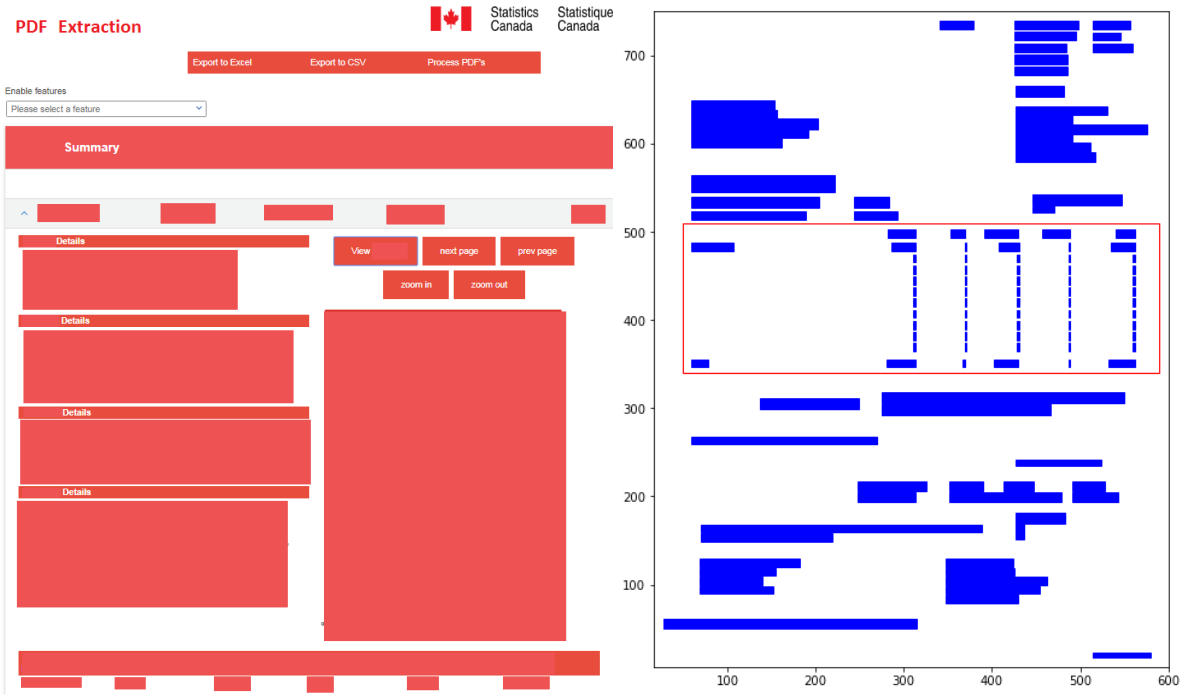

Fig 1: *Camelot table detection framework and template for the front end flask application created*

My supplementary project was fairly simple as the PDF structure was fixed for all input documents. Using open-source python library *Camelot-py* [2] and my own modified patch mounted on it, I was able to build an extraction script that could rightfully identify the position of an open-source variable and find the value associated with it. This was successfully completed and delivered to the client within a week timeframe. The relevant subject matter reviewed and validated the extracted variables and was satisfied with the end result.

2.1. Project Methodology

For my main project, I have come up with a methodology to appropriately counter the variability factor of a PDF Structure by using *logic* and *uniqueness* generally found in a financial report of a company. As part of my methodology, I appropriately identify a subsection of a document that has a large range of tables clustered together. This is done by finding the highest table to page ratio for a page group presented by the rolling mean averages method. Once we have identified the cluster of pages, we use bag-of-words and table to page ratio clubbed with other flag variables as features to generate a *RandomForestClassifier* that can correctly identify the type of a page.

