Assignment I – CS 8803 Data Analytics for Well-Being

Topic	Linguistic Analysis of Georgia Tech campus well-being
Grade	Max 60 points; 10% of overall grade (late policy applies)
Due	March 7, 2016, 2:05pm Eastern Time
What to hand in	A report with answers to the different questions
Where to submit	T-Square
Useful resource	Python's nltk library (http://www.nltk.org/)

Refer to the enclosed text file of 2,641 posts collected between Dec 1, 2015 and Jan 16, 2016 from Georgia Tech's campus subreddit (https://www.reddit.com/r/gatech). It has the following fields, all separated by tabs: <post_id TAB UTC_timestamp TAB username TAB post_title TAB post_body>. Each row is a unique post. On this dataset, respond to the following questions in your submission:

Descriptive Statistics (total – 25 points):

1) Present a distribution of the number of posts per day and the number of unique users per day. The distributions will be charts with *x*-axis as the days and *y*-axis as the number of posts/unique users. Qualitatively discuss the nature of these distributions – are there similar volume of posts across all days; is the user distribution heavy tailed, i.e., a small number of users are overly more active than others? *(5 points)*
2) Report the mean and standard deviation of the length of all post titles and post body. *(2 points)*
3) For posts with body text length greater than the mean, get the top (most frequent) 25 uni-, bi-, and tri-grams (25 each; total 75). Similarly for those with body text length less than the mean, get their top 25 uni-, 25 bi-, and 25 tri-grams. Report the *n*-grams and their associated raw and normalized frequencies for each set in separate tables – this is six tables in all, with three columns each, one for the n-gram, one for its raw frequency and the third for its normalized frequency. Normalized frequency of a uni-, bi-, or tri-gram is its raw frequency in all posts (with length longer or shorter than the mean), divided by the max frequency of any uni-, bi-, or tri-gram in the same set of posts. *(10 points)*
4) Discuss the qualitative differences between the uni-, bi-, and tri-grams of the above two sets. What topics typically fare in longer posts versus shorter posts? Specifically, assess whether longer posts may indicate more candid and honest discourse. *(8 points)*

Campus Affect (total – 20 points):

Use the enclosed LIWC resource files to understand collective affect and tone of the overall campus. Specifically, obtain the levels of mean, median and standard deviation of the following LIWC categories in posts over all of the days of the week (Monday to Sunday) and all of the hours of the day (0 hours to 23 hours): *positive affect, negative affect, anger, anxiety, sadness*, and *swear*[1]. LIWC value of a post corresponding to a category will be calculated as the number of LIWC words or stems that are present in the post body, divided by the total number of whitespaced tokens in the post body. Note that since the timestamps of the posts are in UTC, they will need to be converted to Eastern timezone timestamps for this calculation.

1) Include two charts per LIWC category above, one for day of week and another for hour of day trends, where *x*-axis are the days of week or hours of day, and *y*-axis corresponds to the mean, median and standard deviation values of the category. Thus each chart will contain three separate

[1] Each line in each of the LIWC category files is a word or a word stem.

lines corresponding to the mean, median and standard deviation values of the particular LIWC category. *(12 points)*

2) Discuss days of the week and times of the day when each of the above LIWC categories shows relatively higher or lower values. Is the campus discourse generally positive or generally negative? How do your findings align with your general perceptions of the campus student body? Contrast your findings with that of Dodds et al. 2011[2] that we read in class (specifically, Fig 5 and Fig 10). Are your trends similar to that of Dodds et al.? If so, why so? If not, why not? *(8 points)*

Campus Vibe (total – 15 points):

Use the enclosed Moral Foundations dictionary on the post file and assess the nature of campus vibe in terms of the five dimensions given in the moral foundations theory (http://moralfoundations.org/). The theory seeks to understand why morality varies so much across cultures yet still shows so many similarities and recurrent themes. In brief, the theory proposes that several innate and universally available psychological systems are the foundations of "intuitive ethics." The five dimensions of the moral foundations include:

Care/harm: This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance.
Fairness/cheating: This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy.
Loyalty/betrayal: This foundation is related to our long history as tribal creatures able to form shifting coalitions. It underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's "one for all, and all for one."
Authority/subversion: This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
Sanctity/degradation: This foundation was shaped by the psychology of disgust and contamination. It underlies religious notions of striving to live in an elevated, less carnal, more noble way.

1) For all of the post in the post file, get the mean, median, and standard deviation of the levels of the following dimensions given in the dictionary: HarmVirtue, HarmVice, FairnessVirtue, FairnessVice, IngroupVirtue, IngroupVice, AuthorityVirtue, AuthorityVice, PurityVirtue, PurityVice, MoralityGeneral. For instance, the level of HarmVirtue in a post is given as the number of word or word stems matching HarmVirtue (i.e., dimension id 01 in the dictionary), divided by the total number of white-spaced word or word stems in the post. Report the mean, median and standard deviation for each of the dimensions in a table. *(7 points)*
2) Based on the above calculations, discuss qualitatively the morality of the campus (dimension id 11), e.g., is the tone of the campus generally caring and kind (dimension id 01-02)? Do people tend to fair and impartial (dimension id 03-04)? Is the tone collective or authoritative (dimension id 05-08)? Does the tone show degrading tendency or disgust (dimension id 09-10)? *(5 points)*
3) Pick a sample of five posts with very high morality dimension values (dimension id 11) and qualitatively discuss what they talk about. Similarly, present a discussion of a sample of five posts with very low morality values. Examine on how your qualitative examination of this sample aligns with/deviates from what the dimensional value actually indicates; thereby reflecting on the utility and limitations of dictionary based approaches of detecting a psychological attribute in a community/population. *(3 points)*

[2] http://www.uvm.edu/~pdodds/research/papers/files/2011/dodds2011f.pdf