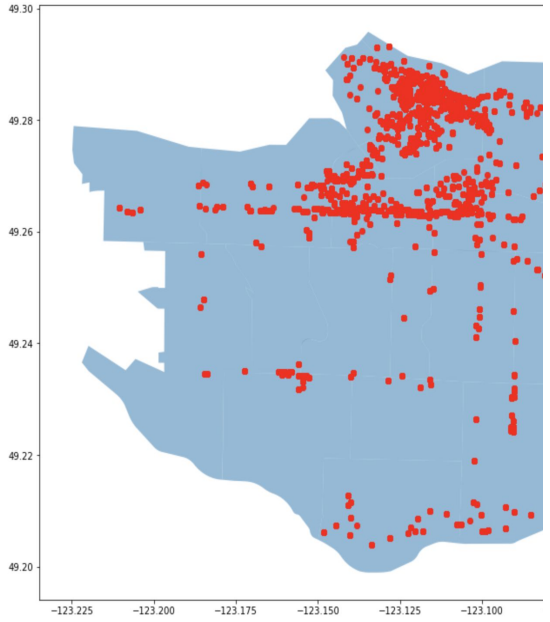
ESDC Integrity Hackathon 2021

UBC Team 13

Asma Al-Odaini, Rachel Wong, Lara Habashy, Simon Zheng, Javairia Raza

Exploring the Data

- Some anomalies (Ex. huge employee increase)
- Some postal codes are mistyped
- Non-Canadian cities found with country listed as Canada
- Possibility to verify mismatch in province/ city/ postal code
- Some missing addresses
- Longitude and latitude are all found within the BC area
- Some business names can be replaced by business trade name for searching purposes.



239319	NOT APPLIC
240388	V67 1N9
241570	V6E 2RA
241765	077190
241800	94304
242407	V6 Z3B2
242573	VGG 1V7
245196	V6B 15G
245239	V6E
245659	V6B 097
248706	V5N 414
249762	V5: 2P5
251568	V6C 3EZ
252336	22314
252437	33434
252511	33434
252564	TIK 6T6
252604	33434
253083	V6 Z3B2
257622	V6Z 21
258405	VV5Y 3X2
260575	V5R 631
260849	NOT APPLIC
261435	VV5Y 3X2
261819	V6 2W5
262443	98682
266761	V6 6B3
266845	V74 1K3

Ideas and Purpose

- Business names from the existing data set
- Ngrams - minimum 3 letter search
  - Keyword searches - “enterprise”, “ltd”, “inc” , etc
- Missing address scraped from Google (Beautiful Soup)
- Two final products:
  - 1. robust data repository and,
  - 2. a dashboard for a user friendly experience

Scraping Approaches

- Scraping process:
  - Canadianbusinessregistries.ca (API)
  - Google.com (BeautifulSoup Python)
  - Opengovca.com (BeautifulSoup Python)

- Limitations:
  - 14+ hours
  - Varying information on different sites

Implementing 3-gram Scraping

```
from itertools import product
from string import ascii_lowercase
keywords = [''.join(i) for i in product(ascii_lowercase, repeat = 3)]
keywords[0:5]
```

```
['aaa', 'aab', 'aac', 'aad', 'aae']
```

```
url = 'https://searchapi.mrasservice.ca/Search/api/v1/search?fq=keyword:%7B%22aaa%22&location=BC&lang=en&queryaction=fi
response = requests.get(url)
content_json = response.json()
```

