Final Project

Derek Yung-Sheng Lin – 301363392

Ogechukwu Kingsley Umeh – 301401702

Sajjad Shamanian – 301379812

Simon Fraser University

CMPT 353 - D100

Dr. Baker

December 11, 2020

The Problem

To what extent can we identify, compare and visualize the number of restaurants and chain restaurants near two locations within Greater Vancouver using Wikidata and the provided OSM data?

First, we started with the following prompt under the Other Questions section of ProjectTour: "I feel like there are some parts of the city with more chain restaurants: is that true? Is there some way to find the chain places automatically and visualize their density relative to non-chains?" (Baker, 2020).

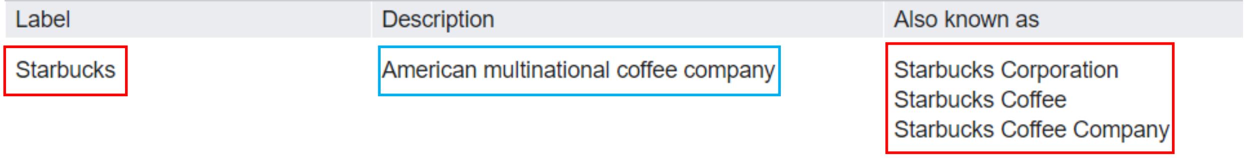
Second, we refined the problem by breaking the problem into smaller questions such as: What do we consider to be restaurants? What data do we have? What data can we get? What features can we use to identify restaurants and chain restaurants? To what extent can these features be used to identify restaurants? What parts of the city will we compare?

Finally, we attempted to answer these questions through brainstorming and exploratory data analysis.

Data Used, Data Acquisition and Cleaning

We used the provided OSM data (amenities-vancouver.json.gz) along with English names and descriptions from Wikidata as shown in Figure 1.

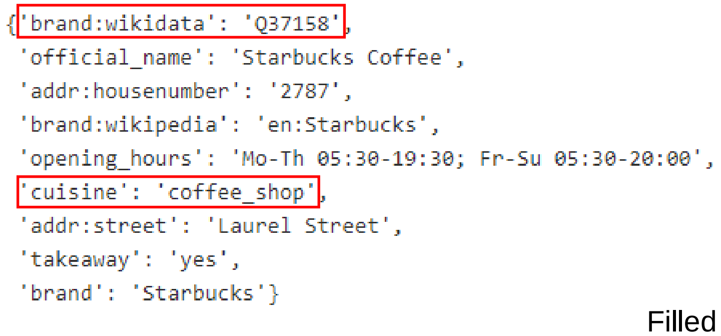
Figure 1. English Wikidata Names and Descriptions for Starbucks. (n.d.), Retrieved from https://www.wikidata.org/wiki/Q37158



Label	Description	Also known as
Starbucks	American multinational coffee company	Starbucks Corporation Starbucks Coffee Starbucks Coffee Company

Processing the OSM data involved extracting the 'cuisine' and 'brand:wikidata' (Wikidata identifier) tags shown in Figure 2 into cuisine and qid columns, creating a dictionary with (name, qid) key value pairs and using the dictionary to fill in missing Wikidata identifiers as shown in Figure 3.

Figure 2. OSM Tags for Starbucks

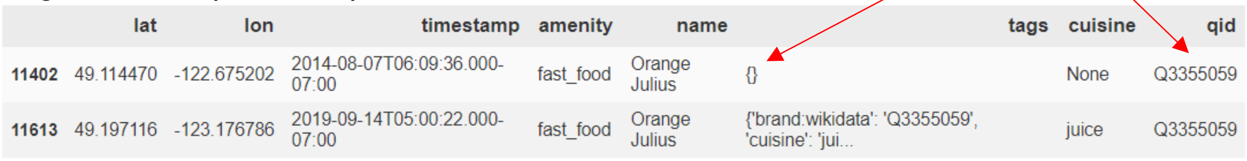


```
{'brand:wikidata': 'Q37158',
 'official_name': 'Starbucks Coffee',
 'addr:housenumber': '2787',
 'brand:wikipedia': 'en:Starbucks',
 'opening_hours': 'Mo-Th 05:30-19:30; Fr-Su 05:30-20:00',
 'cuisine': 'coffee_shop',
 'addr:street': 'Laurel Street',
 'takeaway': 'yes',
 'brand': 'Starbucks'}
```

