Assignment I – CS 4803 Social Computing (Fall 2015) | Instructor: Munmun De Choudhury

Due: October 28, 2015; 3:35pm on T-Square
Materials (scripts and associated resources) need to be downloaded as a zipped folder from T-Square under Assignment I
Deliverable: A report (pdf document) including the items listed below alongside the questions

This assignment involves doing some fundamental analyses around your personal data. Download any one of your personal data of choice (taking a sample is fine, but the least number of items, that is, posts, tweets, Facebook statuses, WhatsApp messages or emails you need to analyze is 500). Please choose something you would be comfortable sharing the findings with the rest of the class and the instructor. If you do not have any such personal data available, get in touch with me within the next few days to obtain a standard dataset of Reddit posts.

(1) First, use a suitable word cloud generator (for example, but you can choose one of your choice too: http://www.wordle.net/) to examine which words are most prevalent in the text data you are analyzing. In your report include a snapshot (as an image) of this word cloud. An example is enclosed to the right:

(2) Second, for the five most frequent words (i.e., of biggest size in the word cloud), find two example pieces of text (i.e., post, tweet or whichever data item you are analyzing) where the word is used. You can pick any examples you like as long as it gets the point across. In the report you need to show this as a list of the five most frequent words, with sub-bullets showing the two example pieces of text for each word:

Word 1	Example text I
	Example text II
Word 2	Example text I
	Example text II

(3) Third, on this data, use the script provided to obtain different LIWC category measures (getLIWC.py). As you have read in the papers covered in class, LIWC is a psycholinguistic lexicon of word categories (http://liwc.wpengine.com/) and is extensively used and validated in the social computing literature to identify different emotional, cognitive and information attributes in written text, including that in social media. In your report include the average values, along with standard deviations[1] of each of these LIWC categories in a table aggregated for all of the text (posts, tweets etc.) in your data.

(4) Fourth, on your data, use the script provided to obtain the lexical density of text (getLexicalDensity.py). It is a measure of how dense the text is – it measures the ratio of content words to grammatical words. Content words are nouns, adjectives, most verbs, and most adverbs. In your report include the average and standard deviation[1] of the lexical density values over all text (posts, tweets etc.) in your data.

(5) Finally, on your data, use the script provided to obtain the automated readability index (ARI) of text (getReadability.py). The ARI is a readability test designed to assess the understandability of a text. Like other popular readability formulas, the ARI formula outputs a number, which approximates the grade level needed to comprehend the text. For example, if the ARI outputs the number 10, this equates to a high school student, ages 15-16 years old; a number 3 means students in 3rd grade (ages 8-9 yrs. old) should be able to comprehend the text. In your report include the average and standard deviation[1] of the ARI values over all text (posts, tweets etc.) in your data.

[1] Definition of standard deviation: https://en.wikipedia.org/wiki/Standard_deviation Use the numpy Python package to get this number: http://docs.scipy.org/doc/numpy/reference/generated/numpy.std.html