**ITSCM 414 Mining Unstructured Data**

**Section 1, Spring 2020**

Prof. Alana Platt

**Lab 1: Text Tokenizing & Normalization**

This lab is due Thursday 3/5/20 at 11:59 pm. Be sure to submit your code, all output files, and lab document to Canvas.

In this lab you will use the corpus you created from Lab 0 to practice your skills tokenizing and normalizing text.

Step 1: Tokenizing and basic processing (50 point):

Write a program that read in each of the 3 files you collected, and then perform the following processing steps:

* Sentence tokenization
* Word tokenization
* Remove special characters
* Expand contractions
* Turn all text to lower case
* Correct repeating characters

Output the text to files. Save the files, and then create 3 new word clouds.

Step 2: Reflection A (15 points)

1. What are your impressions of the raw text generated from Step 1? What do you think was helpful? What do you think had little effect (or perhaps a negative effect)?
2. Compare and contrast your word clouds from Step 1 to Lab 0.

Step 3: Stemming (20 points)

Add stemming to your pre-processing code from Step 1. It is your choice which stemmer to use. Output the text to files.. Save the files, and then create 3 new word clouds.

Step 4: Reflection B (15 points)

1. What are your impressions of the raw text generated from Step 2? What kind of impact did it have on your text?
2. Compare and contrast your word clouds from Step 2 to Step 1.