CS 4372
ASSIGNMENT 5

Names of students in your group:
James Hooper
Hritik Panchasara

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

For API use and pre-processing

https://www.nltk.org/
https://pythonprogramming.net/named-entity-recognition-stanford-ner-tagger/
www.nltk.org/howto/twitter.html
https://www.tweepy.org

REPORT

- The log file that was created by the program was way too long to include in this pdf file, so it is in a standalone text file. The query that was used for the Twitter API call was (“NASA”, ”SpaceX”, ”Crew1”, ”Crew-1”) and was called November 15, 2020 around the exact time of the launch 18:31:20 CST. There were 5 intervals of collection that each lasted 12 seconds. The data was gained progressively so interval 1 was for 12 seconds, interval 2 was for 24 seconds, interval 3 was for 36 seconds, interval 4 was for 48 seconds, and interval 5 was the grand total run time of 60 seconds. As the program gains the data each interval has an increased number of entities that are being tracked. Through pre-processing the tweets after they have been collected, we can get rid of unnecessary text. The program, using the NLTK packages, will printout the count of all entities and the specific count of all entity tagged words (along with the tag) for each interval.
- The results for this query show case a ton of data. A key thing to mention is that if we were doing sentiment analysis this data could be very useful given the amount of emotions tracked in the regular count of words. The entity tagged words showcase about what was expected. Of course, the queried words, specifically “NASA” and “SpaceX”, showed up a lot. As expected, the current President Donald ‘Trump’ and former President Barrack ‘Obama’ were mentioned, along with the owner of SpaceX ‘Elon’ ‘Musk’. From this data one could gather enough info for the date and location of the event given the tags for ‘America’, ‘Florida’, ‘Kennedy’ ‘Center’, ‘Space’ ‘Station’, ‘International’, ‘2020’, and ‘Sunday’. Lastly, we can also see some of the technology used such as the ‘Dragon’ spacecraft utilized within the launch. The tags themselves seem fairly accurate but one of the big issues that was noticed is the inability to differentiate between PERSON and ORGANIZATION with some words. This could be a context thing or a confusion with brands. One example of this is when ‘Elon’ and ‘Musk’ are tagged as ORGANIZATION’s rather than what they truly represent which is a PERSON.