Status Update for Term Project – Team “Classy Flyers”

Problem:

Given business attributes like Business name, business hours of operation, the reviews for each business (reviews are given by users), we would like to predict the business category.

To simplify the problem, we have taken 10 primary categories to predict, but we would like to explore sub-categories as well later on.

In case of sub-categories, a business unit is likely to be present in multiple categories. In these cases, we would like to use one-vs rest classifier to infer all the possible categories.

What is the baseline model you are choosing?

Baseline-1: We assume each user has a certain prior of reviewing businesses from certain categories. For example, certain users more often go to restaurants or write reviews for restaurants than any other categories. Hence if for a new business a review is written by this user, we tend to believe that this business is of category restaurant (if we don’t know anything else). We would like to capture this information and use this as the baseline.

Baseline-2: We also assume Business Hours of Operation (BHO) will help us in determining the business category as a business in category “Active Life” tends to operate late in the day than the category “Education”. We wanted to capture this information in predicting the category information.

Procedure:

Baseline-1: We calculated the prior probability of each user writing a review for each category. Now for each test business we have a list of users who reviewed the business and aggregated the prior probabilities from each user to predict the most likely category. Since we also know the true categories we will calculate the precision, recall, f-measure.

Baseline-2: We calculated the prior probability of each BHO for each category. Now for each test business, we have a BHO, and based on that we predict the most likely category. Since we also know the true categories we will calculate the precision, recall, f-measure.

What options will be exploring to improve upon the baseline model?

We would like to use reviews for our category prediction. For this, we would be exploring various text classification approaches like SVM, Naive Bayes, K-NN classification algorithms for the same.

Since reviews usually contain spelling mistakes, we want to understand the effect of these mistakes on category prediction. If necessary, we plan to use a spell checker to see if this improves category prediction performance.

We would also like to explore the usefulness of POS tags of the review for category prediction.

Since we are dealing with text and believe that the review text can semantically infer the business category, we would like to explore semantic similarity between the reviews and the category using word2vec.

Finally, we would also like to explore how well we can predict, with only the business name and knowing nothing else (business reviews, BHO etc.)

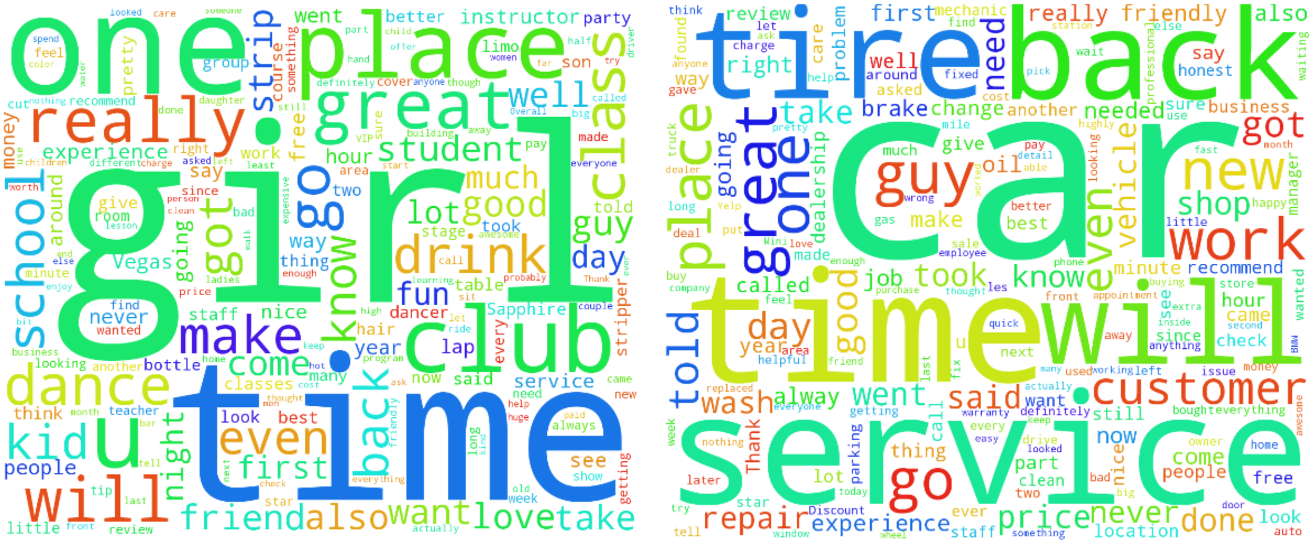
How do you plan to test and compare these models to each other?