Filled in Wikidata identifier

No 'brand:wikidata' tag

Figure 3. Sample of Preprocessed OSM Data

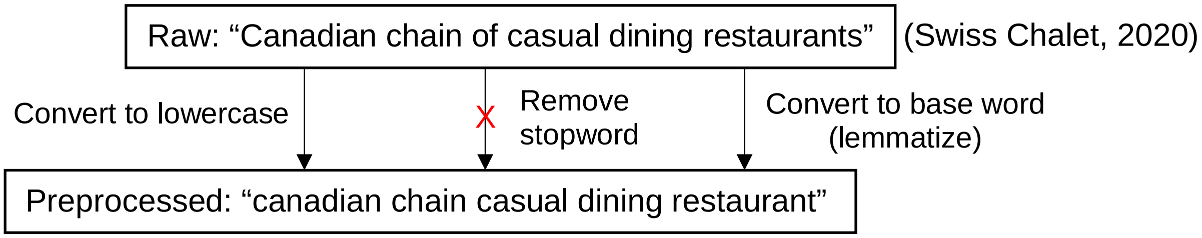


	lat	lon	timestamp	amenity	name	tags	cuisine	qid
11402	49.114470	-122.675202	2014-08-07T06:09:36.000-07:00	fast_food	Orange Julius	{}	None	Q3355059
11613	49.197116	-123.176786	2019-09-14T05:00:22.000-07:00	fast_food	Orange Julius	{'brand:wikidata': 'Q3355059', 'cuisine': 'jui...	juice	Q3355059

Wikidata names and descriptions were scraped using Requests and BeautifulSoup. We first looked at Wikidata's robots.txt for any rules we should follow when scraping. After failing to find a crawl delay, we limited our scraper to 1 request per second and added a failsafe that should stop the scraper upon receiving any bad HTTP responses.

These names and descriptions were then processed using Gensim and NLTK. The names and descriptions were converted into a list of lowercase words, had stop-words removed, were lemmatized and converted back into strings as shown below in Figure 4.

Figure 4. Raw and Preprocessed Wikidata Descriptions for Swiss Chalet



Raw: "Canadian chain of casual dining restaurants"
(Swiss Chalet, 2020)
Convert to lowercase
Remove stopword
Convert to base word (lemmatize)
Preprocessed: "canadian chain casual dining restaurant"

Techniques

Figure 5. Wikidata Description Word Cloud

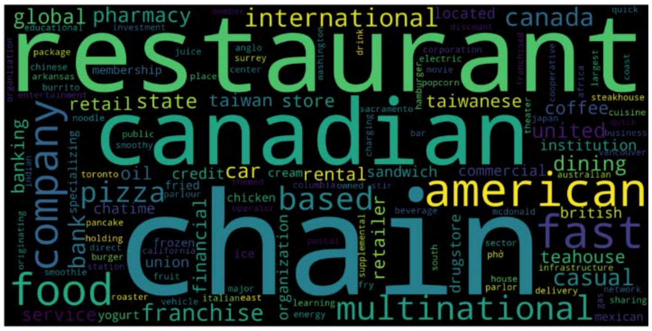
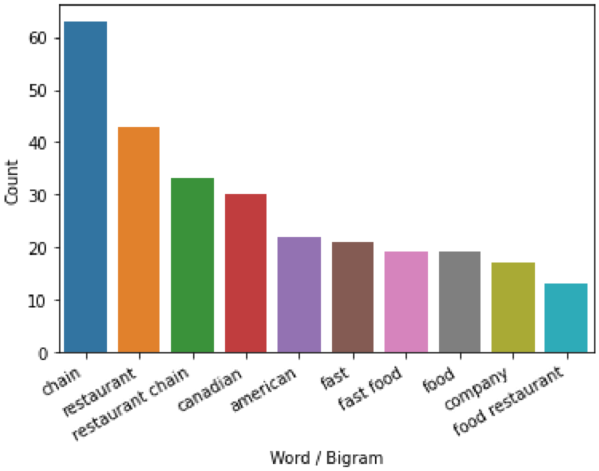


Figure 6. Top 10 Words and Bigrams in Wikidata Descriptions



Word cloud and n-gram counts were used to analyze the frequency at which words appeared in Wikidata names and descriptions. The size of words in Figure 5 and high counts in Figure 6 indicate that words such as ‘chain’ and ‘restaurant’ appear in numerous Wikidata descriptions. This is significant as it suggests that Wikidata descriptions can potentially be used to identify chain restaurants.

