Task:

to explore something interesting in neighborhood description.

Idea:

Given an example below, I first explore what information tends to be included in the neighborhood description.

“ Very safe and quiet residential neighborhood, you can walk around anytime day or night. Great access to **Harvard Square** and all it has to offer - **Gourmet restaurants**, **bars & pubs**, **cultural events**, **historical sites**, shopping, entertainment, etc. Easy access to **MIT**, **downtown Boston**, **area hospitals**, all in about 12 - 15 minutes with Uber or Lyft ”

It can be seen that several locations (marked bold) are referred in the sentences. The locations might attract readers most because those words are what they care about. We should include information with reference to these locations if another description is auto generated by our system which describes the nearby property in this neighborhood.

After we extract the locations, we also interest in the descriptive words for this location. For example,

“ Very safe and quiet residential neighborhood, you can walk around anytime day or night. Great access to **Harvard Square** and all it has to offer - **Gourmet restaurants**, **bars & pubs**, **cultural events**, **historical sites**, shopping, entertainment, etc. Easy access to **MIT**, **downtown Boston**, **area hospitals**, all in about 12 - 15 minutes with Uber or Lyft ”

Those words marked wave underline are to describe the locations. We should extract them too. Combined locations with descriptive words, the basic element of a sentence can be generated.

Take Boston for example. Boston has several neighborhoods, e.g. Allston, Brighton, Chinatown. I am going to extract popular locations for every neighborhood in Boston and compare the frequency of word occurrence. I wrote four python source files.

The python library I used is Spacy. (https://spacy.io/docs/) Some features of this library might help my work:

- Tokenization
- pos tagging
- sentence segmentation
- dependency parsing
- entity recognition
- word vector (Glove)
- sentiment analysis

The function of 4 source files:

- generateData.py

paras: [city]
generate a dict (python structure) which contains lists of neighborhood-related descriptions stored in neighborhood keys. Write data of one neighborhood into one file.

- findNameEntites.py

paras: [city] [neighborhood]
generate keywords of one neighborhood of one city. Choose NameEntity as location-related features.
A good visualizer: https://demos.explosion.ai/displacy-ent/

- findAdj.py

paras: [city] [neighbourhood] [keyPlace]
find descriptive words for a specific key place in one neighborhood of one city. I expect to get adj. words, e.g. far, near, but it does not perform as I expected. Still under construction.