We plan to use precision, recall, f1-measures to evaluate different methods. We will also use AUC curve to understand how well various methods perform. We will also use cross-validation to make sure our results are significant.

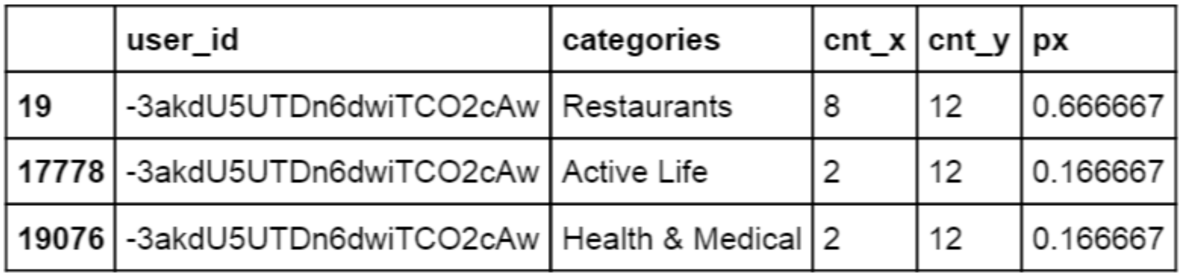
We also plan to understand how many reviews are required to learn a good classifier by producing a learning curve.

Plots and Snapshots:

Word clouds for categories “Education” & “Automotive” from text reviews respectively. We can see that the distribution of the words is very different. Hence we believe text classification form reviews will be very useful.

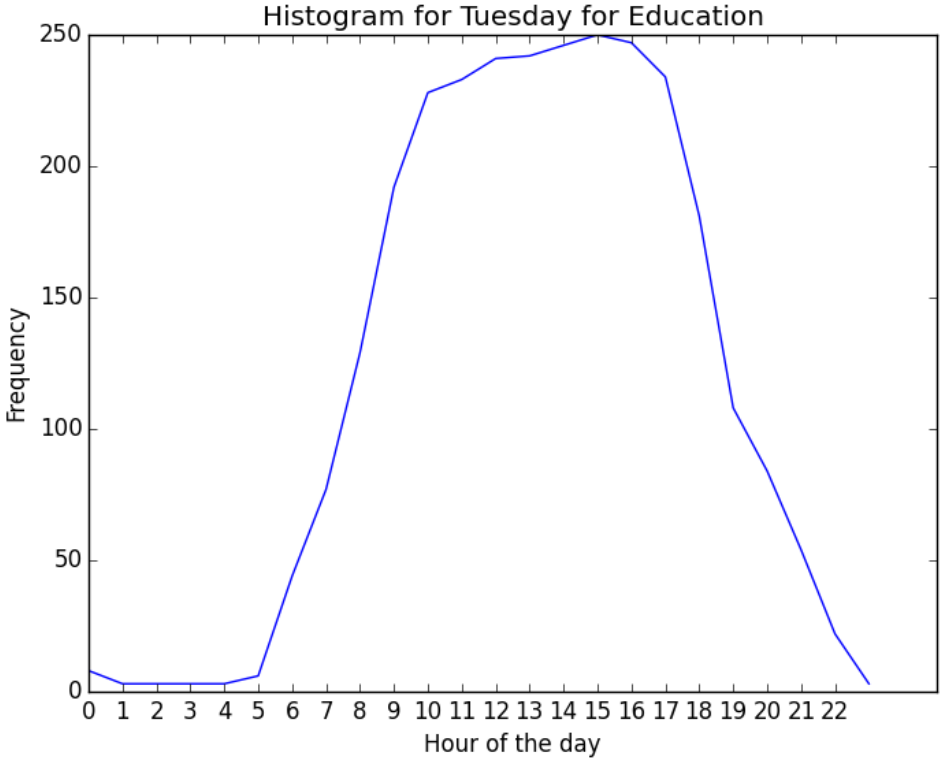


Baseline-1: Category prediction based on user information. Based on the below user-category probability information we understand that this user is most likely to review restaurants.



	user_id	categories	cnt_x	cnt_y	px
19	-3akdU5UTDn6dwiTCO2cAw	Restaurants	8	12	0.666667
17778	-3akdU5UTDn6dwiTCO2cAw	Active Life	2	12	0.166667
19076	-3akdU5UTDn6dwiTCO2cAw	Health & Medical	2	12	0.166667

Baseline-2: Category prediction based on BOH. Below plot is a graph of the category "Education" and the distribution of business hours of operation on "Tuesday".


Histogram for Tuesday for Education
Frequency
Hour of the day