After discovering that Wikidata descriptions could potentially be used to identify chain restaurants, we analyzed the descriptions using document term matrices, cosine similarities, clustering and regex. Refer to Appendix A for a flow chart of the analysis and Appendix B for a sample cluster.

Results

We managed to identify approximately 5100 potential restaurants and 746 potential chain restaurants in Greater Vancouver using OSM cuisine tags and Wikidata descriptions, compare the density of restaurants and chain restaurants near two locations using chi-squared and visualize the density restaurants and chain restaurants.

Results produced by 05-analyze-and-visualize.py with default inputs. The interactive map can be viewed by opening *map.html* with your browser.

Restaurants and Chain Restaurants Within 5km of SFU Burnaby and SFU Vancouver

	Non-Chain Restaurant	**Chain Restaurant**
SFU Burnaby	116	21
SFU Vancouver	1671	154

Chi-Square p-value: 0.010067778591629453

Figure 7. Non-Chain Restaurants (Red) and Chain Restaurants (Blue) by SFU Vancouver

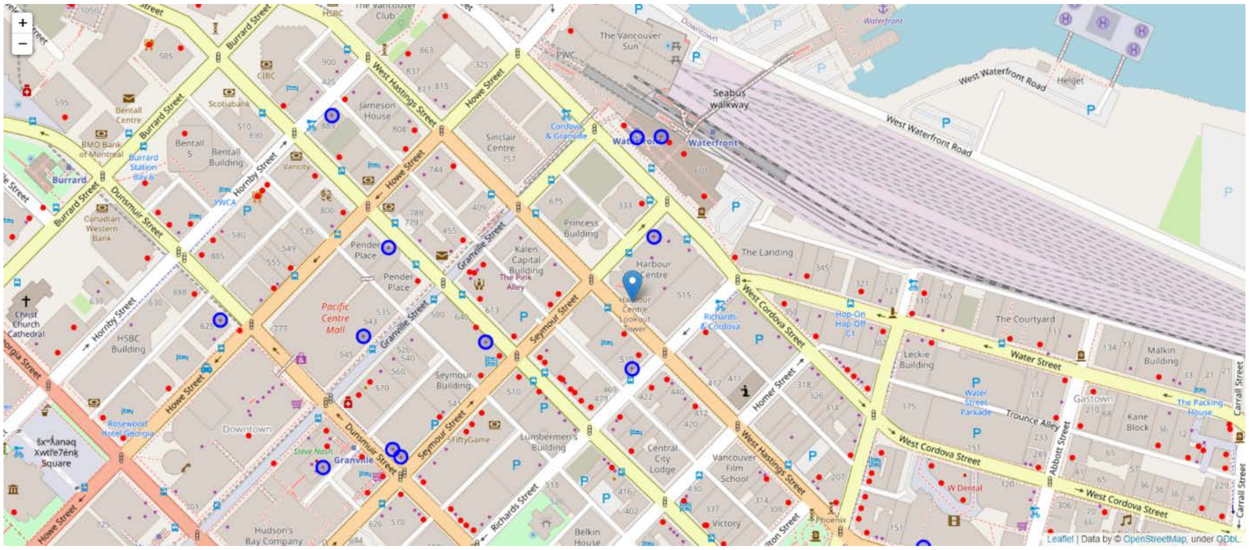
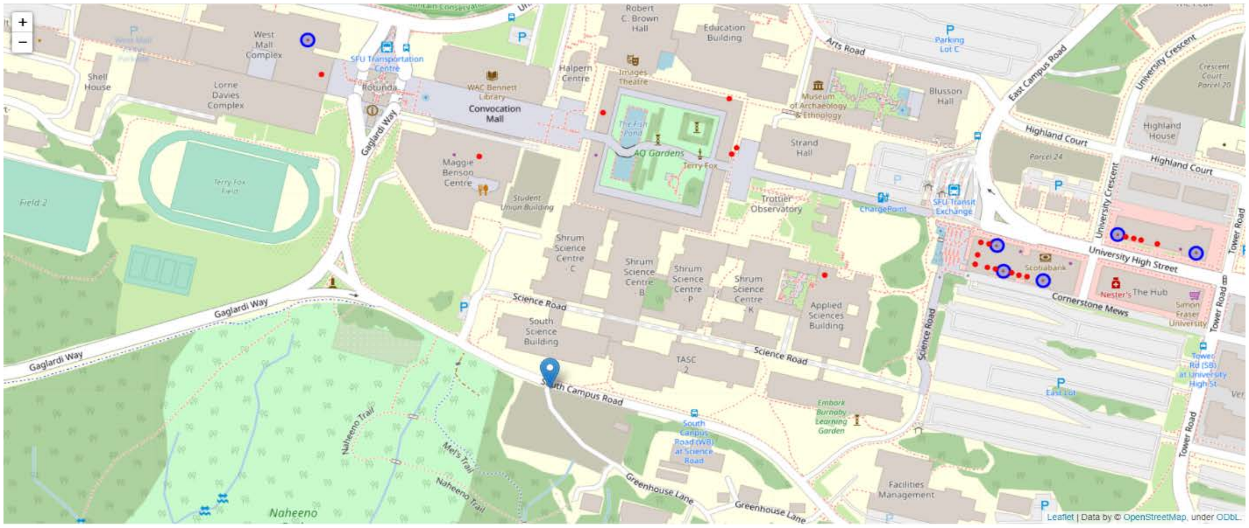


