**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.

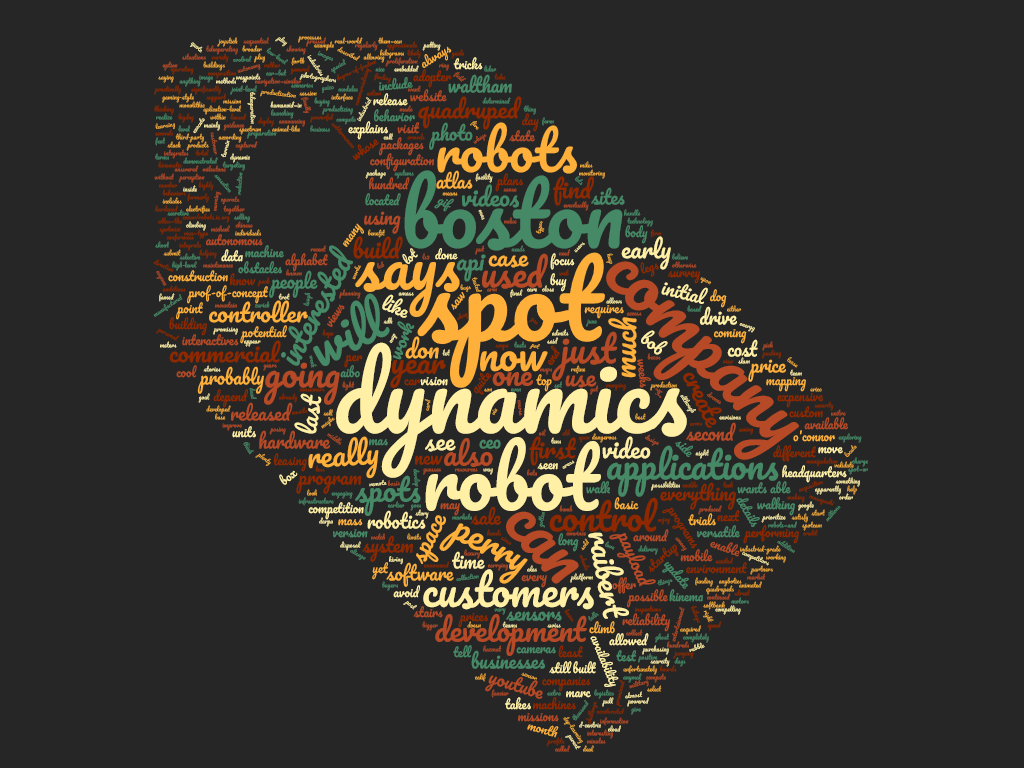
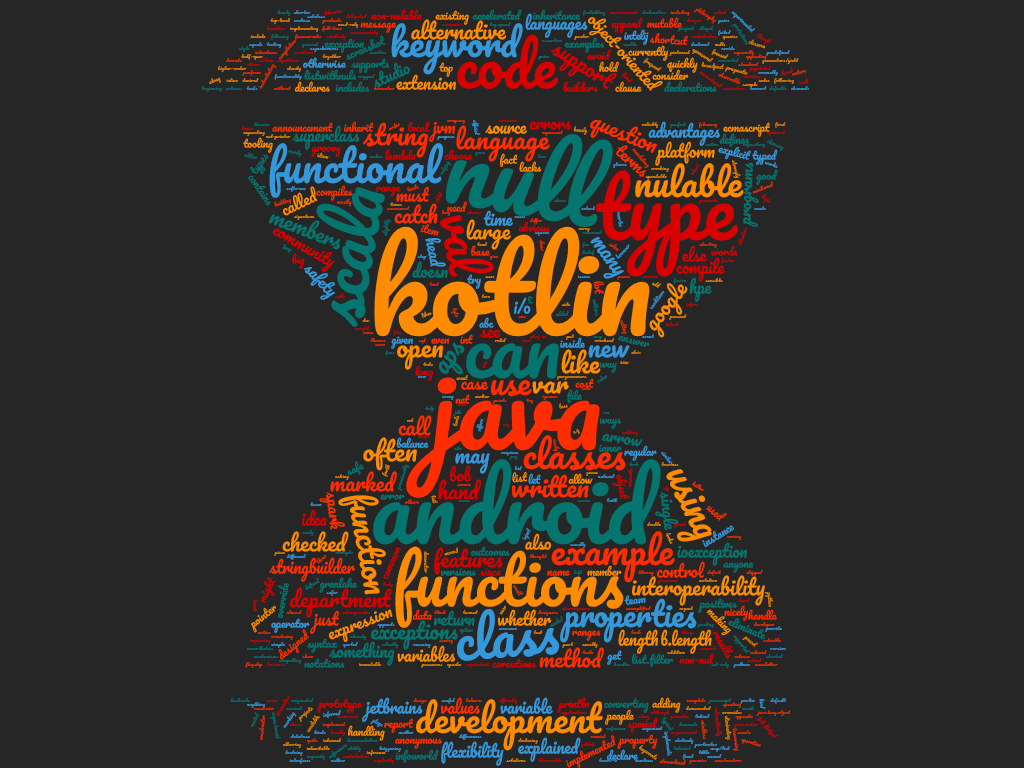
In this lab you will use the corpus you created from Lab 0 to practice your skills tokenizing and normalizing text.

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)?

This process will definitely make text easier to analyze in the future. By removing commas, periods, special characters, and ensuring everything is the same case and not contracted, it allows for a greater analysis of what is really in the text without having to account for the writing style. Some things I noticed had little effect were the contractions. The provided contraction map is pretty small, and some entire articles got through without triggering the replacement method. I was actually unsure as to why the method was not getting called until I realized the map was just too small. Additionally, the method which removes repeating characters removed the second ‘n’ from danny, which makes for an odd analysis considering danny is the main focus of the article.

1. Compare and contrast your word clouds from Step 1 to Lab 0.

Not a huge difference. Some contractions are expanded. Everything is the same case and that made a pretty big difference in the main focus words of articles. “Boston Dynamics” became “boston dynamics.” The word cloud website already did a lot of handling for special characters

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?

Huge difference. Just about every word of the text is different. It looks like someone with marbles in their mouth attempted to use a speech-to-text service. As the word “stemming” describes, only the stems of the words are left behind. Why use many word when few do.

1. Compare and contrast your word clouds from Step 2 to Step 1.

Also a huge difference. These are just silly to look at. None of the words make much sense. I could understand using stemming for analytics purposes, but for an artistic purpose like wordclouds, there is not much a use.