Homework 2

COMP 7150/8150

FALL 2016

Instructor: Deepak Venugopal

Due Date: October 12, 2016 (submit code, scripts etc. to ecourseware)

1. (40 points) Cassandra Exercise
   Download the devstudio from datastax.com, it contains a utility called Cassandra CQL shell. Use this to design a keyspace for restaurant reviews. The main types of queries that you will encounter are
   - Get all the customer reviews for a specific restaurant-name ordered by date
   - Get all the restaurant-names in a particular zipcode ordered by rating
   - Get all the restaurant-names in a particular zipcode for a specific cusine

   Make a script that we can automatically execute. Specifically, the script should create the keyspace, column families, insert dummy data in the column families and run the 3 queries. You can run any file with CQL sommands using the "source" command in the CQL shell.

2. (60 points) Supervised Machine Learning: Build a spam filter using the provided spamassasin dataset.
   You need to first pre-process the data and remove the http headers from each email message. Get it into a form where for each email, you only have the subject of the email and the body of the email. Then you will extract features for each email by vectorizing each email using two separate methods, counting-based vectorization (sklearn.feature_extraction.text.CountVectorizer) and tfidf-based vectorization (sklearn.feature_extraction.text.TFidfVectorizer). Finally, you will evaluate the performance of spam filtering using the Naïve Bayes as well as logistic regression algorithms using 5-fold cross validation. Specifically, print the average and standard deviation of precision, recall and F1-scores for the following cases.
   - Naïve Bayes (count vectorized and tf-idf vectorized)
   - Logistic regression with L2 regularization with C=1 and C=0.5 (count vectorized and tf-idf vectorized)
   - Logistic regression with L1 regularization with C=1 and C=0.5 (count vectorized and tf-idf vectorized)