Figure 8. Non-Chain Restaurants (Red) and Chain Restaurants (Blue) by SFU Burnaby



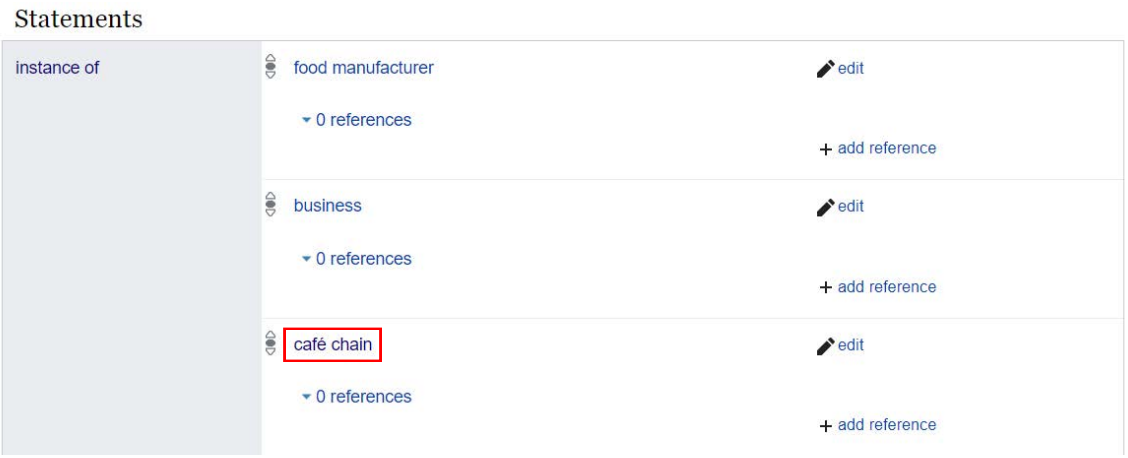
Limitations

Key limitations of our results include how we attempted to use clustering to identify chain restaurants, the various assumptions that we made to simplify the problem along with how we implemented parts our analysis.

First, using clustering to identify chain restaurants is a key limitation as we cannot quantify the uncertainty associated with our results using metrics like accuracy precision, recall, or F1 scores. This is significant as it may reduce the reliability of our subsequent chi-squared test and visualizations or even invalidate our results.

