Extracting data from Unstructured PDFs &

Analysis on Emergency Food Security Fund Data

Team #02:
Ravi Chandan
Jeffy Joseph

Objective

The project involves the conversion of food security funds data from PDF files via OCR and storing them in a database for further analysis. Even though PDF is one of the most used formats for document storage, it is not standardized. There are many types of PDFs, from computer-generated articles to scanned copies of old documents, all of which affect how well a set of instructions from software can "read" the text embedded in a PDF.

Methodology

To parse the PDFs, we used open-source OCR (Optical Character Recognition) engines and python as the programming language of choice. After exhaustive testing, ***EasyOCR*** and ***pyTesseract*** were determined as optimal OCR engines and ***OpenCV*** for the processing of PDF files. Each page of the PDF file is converted into an image file in PNG format, which is pre-processed before employing the pytesseract OCR engine to convert the text in the image to data points. Markers for odd and even pages are defined separately, the data output generated from the pytesseract engine is cleaned, and relevant height differences and width differences are generated. Data from each page is extracted and appended with the previous data to return a list of data rows which is stored as a CSV.

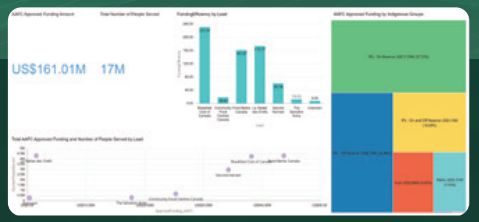
Introduction

The PDFs demonstrate the allocation of funds and their respective management by local food banks and other food relief organizations as a part of the Emergency Food Security fund (EFSF) program, a federal investment fund to support Canadians suffering from food insecurity as a result of the COVID-19 pandemic.

Upon minor inspection of data in the PDF, one could find an important pattern: text at the intersection of column names is stacked and shifted so that it could hardly be recognized as the additional feature of the same column name. Nonetheless, any data that does not fit rightly into a column or a row is widely considered unstructured, we can identify data from these given PDFs as semi-structured.

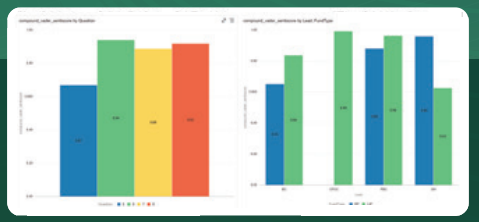
Results and Validation

A write-back of the extracted data is performed to store the information in a *CSV* file. Although the extracted data is accurate, some minor changes are made manually where OCR fails to recognize certain parts of the text. A combination of Excel spreadsheets and Python libraries, such as pandas, are used to verify data, as is manual validation by traversing the original pdf pages and comparing them to the data points.



Descriptive Analysis of AAFC Approved Funding Amount and the number of people served by Lead and Indigenous Groups. Funding Efficiency is computed as People served by Approved Funding Amount to understand the efficiency in distribution of funds. It is important to note that this is a suggested metric as there were incomplete data for some rows of AAFC Approved Funding and Number of People Served as well as some unknown lead names.

Vader based Sentiment Analysis of feedback form with English and French text to understand opinion of Leads for the Emergency Fund Security Project. Also, Machine Learning techniques have been used to predict approved funding status which yielded 83% accuracy but again there were some data limitations so we down sampled the data to fit the model.



This word cloud highlights the key words used in the PDFs. The largest and most prominent words in the cloud are "youth," "women," and "income," which reflect the means in which funds are distributed. The use of different colours and fonts adds visual interest and draws attention to certain words.