CSCI 3907/6907: Introduction to Statistical NLP

# Assignment #3

**Instructions:**

- ✓ This assignment is due on **Friday November 8, 2019, by 11:59 pm**. Refer to course information about late submission policy. Late days are counted from 00:00 a.m. onwards.
- ✓ Submit your solution in a zipped folder through blackboard.
- ✓ All code must compile and run to receive full credit for coding parts.
- ✓ Include citations for any online resources used or group discussions.

## Text Classification

In this assignment, you will use **scikit-learn**, a machine learning toolkit in Python, to implement text classifiers for sentiment analysis. Please read all instructions below carefully.

## Datasets and evaluation:

You are given the following customer reviews dataset: `CR.zip`, which includes positive and negative reviews. CR is a small dataset that doesn't have train/test divisions, so you are required to evaluate the performance using **10-fold cross-validation**. Please use the following scikit-learn modules in your implementation:

## scikit-learn documentation:

Bag-of-words (or ngrams) feature extraction using CountVectorizer:
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Use **binary features** (1/0 rather than counts).

Naïve Bayes classifier:
http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

Logistic Regression classifier:
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Cross validation:
http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html

Classification report:
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

**Question 1:**
Using the scikit-learn modules described above, Implement the following models and report the performance (accuracy and F1) for the CR dataset:

a) [10 points] A Naïve Bayes classifier with add-1 smoothing using **binary** bag-of-words features.

b) [10 points] A Naïve Bayes classifier with add-1 smoothing using binary bag-of-ngrams features (with unigrams and bigrams).

c) [10 points] Logistic Regression classifier with L2 regularization (and default parameters) using binary bag-of-words features.

d) [10 points] Logistic Regression classifier with L2 regularization using binary bag-of-ngrams features (with unigrams and bigrams).

Performance report [10 points].