**Natural Language Processing for the Social Sciences (GR5067)**

**HW#2**

1. A positive and negative word list (Hu and Bing 2004) exists on CourseWorks (Files 🡪 hw 🡪 hw2 🡪 negative-words.txt and positive-words.txt). Please create a function called *gen\_senti* that Tokenizes arbitrary text and compares each token with the positive and negative lexicons of each dictionary and outputs the sentiment score, S. Positive and negative words, pw and nw, count as a score of 1 and -1 respectively for each word matched. The total count for pw and nw are pc and nc, respectively. Each message sentiment, S, is normalized between -1 and 1. Any text that does not any positive AND negative words would have to be ignored, and not scored. (**60 points**)

where

For example: Let us say the following sentence was an input into the function “The darkest hour is among us in this time of gloom, however, we will prevail!”. Let’s *pretend* the negative words were *darkest* and *gloom* and positive words were *prevail*

S = (-1 + -1 + 1) / 3 = -1/3 = -0.3333

1. Using the dataframe from lecture, *the\_data*, column *body*, apply this function to each corpus and add a column called “simple\_senti” (**15 points**)
2. Using vaderSentiment, apply the “compound” value of sentiment for each corpus in column *body* on a new column of *the\_data* called “vader” (**15 points**)
3. Compute the mean, median and standard\_deviations of both sentiment measures, “simple\_senti” and “vader” (**10 points**)

References:

Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.

<https://pypi.org/project/vaderSentiment/>