City Clustering from Craigslist Posts

Charles Johnson and Michael Kim, *Stanford University - CS 229*

Project Milestone - 16 November 2012

We aim to cluster cities based on postings from the classifieds website Craigslist. Our results will provide a metric for evaluating the similarity between cities based on the posting content on Craigslist, a community driven marketplace. Of course, the clustering will be agnostic to geographic location so as to determine a more nuanced indicator of similarity. Here, we report our initial results using k-means clustering on a preliminary dataset with a preliminary feature space. We will also discuss our plans for future work to improve the feature space and the results of the clustering algorithm.

Current Work

In this section, we summarize the work completed so far in building our data set, determining features, and clustering cities.

Data Collection

To obtain data from Craigslist, we elected to use the 3taps API (`3taps.com`), a RESTful API for querying up-to-date and historical posts from all of Craigslist in a JSON format. The JSON response includes source data from the posts such as post titles, post bodies, and post categories, as well as additional annotations such as location codes at varying granularity. From these data, we also calculated metadata about the collections of posts associated with a given location. For our preliminary analysis, we collected a dataset of the 11,000 most recent posts at the time of collection. While many millions of posts are available, we started with this more manageable dataset to develop and iterate our algorithms quickly on a reasonably representative sample space. The discussion herein refers to this dataset exclusively.

Originally, we had hoped to cluster neighborhoods using these posts, but the annotated neighborhood codes were clearly inadequate for proper analysis. A quick descriptive analysis showed that there were only 198 unique neighborhood codes, and 54% of posts did not have a neighborhood code. In contrast, there were 399 unique city codes, and only 5.9% of posts were missing a city code. In light of this discovery, we decided to group posts by the city they originated from, based on the city code provided. Additionally, the dataset happened to only contain city codes from the US.

Feature Selection

One of our main goals for the milestone was to determine what features we could reasonably pull out of the data and how effective they would be at clustering cities. Initially, we segregated the dataset of posts into groups based on the city code and calculated a few statistics on these data to use as features in our city vector. Specifically, we computed the frequency of posting, the average title length (in characters), the average body length (in characters), and the average number of images per post. We created a 4-dimensional feature vector with these metadata for each of the 399 US cities.

Instead of using these raw features, we normalized them in order to weight the features equally against one another. Otherwise, a feature which is naturally larger (like post body length compared to post title length) would artificially create larger distances between vectors. As a first pass, we simply normalized all values of a feature by the maximum observed value for that feature, moving the values of all features onto a $[0, 1]$ scale. While performing

k-means clustering we experimented with using different weights for each of these features, which showed no significant effects.

Eventually, we intend to pull in some aspects of analyzing the natural text of the posts, but we wanted to first evaluate how a model using simple features performed in clustering cities. We will discuss other features we intend to investigate in the section on Future Plans.

Clustering Results

After building up a vector of metadata features for each of the 399 cities, we then performed k-means clustering on these input vectors for various values of k. Overall, the results of the clustering were qualitatively poor. Below we analyze different aspects of the initial results in order to motivate future directions.

A natural initial question in running the k-means algorithm is what value to use for k. Ideally, if there are natural clusters, there will be an ideal value to use for k. A good measure to evaluate the quality of a given clustering is the average euclidean distance between a feature vector and its nearest cluster centroid. We ran the clustering algorithm on the input data for various values of k producing the plot in Figure 1.

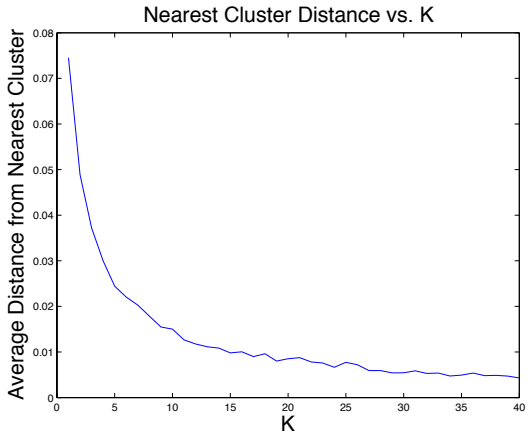

Figure 1: *Average Euclidean Distance from Nearest Centroids.*

The results in Figure 1 show that there is a relatively smooth function depicting the decline in the average distance to the nearest centroid as the value of k increases, which is a very natural result. However, if there is a natural set of clusters among the feature data, we might expect there to be a non-smooth, sudden decrease in the average distance as the value of k approaches the natural number of clusters. From the plot we do not see any such behavior, and so initially there is no obvious choice for value for k.

In order to qualitatively analyze the clustering of the feature data, we performed principal component analysis on the feature vectors so that we could plot them in two dimensions. In visually analyzing the clustering for various values of k, the clustered data in PCA seemed most compelling when clustered around 5 centroids.