Second, we assumed that chain restaurants would have (1) Wikidata descriptions that contain both the words ‘chain’ and ‘restaurant’ or (2) Wikidata descriptions similar to descriptions that contain both ‘chain’ and ‘restaurant’. This is a key limitation as (1) not all chain restaurants have a Wikidata entry and (2) not all chain restaurant Wikidata descriptions are similar to ones that contain ‘chain’ and ‘restaurant’. A specific example of this would be the Wikidata description for Starbucks which contained neither words and was different from descriptions which did contain ‘chain’ and ‘restaurant’. This example is significant as demonstrates how our assumptions caused us to mislabel approximately 200 Starbucks locations as restaurants instead of chain restaurants. A possible solution to this would be to scrape and incorporate data from the Statements section of Wikidata (shown below in Figure 9) into our model.

Figure 9. Wikidata Statements for Starbucks. (n.d.), Retrieved from https://www.wikidata.org/wiki/Q37158



Statements

instance of	food manufacturer 0 references
	business 0 references
	café chain 0 references

Third, our main pipeline is limited by when we scraped Wikidata along with the final step in our main pipeline. We scraped Wikidata before filtering non-restaurants which is a limitation as it causes us to have to scrape more Wikidata entries. In retrospect, we should have first filtered the OSM data before scraping. The final step in our main pipeline 05-analyze-and-visualize.py is also a limitation as it assumes that inputs are well-formed and that there is no overlap between the two locations. If there is an overlap between the two locations, it will still perform the chi-squared test and return an invalid p-value.

In conclusion based on these limitations, we could only identify, compare and visualize the number of restaurants and chain restaurants near two locations within Greater Vancouver using Wikidata and the provided OSM data to some extent.

Accomplishment Statements

Derek

- scraped Wikidata using Requests and bs4
- preprocessed text data using Gensim and NLTK
- visualized text data using wordcloud and Seaborn
- calculated cosine similarities of Wikidata descriptions using Sklearn
- clustered Wikidata descriptions using Scipy

Kingsley

- Filtered OSM data to remove irrelevant data
- Calculated distances between OSM data locations and provided latitude and longitude using haversine: sample Lat Lon: 49.284478, -123.112349 (SFU Vancouver). For example, Location 1: Distances between location 1 and OSM data Location 2: Distances between location 2 and OSM data
- Filtered OSM data locations outside of a given radius for some latitude and longitude, for example with 5km
- With the assistance of a teammate, Derek, we used Chi-Squared together with the filtered data to compare the density of chain restaurants by two locations
- Finally, created visualizations for restaurants/chain restaurants using folium.

Sajjad

- Compute working hours based on the provided data in the OSM dataset
- Create new data set as a working hour values
- Analyze the created working hour data
    - Visualization of working hours distribution
    - Comparing the working hours comparison between chain and non chain restaurants
    - Performing Few Statistical Tests to Reveal some Underline Behavior
    - Working Hour Analysis For Each Day of Week

References

Baker, G. (2020, August 28). Project Topic: OSM, Photos, and Tours. Coursys. https://coursys.sfu.ca/2020fa-cmpt-353-d1/pages/ProjectTour

