CS 506 FALL 2019 - HW1

Clustering and Visualization

Due date: October 16, 2019

1 Find clusters

- In this assignment we will be working with the AirBnB dataset, that you can also find here. Our goal is to visualize areas of the NYC with respect to the price of the AirBnb listings in those areas.
  From the detailed *nyc_listings.csv* file, you will use **longitude** and **latitude** to cluster closeness and **price** to cluster for expensiveness.
  Note that spatial coordinates and price are in different units, so **you may need to consider scaling** in order to avoid arbitrary skewed results.

a) [8pts.] **Find clusters using the 3 different techniques we discussed in class: k-means++, hierarchical, and GMM**. Explain your data representation and how you determined certain parameters (for example, the number of clusters in k-means++).

A few hints:
-Some listings contain missing values. Better strategy for this assignment is to completely ignore those listings.
-Pay attention to the data type of every column when you read a .csv file and convert them to the appropriate types (e.g. float or integer).

2 Data visualization

a) [1pt.] Start by producing a Heatmap using the Folium package (you can install it using pip). You can use the code below to help you (assumes the use of Pandas Dataframes):

```
def generateBaseMap(default_location=[40.693943,
    -73.985880]):
    base_map = folium.Map(location=default_location)
    return base_map
```

```
base_map = generateBaseMap()
HeatMap(data=df[['latitude', 'longitude', 'price']].
    groupby(['latitude', 'longitude']).mean().
    reset_index().values.tolist(), radius=8, max_zoom
    =13).add_to(base_map)
base_map.save('index.html')
```

Is this heatmap useful in order to draw conclusions about the expressiveness of areas within NYC? If not, why?

b) [2pts.] Visualize the clusters by plotting the longitude/latitude of every listing in a scatter plot.

c) [2pts.] For every cluster report the average price of the listings within this cluster.

d) Bonus points [1pt.] if you provide a plot on an actual NYC map! You may use Folium or any other package for this.

e) [1pt.] Are the findings in agreement with what you have in mind about the cost of living for neighborhoods in NYC? If you are unfamiliar with NYC, you can consult the web.