**Recommending locations and categories relative to business requirements.**

**Team Members**:

1. Suyog Swami
2. Vineet Nair

**Objective and Tasks:**

* To recommend business categories for a given location requirement and suggest best Working Hours:

We are using K-means clustering method to make clusters of Yelp Dataset. After generating the clusters we will calculate the centroid for each cluster. This centroid will further be used to compare with user specified Location. Next we will use this cluster to find out groups in the cluster by business categories. For each group we will find out best working hours i.e. the working hours with highest frequency. Based on the number of records in each group of Business Category we will output the business categories in descending order along with their best working hours.

* To recommend locations for a given business category input and suggest best area for the business in the suggested location:

We are using K-means clustering method to make clusters of Yelp Dataset. After generating the clusters we will calculate the centroid for each cluster. This centroid will further be used to compare with user specified Business Category. Next we will use this cluster to find out groups in the cluster by Locations. For each group we will find out best area i.e. street name with highest frequency. Based on the number of records in each group of we will output the locations in descending order along with their best area.

**Deliverables:**

* We have converted the dataset file from json to csv format. This provides us ease during performing data mining tasks.
* We have performed K-means clustering on the business dataset and calculated the centroid for each cluster.

**Challenges:**

Following are the challenges faced during the project implementation.

1. Finding the appropriate value for K for K-means clustering so that dataset is properly partitioned without much overlapping.
2. Matching user query with appropriate attribute value in dataset. E.g. If user enters “schezwan” as business category then we should be able to map “schezwan” with “Chinese Restaurant”.

Accurately recommend locations for a given business category which are nearby the users specified location. E.g. If user specifies Arlington as the location the resulting categories of businesses should be from nearby cities like Irving, Plano, Dallas etc.

**Algorithm:**

Given an initial set of *k* means *m*1(1),…,*mk*(1) , the algorithm proceeds by alternating between two steps[1]:

**Assignment step**: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared [Euclidean distance](http://en.wikipedia.org/wiki/Euclidean_distance), this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

S_i^{(t)} = \big \{ x_p : \big \| x_p - m^{(t)}_i \big \|^2 \le \big \| x_p - m^{(t)}_j \big \|^2 \ \forall j, 1 \le j \le k \big\},

where each x_p is assigned to exactly one S^{(t)}, even if it could be assigned to two or more of them.

**Update step**: Calculate the new means to be the [centroids](http://en.wikipedia.org/wiki/Centroids) of the observations in the new clusters.

m^{(t+1)}_i = \frac{1}{|S^{(t)}_i|} \sum_{x_j \in S^{(t)}_i} x_j 

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitionings, the algorithm must converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm.

The algorithm is often presented as assigning objects to the nearest cluster by distance. The standard algorithm aims at minimizing the WCSS objective, and thus assigns by "least sum of squares", which is exactly equivalent to assigning by the smallest Euclidean distance.

The algorithm is composed of the following steps [2]:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

**Initial Implementation:**

Dataset file was converted to csv format. This eased the implementation of K-means clustering we implemented in the next step. We have performed K-means clustering on the business dataset and calculated the centroid for each cluster. The centroid was used to match with the user specified query to find the target cluster.

**Evaluation Plan:**

Estimating the value of K for the K-means clustering is one of our main points of evaluation. This is important because we need to divide yelp business dataset into distinct clusters.

**Change of Plan:**

We decided to drop the plan of using longitude and latitude for comparing the locations because we decide to consider broader area of user specified location rather than concentrating on a well specified address. We were able to perform same tasks without using longitude and latitude and provide user with much more options to consider.

**Difficulties encountered:**

We were not able to finalize on a specific clustering technique for this task. We researched and found out that K-means would be appropriate for our task.

Also estimating the value of K for the K-means clustering is a challenge. We are working on it to find a value of K such that we partition the Yelp dataset distinctively.

**Tasks to be accomplished:**

After finding the appropriate cluster for our task we will group them according to user specification. For the first task we will use this cluster to find out groups in the cluster by business categories. For each group we will find out best working hours i.e. the working hours with highest frequency. Based on the number of records in each group of Business Category we will output the business categories in descending order along with their best working hours. For the second task we will use this cluster to find out groups in the cluster by Locations. For each group we will find out best area i.e. street name with highest frequency. Based on the number of records in each group of we will output the locations in descending order along with their best area.

Representing the obtained result from the cluster analysis to the user in a meaningful way will be our next challenge. We are planning to host a website that will be interactive with the user and represent the results of our project in elegant manner.

**References:**

1. <http://en.wikipedia.org/wiki/K-means_clustering#Standard_algorithm>
2. <http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html>