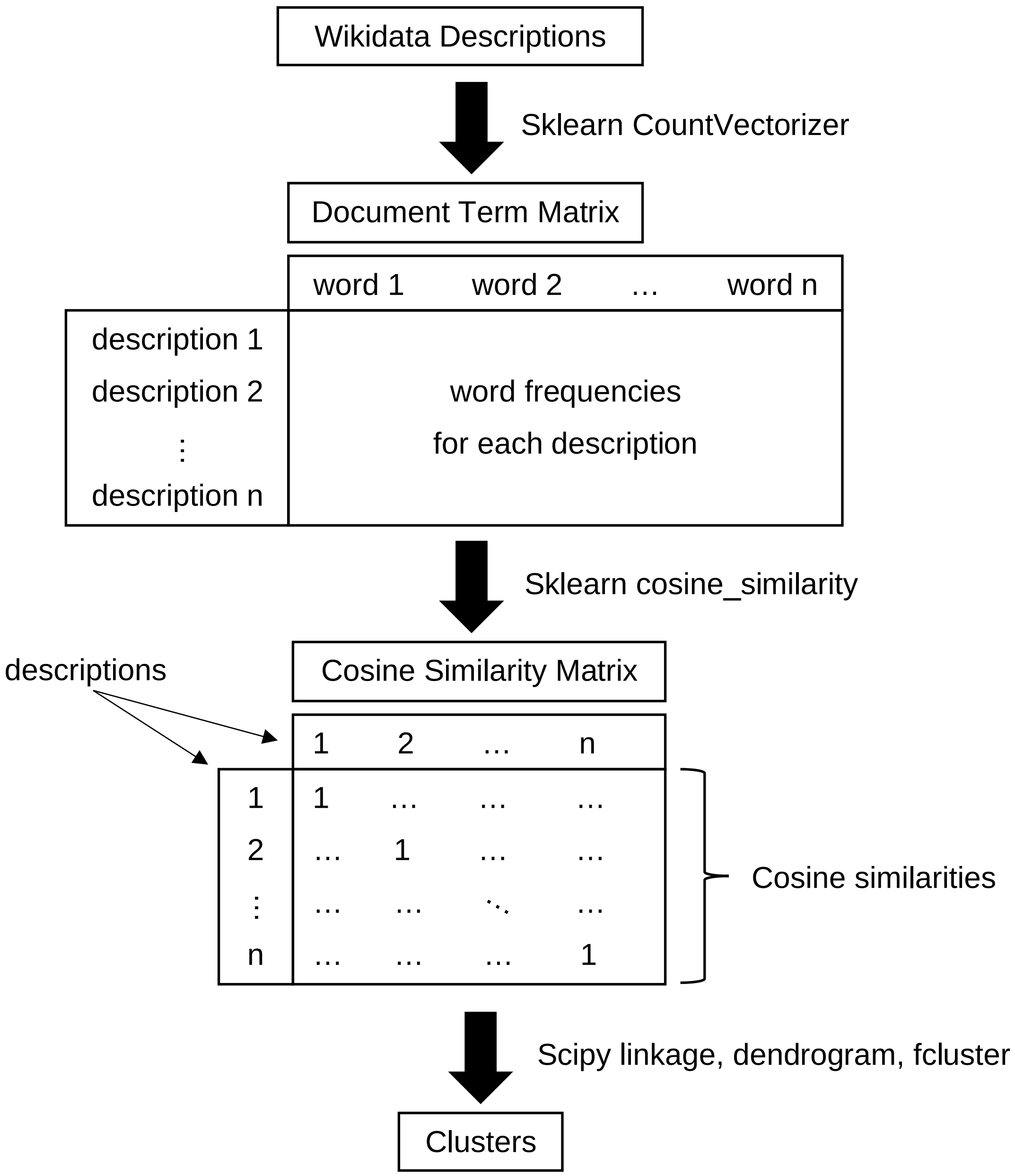
Figure 1. [English Wikidata Names and Descriptions for Starbucks] [Screenshot]. (n.d.). Retrieved from https://www.wikidata.org/wiki/Q37158

Figure 9. [Wikidata Statements for Starbucks] [Screenshot]. (n.d.). Retrieved from https://www.wikidata.org/wiki/Q37158

Scipy. (2020, November 4). scipy.cluster.hierarchy.linkage. SciPy. https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

Swiss Chalet. (2020, November 16). Retrieved from https://www.wikidata.org/wiki/Q2372909

