**COS 424 Project Proposal** 4/6/2012

*Members*: Stephen Beard, Joey Dodds, Soumyadeep Ghosh, Matt Zoufaly

*Project Name*: *Nightout*

**Problem**

Exploring where you are isn’t as easy as it sounds. Word of mouth combined with reviews from places like Google and Yelp help to find good places to eat and fun things to do. However, it would be nice to have a simpler approach: create an application that finds things for you to do based on your interests and tastes.

We propose a rating system for activities in your area. Essentially, an activity can be anything ranging from a place to grab food, a place to go shopping, or a trip to a park. To put it simply, we’re going to combine a rating system from sites like Netflix with rating applications like Yelp.

Fundamentally, people enjoy different types of things. So recommending activities based on what “most people” like (such as Yelp) does not necessarily lead searching users to activities they’ll enjoy. Specific interests / parameters can drastically change the recommendation. Thus we plan to make recommendations based on ratings of similar types of activities.

**Data**

Ideally, we would be able to populate our databases with real user data. To this end, we’re creating a website that is slowly being populated. In order to evaluate our model, we can leverage existing ratings and information in internet databases (including Yelp, Google and FourSquare) to “seed” our databases with ratings. In order to do this, we’re going to create a web-crawler that will scrape businesses and user ratings from existing internet database. For each internet database, we will translate their rating system to ours (we intend to use a like / dislike approach, but an X star rating system is also something we might explore) and insert the data as if we had collected it ourselves.[[1]](#footnote-1) This will be customized for each database, but since there are only a couple of possible rating systems (and most use a 5-star approach), this is fairly simple. We can evaluate on, essentially, an arbitrary number of users / businesses. To start, we’ll evaluate on several dozen users / businesses, refine the model, and then continually expand until we feel the results are indicative of a realistic workload (we estimate this will be around several thousand users and businesses).

**Methods**

**Evaluation**

To evaluate our project, we will employ the methods discussed in class. More concretely, we plan to keep 20% of the dataset as a testing set. The model will then be fit using five fold cross validation on the remaining 80% of the data.

The natural measure of success is how well we can predict the held out users ratings of the various places and activities in the data set.

**Contingency Plan**

If scraping the various sites proves to be too difficult, we have a few options for gathering the data.

* We can generate data by creating classes of people with various levels of ratings for different business types, then generate a dataset by randomly selecting a user class and filling in ratings by adding a Gaussian term to the base rating level.
* Generate the data using real people (friends, families, victims) or potentially something like Amazon’s Mechanical Turk

1. Note, this is *only* for testing the validity of the model. This would not be legal in a live web application [↑](#footnote-ref-1)