- findKeywordsInCity.py

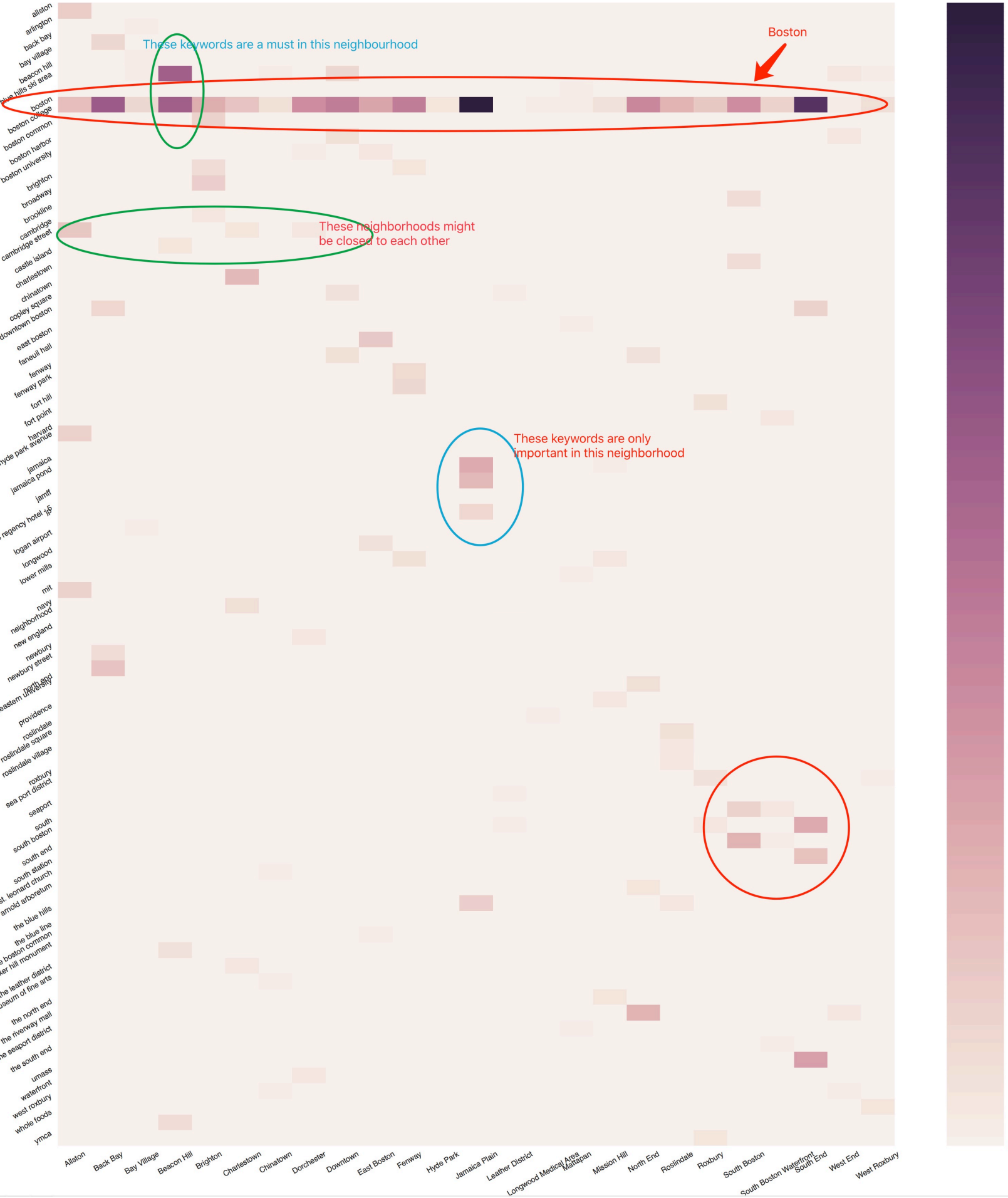
paras: [city]
find all the location-related keywords in one city. Draw a heat map for visualization which shows how frequent a location appears in the description of one neighborhood in one city.

What I can get from the data visualization

- The heat map is sparse. only several locations are frequently appear in all neighborhoods.
- Some locations only appear in one neighborhood. It means that that location must be described in this neighborhood.
- If one location appears in several neighborhoods, it is suggested that those neighborhoods might be closed to each other.
- City name is always the most frequent word.
- It seldom exists many locations which all appear very frequently in one neighborhood. Seems good because we don't have to struggle in selection.

Data Visualization:

see below


These keywords are a must in this neighbourhood
Boston
These neighborhoods might be closed to each other
These keywords are only important in this neighborhood
allston
arlington
back bay
bay village
beacon hill
blue hills ski area
boston
boston college
boston common
boston harbor
boston university
brighton
broadway
brookline
cambridge
cambridge street
castle island
charlestown
chinatown
copley square
downtown boston
east boston
faneuil hall
fenway
fenway park
fort hill
fort point
harvard
hyde park avenue
jamaica
jamaica pond
jamfr
jp
regency hotel - jp
logan airport
longwood
lower mills
mit
navy
neighborhood
new england
newbury
newbury street
north end
northeastern university
providence
roslindale
roslindale square
roslindale village
roxbury
sea port district
seaport
south
south boston
south end
south station
st. leonard church
arnold arboretum
the blue hills
the blue line
the boston common
the bunker hill monument
the leather district
the museum of fine arts
the north end
the riverway mall
the seaport district
the south end
umass
waterfront
west roxbury
whole foods
ymca
Allston
Back Bay
Bay Village
Beacon Hill
Brighton
Charlestown
Chinatown
Dorchester
Downtown
East Boston
Fenway
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury
150
120
90
60
30
0