## Twitter Scraper

In this assignment, you will develop a python program to obtain data from Twitter and provide basic statistics about named attributes (called *tags*) in the resulting tweets. Your program should take as input one or more hashtags that refer to the exact same topic (and hence can be thought of as synonyms) and return tweets that contain any of the hashtags.[1] You should find and submit hashtags that (a): indicate a reasonably *concrete topic*, (b): include the *preferred* or commonly-used hashtag for the topic of interest (among all other possible hashtags for the topic) and (c): result in a *good number of tweets* on the platform, e.g., more than 500 tweets. To find hashtags, go to `https://twitter.com/search`, search for a few topics (based on your interest) and pay close attention to pertinent hashtags. Alternatively, you can find good hashtags through Twitter's trending topics at `https://twitter.com/explore/tabs/trending`. Here are examples of good hashtags that satisfy the above requirements:

- {*#CambridgeMASNOW, #CambMASNOW* }: tweets about snow in Cambridge, MA
- {*#NLP, #NLProc*}: tweets about natural language processing (NLP)
- {*#EmergingTechnology, #EmergingTech*}: tweets about emerging technology

Please avoid general hashtags that return too many tweets. For example *#apple*, *#fashion* or *#goodmorning* are not good hashtags for this assignment as they are too general and cover a wide range of topics.

After finding appropriate hashtags, give them to your twitter scraper program and store the returned tweets in JSON format (do not collect more than 5000 tweets). Process the tags of the returned tweets and report Twitter's organic performance metrics such as `retweet_count`, `favorite_count`, `followers_count`, `friends_count`, and, if available, `lang` and `geo` tags. *Optionally*, you may extract statistics on the most frequent words or phrases from the `text` tags. You may use the following commands to obtain a contiguous sequence of $n$ words from the input text (they are called n-grams[2] in NLP; this link[3] provides ideas and tools for exploratory text analysis):

```
from nltk.util import ngrams
text = "this is a tweet"
list(ngrams(text.split(" "), 2)) # lists n-grams for n=2
>> [('this', 'is'), ('is', 'a'), ('a', 'tweet')]
list(ngrams(text.split(" "), 3)) # lists n-grams for n=3
>> [('this', 'is', 'a'), ('is', 'a', 'tweet')]
```

In addition, process all the returned tweets to (a): remove all the re-tweets and (b): remove tweets that contain less than five (5) words[4] (excluding any hashtags). Submit a sample of exactly 300 tweets from the resulting tweets in JSON format. You may randomly sample 300 tweets or sample based on Twitter's organic performance metrics.

[1] Hashtags are words or phrases prefixed with the # symbol. They are user generated and can be used to label tweets based on their topics or content.

[2] `https://en.wikipedia.org/wiki/N-gram`

[3] `https://neptune.ai/blog/exploratory-data-analysis-natural-language-processing-tools`

[4] A good tokenizer for tweets: `https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/ark-tweet-nlp/ark-tweet-nlp-0.3.2.tgz`

## Important Instructions

You must submit a single zip file named `[STUDENTID].zip` that contains the following files in its root directory:

1. `scraper.py`: a script to run your python code.
2. `hashtags.txt`: a text file that contains the list of hashtags used to collect the data. Hashtags should be separated by new lines and the list should start with the preferred hashtag, which is the commonly-used form of these hashtags.
3. `data_full.json`: a `JSON` file containing all tweets (a maximum of 5000 tweets) returned from twitter.
4. `data_sample.json`: a `JSON` file containing 300 sampled tweets.
5. `report.pdf`: a `PDF` file reporting tweet sampling strategy and data statistics.
6. `readme.txt`: a text file that briefly describes steps to run your `scraper.py` program.

Your `data_sample.json` file should contain exactly 300 tweets in the original format returned by the Twitter API. Do not add/omit any tag and do not change tag values. Your Zip file must be submitted to the link available on Blackboard, otherwise it will be ignored.

Good luck with the assignment!