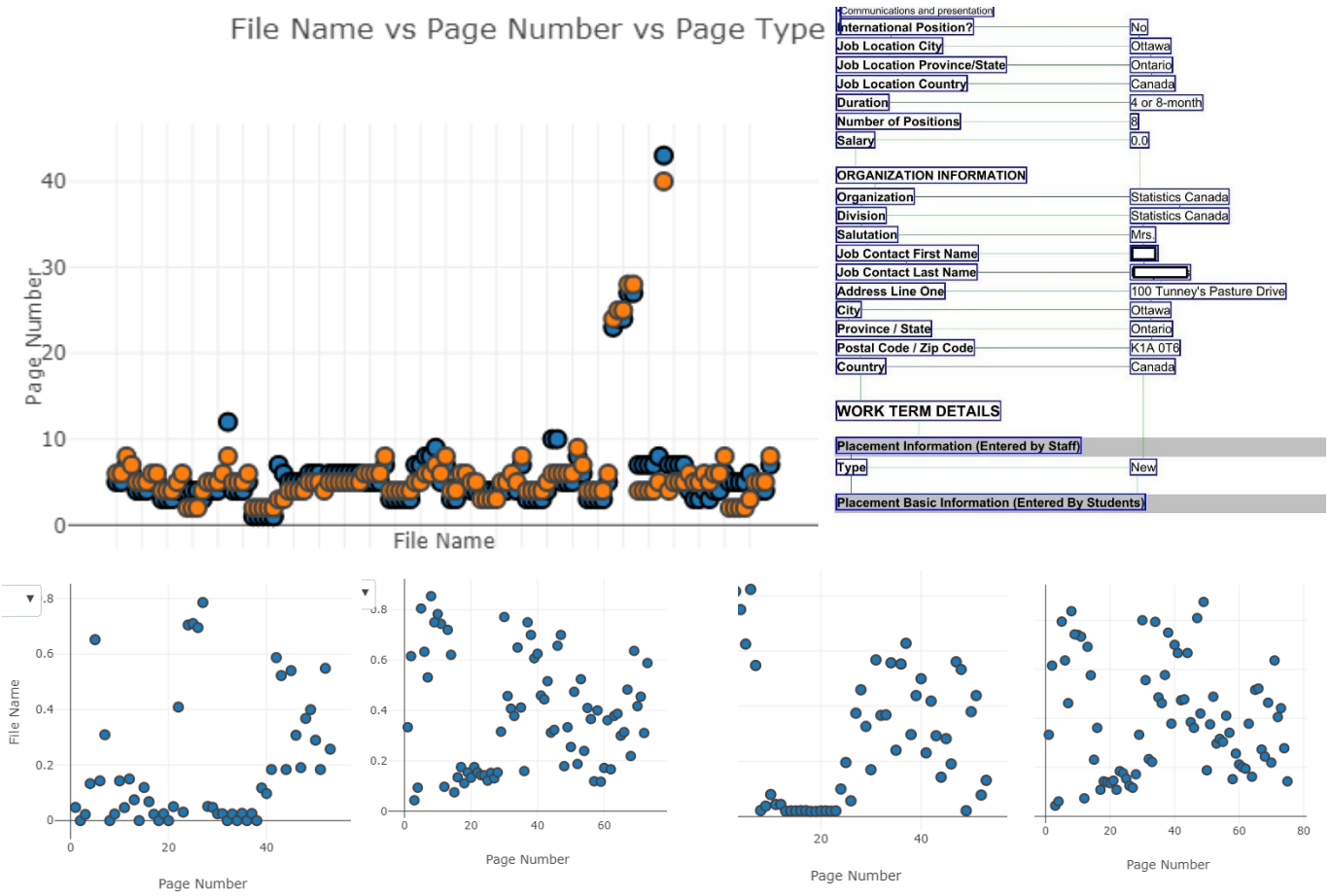

Fig 2: *Exploratory Data Analysis for Page Clustering and PDF-to-Graph Example*

With the correct Page detected, we transfer the table present on that page to its associated graph using state-of-the-art technique] The authors have found a new algorithm that uses spatial areas in association with text to create nodes with horizontal or vertical edges. Using this generated graph, we created extraction rules that traverse through the nodes and correctly identify the value of a financial variable.