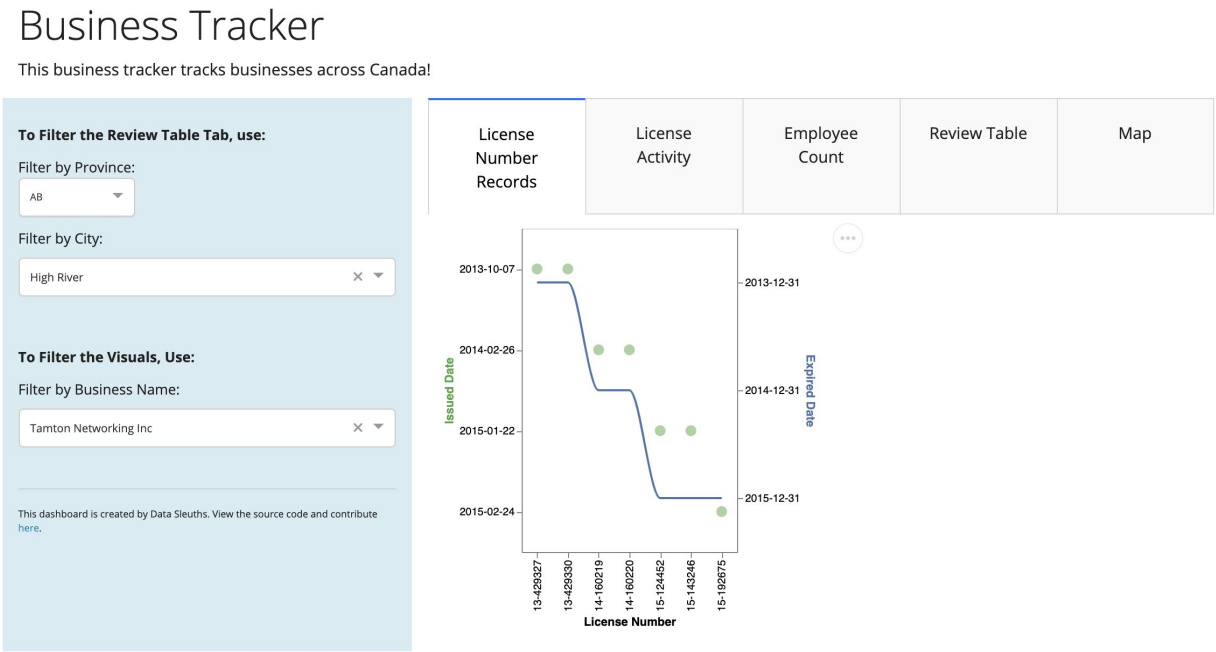
```
for ngram in keywords[0:5]:
    url = 'https://searchapi.mrasservice.ca/Search/api/v1/search?fq=keyword:%7B%22' + ngram + '%22&location=BC&lang=en&
    response = requests.get(url)
    content_json = response.json()
    time.sleep(3)
content_json['docs']
```

3-gram Scraping Dataframe

Jurisdiction	MRAS_ID	Company_Name	Status_State	Status_Notes	Status_Date	Reg_office_city	City	Reg_office_province	Entity_Type	Date_Incorporated
MB	MB_5812438	AAE TECH SERVICES INC.	Active	Active	2008-12-29	WINNIPEG	WINNIPEG	MB	MB SHARE CORPORATION	2008-12-29
BC	BC_A0102220	AAE TECH SERVICES INC.	Active	Active	NaN	La Salle	La Salle	MB	Extraprovincial Company	2017-04-18
BC	BC_BC0500961	AAE STRUCTURAL LTD.	Active	Active	NaN	Duncan	Duncan	BC	BC Company	1995-07-19
BC	BC_BC0750917	AAE EXPRESS CORP.	Inactive	Dissolved for Failure to File	NaN	Richmond	Richmond	BC	BC Company	2006-03-07
BC	BC_BC0871429	AAE HOLDINGS LTD.	Active	NaN	2010-01-15	LANGLEY	LANGLEY	BC	BC Company	2010-01-15
BC	BC_BC0522297	AAEA APPLICATION ASSISTANCE AND ENVIRONMENTAL	Inactive	Dissolved for Failure to File	NaN	CHARLIE LAKE	CHARLIE LAKE	BC	BC Company	1996-06-19

Dashboard

- Live Demo



Thank you for your time!

The Pitch!

Our dashboard allows investigators to view license number records for a specific business which could help then identify businesses with multiple license numbers in different years. They can also view the number of days a business had an active license across multiple years. It also allows them to compare employee numbers across the years to help detect anomalies.

In addition, the investigator can filter the dataset by provinces and cities to browse the data for that area and select features to view.