Homework-2-2

March 16, 2017

1 Clustering Yelp Restaurants

Part 2: To be completed INDIVIDUALLY and due on March 24 at 3pm.

In this assignment, we will continue to work with the Yelp dataset that we used in Homework 2-1.

We will continue to try to find culinary districts in Las Vegas.

(As a reminder from last time, these are characterized by closeness and similarity of restaurants. Use the "longitude" and "latitude" to cluster closeness.)

However, in this analysis we will not use the Yelp-supplied "categories" to cluster for similarity as we did in Part 1.

Instead we will cluster the reviews themselves, extracting categories in an unsupervised fashion.

Specifically, you are to use PCA/SVD on the Yelp reviews to cluster restaurants based off on their reviews. As a reminder, LSA consists of using PCA applied to the document-term matrix.

(20 pts)

```
In [ ]:
```

Find clusters using the 3 different techniques we discussed in class: k-means++, hierarchical, and GMM. Visualize the clusters by plotting the longitude/latitude of the restaurants in a scatter plot and label each cluster.

Note that to label each cluster, you will need to think about how to extract labels from the LSA results. **(25 pts)**

```
In [ ]:
```

Compare your clusters with the results you obtained in Part 1. Use cluster comparison metrics, and also comment on which clustering appears (from your inspection of the clusters) to be more informative, and why. **(15 pts)**

```
In [ ]:
```
