**Alexandria University**
**Faculty of Engineering**
**Computer and Communications Program**

**Due : Monday 3/29/2019**
**CCE :Pattern Recognition**

# Assignment#2 Classification for NLP Tasks (Total 150Points)

**Submit a report and the codes used. Report should detail and illustrate every step in the assignment. Report will worth (25 points).**

# Problem Statement

We intend to perform NLP task. We will work on the Sentiment Analysis Benchmark on IMDB dataset.

## Movie Review IMDB

The IMDB movie review dataset was first proposed by Maas et al.[1] as a benchmark for sentiment analysis. The dataset consists of 100K IMDB movie reviews and each review has several sentences.
The 100K reviews are divided into three datasets: 25K labeled training instances, 25K labeled test instances and 50K unlabeled training instances. Each review has one label representing the sentiment of it: Positive or Negative. These labels are balanced in both the training and the test set.
The dataset can be downloaded at http://ai.Stanford.edu/amaas/data/sentiment/index.html

### 1. Download Data, Apply Text Pre-processing (30 points)

Text pre-processing is essential for NLP tasks. Your job here will be to apply pre-processing modules such as

a. like stopword removal, tokenization, stemming, lemmatization., etc.
b. Surveying available tools is an important step. Check NLTK.

### 2. Create Vector Spaces and Data Matrix (40 points)

You will need to convert the text of the review after pre-processing into a vector form for further steps. You will need to consider different alternatives for text representation such as traditional methods (BOW, TF-IDF,...) and modern methods such as word embedding based vectors.

a. Understand n-grams.
b. Understand sklearn feature vectorizer (Count, TF-IDF). Make sure to consider n-gram vectorizers.
c. Get familiar with gensim library for generating word embeddings. **Never use test set in the generation of word embedding models.**
d. Understand the fast text wrapper in gensim and use it to generate word embeddings. **Now every word can be a vector. How about the review itself??**

You might use gensim embeddings or pre-trained fasttext models.

## 4. Supervised Learning Step (30 Points)

In this step you will learn multiple of classification models. You will need to tune their parameters**. Every group is required to learn at least 6 models to get the 30 points.** It is also required to **tune** these models parameters on 10% subset of the training set.
Model choices are

a. KNN
b. Decision Trees
c. Random Forests
d. Naive Bayes
e. SVM Linear
f. SVM non-linear
g. Adaboost
h. Logistic Regression
i. Bagging

## 5. Big Picture (25 Points)

You need to compare the performance of the learned models against multiple of factors.
Factors are

a. Pre-processing effect.
b. Features choice.
c. Classifier choice.

**Show success and failure cases. Plots of the evaluation measures should be presented as well.**

## 6. Bonus (20 Points)

Results better than 10% error rate are getting max of 10 points bonus.
Results better than 7% error rate will get 20 points bonus

**GOOD LUCK**