October 27, 2013

Problem

The problem I've selected for my final project of Data Science course is Yelp-recruiting challenge from Kaggle's competition. Yelp was created to find to help consumers find great local businesses. Yelp is an online urban city guide that helps people find cool places to eat, shop, drink, relax and play, based on the informed opinions of a vibrant and active community of locals in the know.

Yelp has a very large community of users who write reviews about any local business, place or service ranging from restaurants to bars, shops to salons, spas to dentists and so on. Users can vote on reviews typically in three categories – Useful, Funny or Cool. The goal of this competition is to estimate the number of useful votes a review will receive without even waiting for the community to vote. It's expected that the solution of the problem should automatically elevate the high quality review with its freshness intact.

Data

The data provided consists two datasets training and test as similar to any prediction problem. Each dataset contains four files for businesses, users, reviews & check-ins and each file is composed of a single object type, one JSON object per line. The training data was recorded on 2013-01-19 whereas the test data is collected from the period between 2013-01-19 and 2013-03-12.

Evaluation

The prescribed model should predict, for each review from the dataset, the number of useful votes made at a specific point in time. The training data contains reviews with votes measured at time=2013-01-19. The testing data contains reviews with votes measured at time=2013-03-12. The dataset contains example reviews with the number of votes they've received, as well as additional information about the business and users. Root Mean Squared Logarithmic Error ("RMSLE") would be the evaluation criteria to measure the accuracy of an algorithm.

Approach and Challenges

Having a quick look at the data and its structure, it's clear that this problem requires text based analytics and techniques as the prediction is based on categorical variables. There's also geographic coordinates about the business addresses though at the moment I'm not sure how does geography matter much in this problem. I might start with visualizing each dataset and look for features that are representative of the problem. For modelling, I might well start with regression by factorizing categorical variables and see how good fit it would be.

In terms of challenges, looks like it would require a bit of pre-processing work like parsing JSON objects and dealing with dates & timestamps, in any case Python (pandas, numpy, and scikit) should be in rescue. Also I expect some computational limitations due to size of the data, need to figure out some best ways of memory management while training models.