A detailed outline of what I worked on this semester

English tweets

1. Read all the csv files and get all the English tweets
2. Loop through every word in every tweet (excluding retweets), and do the following (two approaches):
    a. If a word has a similarity >= 0.7 to one of the key words ["mosquito", "protect", "repel", "spray"], include that tweet to the filtered tweets, then break out of the loop. The problem with this approach is that there are many words that have nothing to do with the key words were included, such as pray and project.
    b. Hence, the better approach is to check:
        i. If a word is a misspelling of mosquito (by matching it with the regex 'm[a-z]+sq[a-z]+to'), include the tweets
        ii. Or if any of the keywords (repel, protect, spray) appear in that word. This will catch words like repelling, protection, spraying, etc.
3. Remove redundant tweets by applying the following algorithm:
    a. Go over all the filtered tweets, check the previous 10 tweets to that certain tweets. If a tweet found with similarity >= 0.7, consider it a redundant tweet. Hence, this algorithm will remove redundant tweets to some extent. A better approach is to create two nested loops, but the running time for that is n^2, which will take forever.
4. Now we have filtered tweets that we can work with. We need to find the personal tweets among these tweets, hence:
    a. Look for all the tweets that have any of these pronouns ['i', 'me', 'my', 'we', 'us', 'our']
    b. Use nltk python package to use part of speech taggers to find the tweets that begin with verbs (some of those tweets can be personal tweets)
5. At the end, I created a file for each tag. Ashlynn annotated random samples of those tweets, so Ahmed can use them to build the classifier.

Spanish and Portuguese tweets

1. We tried to tackle the Spanish and Portuguese by applying to approaches:
    a. Translating the tweets to English, the follow the exact same approach for English tweets mentioned above. However, this was not possible because all the tools found online have a limit to the number of words/sentences translated.
    b. The second approach is to deal with the original tweets without translating them, the approach taken is similar to the one mentioned above, but with some differences:
        i. Translated the keywords and the pronouns to Spanish and Portuguese
        ii. Followed the same approach except instead of only using nltk, used stanfordnlp package to look for tweets that begin with verbs.

iii. Applied both tools to see which one is better.

Statistics

1. Grouped all the tweets by location and created a new dataframe that has the following attributes:
    a. City
    b. State
    c. Country
    d. Total number of tweets
    e. Number of filtered tweets
    f. Percentage (filtered/total)