CS410 Final Project Presentation

An analysis of common restaurants mentioned in the UIUC subreddit given a specific query

How to Install

The main code for our project is located in a Python Notebook file (CS410FinalProject.ipynb). To use the software, it is sufficient to open the file in Google Colab and run the cells.

Dependencies

This project makes extensive use of Python libraries, which are set up in Part 1 of the project. If any of the libraries fail to import or have not yet downloaded, you may have to add

```
!pip install [library]
```

to the first cell of the notebook.

How to Use

The data we used for our project is located in the file: reddit_data_big.json, which is included in the repository. In order to use this data, the user should upload the file to Google Drive and update the path in the first cell of Part 2 to the location of their file.

```
# Load data
f = open('/content/drive/MyDrive/CS410 Final Project/reddit_data_big.json')
data = json.load(f)
f.close()
```

Reddit Scraping

We used PRAW (Python Reddit API Wrapper) to pull posts from r/UIUC that mentioned key words like “restaurant”, “food”, “eating out”, etc.

We then translated that data to JSON so it could be easily used and manipulated in our algorithm.*

*To rerun the scraping script, run the file “cs410_streamredditdata.py”. Note that this file takes a while to run due to the volume of posts.

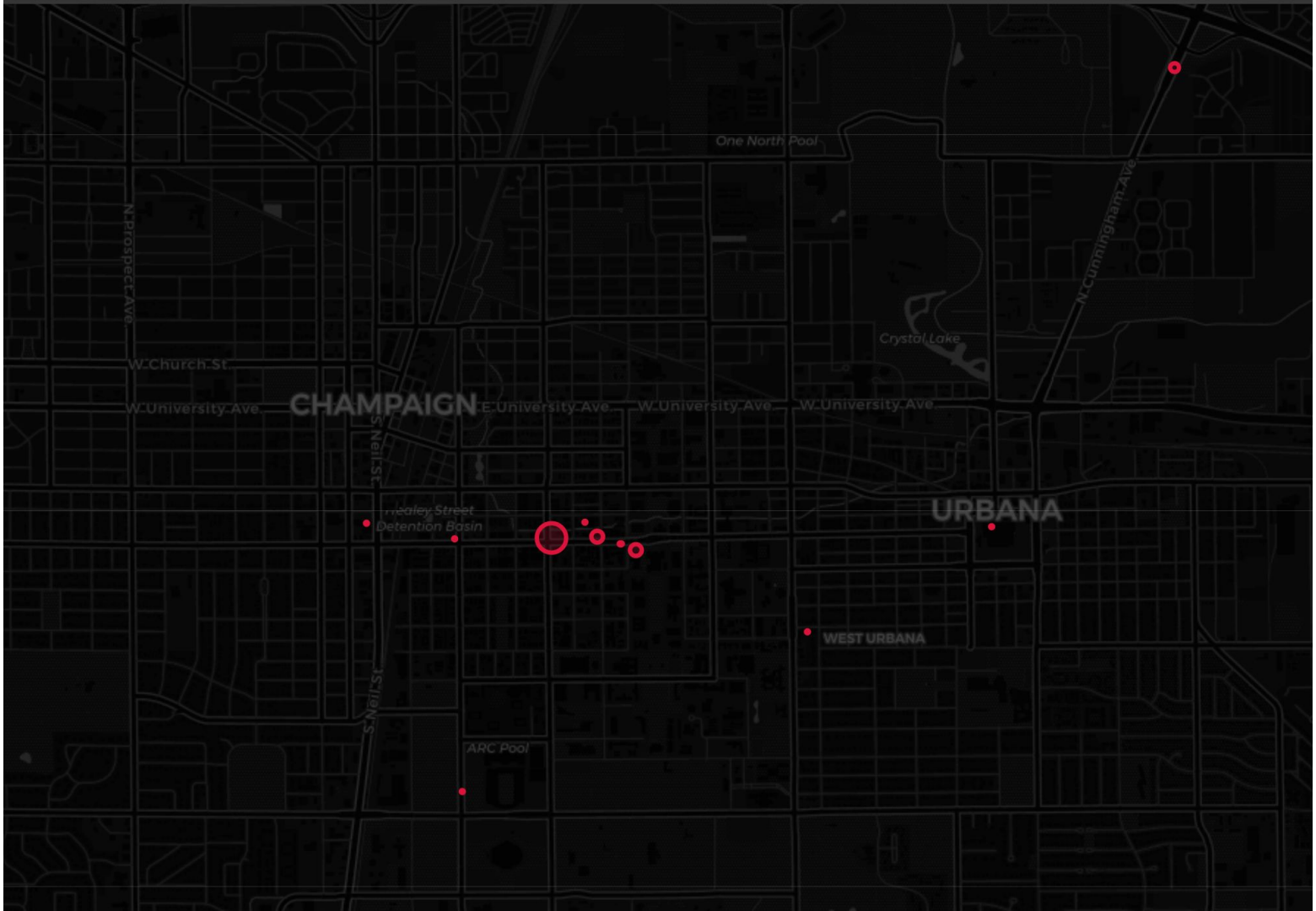
Change Parameters

The user can change the queries used for finding relevant documents by adding to or modifying the ‘queries’ variable in the first cell of Part 3. The user can also change the number of relevant documents used in finding the most common restaurants by modifying the variable ‘N’ in the same cell.

```
# Queries for finding relevant documents
queries = ["mexican restaurants", "chinese restaurants", "good restaurants", "cheap restaurants"]
names = ['mex','chi','good','cheap']
N = 100
```

Results

The program works by ranking the documents by relevance to a query using the BM25 Okapi algorithm. It then extracts the restaurant names from each document by finding the noun phrases that contain a proper noun and classifying those phrases using a Hugging Face Zero-Shot Classification model. Finally, it takes the most common restaurants from the top N most relevant documents and displays them on a map.


One North Pool
N Prospect Ave
N Cunningham Ave
Crystal Lake
W Church St
W University Ave
CHAMPAIGN
E University Ave
W University Ave
W University Ave
S Neil St
Healey Street
Detention Basin
URBANA
WEST URBANA
S Neil St
ARC Pool