**CS 536/CS 432 – Data Mining**

**Assignment 4**

**Due: May 2 (Tuesday) at 12 midnight**

**Instructions:** (1) You may discuss the assignment with others. However, you MUST do and submit your OWN work. (2) Submit a soft-copy report to the submission folder on LMS. Include report and code needed to reproduce your results.

**1. Text Preprocessing and Descriptive Analysis (35 points)**

Download the Movie Review dataset from:

<http://www.cs.cornell.edu/people/pabo/movie-review-data/>. The dataset contains 1,000 positive and 1,000 negative polarity reviews on movies. Perform the following tasks on this dataset:

1. Preprocess the dataset and construct its vocabulary. Report the vocabular size for the following cases:   
   (i) tokenization without any additional preprocessing steps,   
   (ii) after lower-casing, punctuation removal, stopwarod removal,   
   (iii) after   
    (ii) plus stemming,   
   (iv) after   
    (ii) plus **lemmatization**.
2. Construct the TF and TF-IDF term-doc matrix for (a)(iv).
3. Report the top 10 words based on TF and TF-IDF in the positive and neagative class.
4. Compute the discriminative term weights for words in (a)(iv). Then, list the top 10 words based on these weights in teach class.

Use NLTK, RapidMIner, or any other tool of your choice for this exercise. Provide the necessary readme and code to run the tasks.

**2. Dimensionality Reduction for Sentiment Classification (35 points)**

This exercise explores some dimensionality reduction techniques for sentiment classification. Use the TF-IDF TDM from (a)(iv) above as the base case.

1. Segment the dataset into 60% training and 40% test sets ensuring that the distrution of positive/negative remains approximately 50-50 in each set.
2. Use LSI/SVD to reduce the dimensions to 50%, 25%, 10%, and 5% of the original dimensions.
3. Learn a logistic regression classifier on the training set and test it on the test set. Report the accuracy, false positive rate, and false negative rate of the predictions.
4. Repeat (b) and (c), but this time using random projections.

Perform this task using RapidMIner or MATLAB. Provide the necessary info to re-run your experiments.

**3. Topic Modeling (30 points)**

Develop topic models to understand document collections. Here we will use LDA and its implementation in the MATLAB Topic Modeling Toolbox <http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm>. Download the TIME news article dataset from

<https://pern-my.sharepoint.com/personal/akarim_lums_edu_pk/Documents/Datasets/TIME%20dataset-20170417T074809Z-001.zip?web=1>

and perform the follwng tasks:

1. Perform topic modeling with number of topics equal to 2, 5, 10, 50, and 100. Report the top 5 words from each topic.
2. Take a few sample articles and report their topic distribution.

**4. Lexicon-based Sentiment Classification (30 points)**

This exercise uses an unsupervised lexicon-based sentiment classifier to understand the sentiments expressed in short texts. We will use SentiStrength (<http://sentistrength.wlv.ac.uk/>). Register and download the code/data and run some examples. Then do the following:

1. Parse the Movie Review dataset into sentences, i.e., each review will now consist of one or more sentences with a polarity label for the entire review.
2. Apply SentiStrength on each sentence. Report the distribution of strengths of sentences in the dataset.
3. For each review, average the strengths of its sentences, and label the review by thresholding the average. Report the accuracy, false positive rate, and fale negative rate of the predictions.