2.2 Validation and Results

To build our RandomForestClassifier model, we have split the entire training set to *70-30 ratio*. Based on this split, we had a training accuracy of roughly *99.15 %* for our page detection task. Upon testing it on a larger unknown test set and completing the validation process, I was able to achieve close to *98.5 % test accuracy* for each type of page. With the page rightfully detected, I developed a graph retrieval algorithm that extracts values associated with a financial variable. This algorithm when tested on the largest test set, it generalized well and correctly identified *94%* of the time.

3. Personal Perspectives & Learning

3.1. Learning objectives:

Over the course, I was successfully able to achieve the following three main learning objectives.

- The first learning objective was to use all the knowledge and techniques learned in the last two terms of my professional masters in big data program at SFU to solve real-world problems. With the help of the foundation set from courses like big data 1 and 2 programming, I was able to create an entire big data pipeline and ensure each step was carried out efficiently.

- Communication is key while working on a team. Since there are multiple stakeholder's part of my project, communication was paramount to ensure the milestones are reached in timely manner. With constant feedback provided by my supervisor, I was able to excel at it.

- The last learning objective would be to *innovate*, *learn* and *implement* new techniques as part of our work. With the projects given to us, we had multiple opportunities to learn about the best practices in the data science world. Things like machine learning meetings, conferences and hackathons motivated us to grow our knowledge base in this field.

3.2. Work Place Culture:

With Statistics Canada putting a huge emphasis on health, safety, respect and fairness while creating their workplace culture, it gave me a sense of belonging and recognition and assured me that workplace wellness is at the heart of their organizational culture. With communication, cooperation and continuous improvement being part of their core values, I really liked working here as an intern. By providing opportunities to sit in on all group meeting with other full-time employees, it helped me understand the vision of the entire company as a whole. From fun barbeque lunches to a picnic at Britannia park, I thoroughly enjoyed every moment spent working here over the last few months. With the motivation and guidance provided by both my workplace mentors – *Monica* and *Saeid*, I was able to successfully showcase my talent to the entire organization as well as gain recognition for the work I have done.

3.3. Personal Development:

With most interns at Statistics Canada being assigned a full-term project, it allows us to not only extend our technical knowledge but also provides us with opportunities to grow our personality. From the very initiation, my supervisors had encouraged me to speak freely and share my thought process with them. With each project update requiring me to present my work in front of subject matter, I familiarized myself to communicate better and gain confidence to put my views clearly. This co-op has also pushed me to reach out of my comfort zone and allowed me to listen, share, debate and work with other data scientists across the organization. I was able to make a few amazing friends and participated with one colleague in a hacking health hackathon that we won. All these experiences throughout my journey, helped me immensely to grow as a person.

4. Acknowledgement:

Support on-demand, encouragement at the needed moment and guidance in the right direction are indispensable for the success of a project. I have received these in excess from all corners from various people, I am glad to submit my gratitude to them. I would like to first thank the entire SFU co-op team (*Eunice, Paula, Wendy*) and Big Data team (*Greg Baker, Greg Mori, Steven Bergner, Jiannan Wang, Kattie Knor*) for always providing the support and guidance needed for me throughout my co-op journey.

My sincere gratitude goes to *Saeid Molladavoudi* and *Monica Pickard* for constantly supporting and motivating me throughout my project. I was able to learn and thrive at my work because of the guidance provided by you. Last, but not the least, I take this opportunity to thank all the members of the Data Science Accelerator family who offered an unflinching technical and moral support during the entire course of my project.

5. References

[1] *https://en.wikipedia.org/wiki/Statistics_Canada*
[2] *https://camelot-py.readthedocs.io/en/master/*