Appendix A – Flow Chart for Wikidata Description Analysis



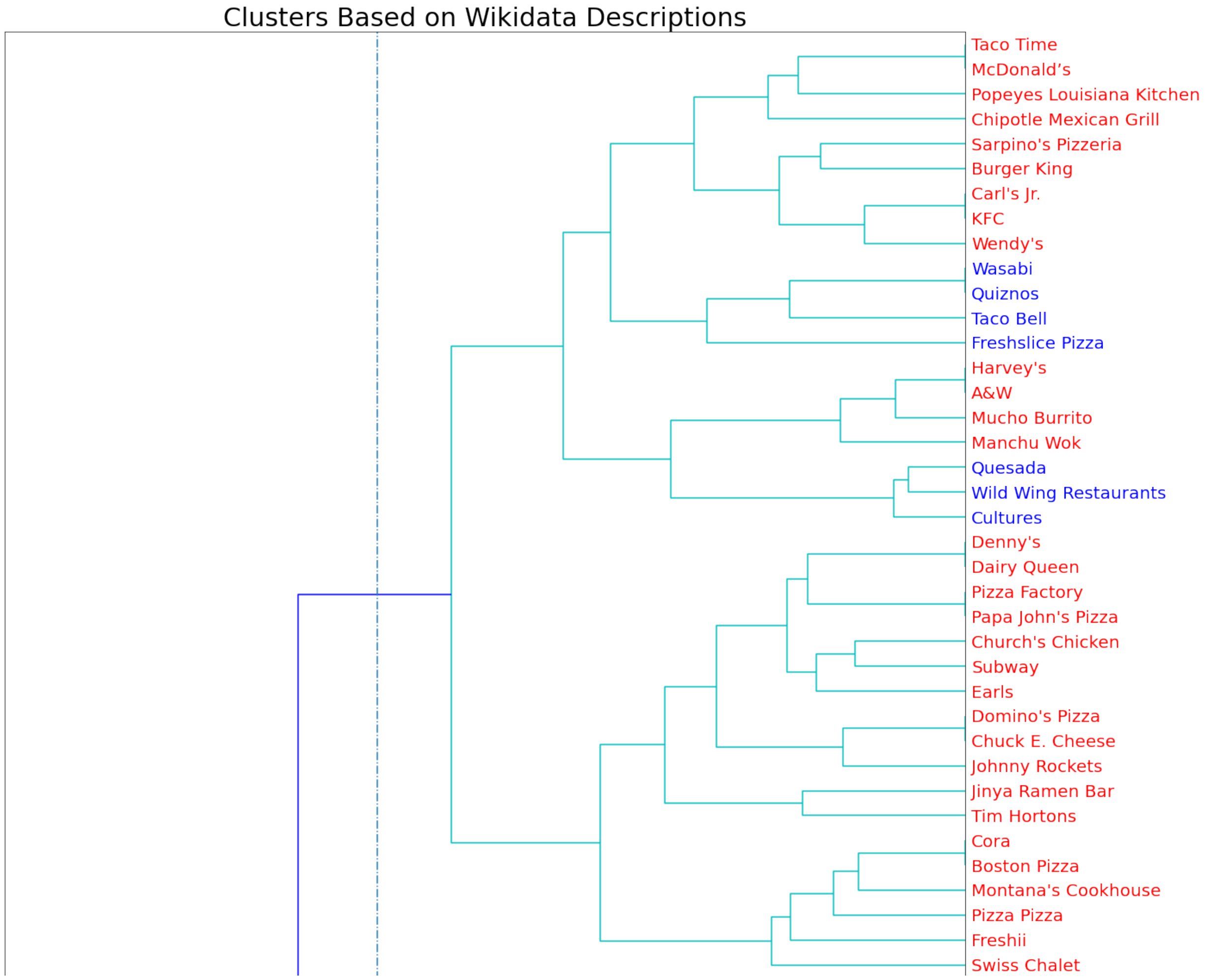
	word 1	word 2	…	word n
description 1 description 2 ⋮ description n	word frequencies for each description			

	1	2	…	n
1	1	…	…	…
2	…	1	…	…
⋮	…	…	⋱	…
n	…	…	…	1

Note: Since the cosine similarity of two documents ranges from 0 (no word matches) to 1 (identical), our linkage uses the argument method = ‘complete’ which is “also known [as] the Farthest Point Algorithm or Voor Hees Algorithm” according to Scipy (2020) to combine descriptions with the highest cosine similarities into clusters first.

Appendix B – Sample Cluster

Cluster containing most locations with Wikidata descriptions which contain the words ‘chain’ and ‘restaurant’. Locations with Wikidata descriptions that contain both words are shown in red. Refer to identify_chain_restaurants.ipynb for the full dendrogram.


Clusters Based on Wikidata Descriptions
Taco Time
McDonald’s
Popeyes Louisiana Kitchen
Chipotle Mexican Grill
Sarpino's Pizzeria
Burger King
Carl's Jr.
KFC
Wendy's
Wasabi
Quiznos
Taco Bell
Freshslice Pizza
Harvey's
A&W
Mucho Burrito
Manchu Wok
Quesada
Wild Wing Restaurants
Cultures
Denny's
Dairy Queen
Pizza Factory
Papa John's Pizza
Church's Chicken
Subway
Earls
Domino's Pizza
Chuck E. Cheese
Johnny Rockets
Jinya Ramen Bar
Tim Hortons
Cora
Boston Pizza
Montana's Cookhouse
Pizza Pizza
Freshii
Swiss Chalet