Figure 2 shows the centroid assignments when clustering the original 4-dimensional feature vectors. The feature vectors are plotted according to their first and second principal components, and the colors correspond to the centroid assignments of the 4-dimensional vector. The clustering seems to

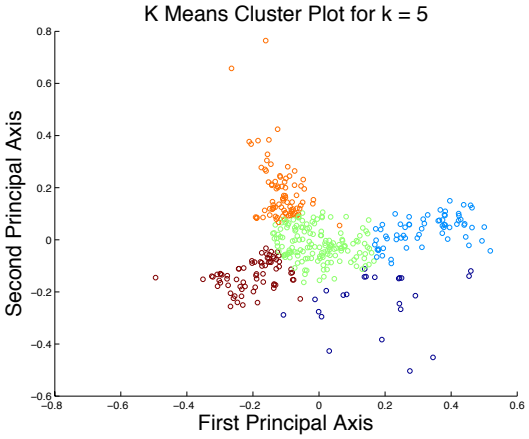

Figure 2: *5-means clustering plotted with PCA.*

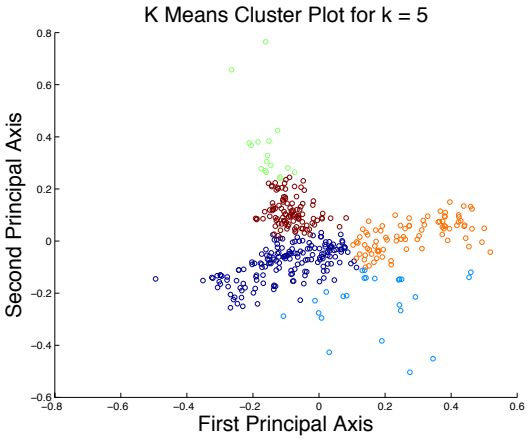

Figure 3: *5-means clustering second iteration.*

have reasonably separated the data into 5 clusters,

but clearly the separation of these clusters is not compelling. Furthermore, additional iterations of 5-means clustering shows that the algorithm does not consistenly converge on the same centroids and assignments, as shown in Figure 3, which is further evidence that the clustering is not natural nor compelling.

One hypothesis to explain the low initial success would be that only some of the 4 features provide any separability to the data. To explore this we experimented with 5-means clustering the input data for only one input feature at a time. The resulting clusters, shown in Figure 4, proved similar to the 4-dimensional case in that the data is relatively inseparable, forming no natural clusters.

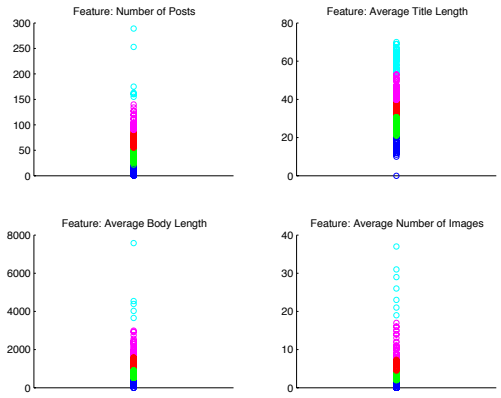

Figure 4: *5-means clustering on each feature.*

From the analysis, it seems that none of the input features are particularly distinct nor useful in clustering the cities into disparate groups. The most likely underlying explanation for this poor clustering performance is precisely because we analyzed metadata for which there is no intuitive geographic correspondence. That is, there is no intuitive reason to believe that some group of cities might have significantly shorter or longer posts than the others, for example. In fact, all of the input features were distributed roughly as a Gaussian, which is a distribution not so conducive to clustering.

Future Plans

Our primary focus in our future work will be in selecting more indicative features, but we will also strive for new ways to improve the performance of the clustering on the imput data for any given feature set. Here, we specifically discuss our plans to improve our selection of features.

Feature Selection

It is clear that the metadata features that we used to cluster cities are not differentiating collections of posts enough to provide obvious clusters. Moving forward, we intend to test the performance of other features more closely related to the post data rather than data about sets of posts. In particular, the categorical information of the posts, prices of listings, and the natural language are three compelling features that may intuitively relate to characters of cities. For example, perhaps cities with higher average incomes post listings at higher prices, or perhaps cities with lower literacy have more typos.

We intend to further organize the posts by their categories. This will allow us to create a few new features. First, the number of independent categories used, which seems like it would naturally represent a city's degree of engagement with Craigslist (and thus, perhaps more generally technology/the internet in general). Beyond this we could look at the percentages of posts in specific categories such as housing or specific job postings.

We intend also to experiment with adding average price of a listing as a feature. Looking further into the posts themselves, we could also pull out average listing prices in different categories. This feature will be straight forward to implement once we have posts separated by category and will hopefully provide an indication of the cost of living per city.

Furthermore and perhaps most interesting, we intend to start analyzing the content of the post titles and bodies. Hopefully, we will be able to identify features of the natural language that will provide more distinction between regions. There are many ways to analyze and represent the natural language features, and so we intend to spend a significant effort in researching and experimenting existing methods.

Finally, after a push to expand our feature space and given adequate time, we will also experiment with feature selection algorithms for determing the most informative features to provide the most compelling clustering.