**SI 618 Fall 2017 Lab 5 – Using Facebook API to collect news discussion data**

In this lab, we will collect data via the Facebook API. We will later utilize this dataset in HW5 and get human judges to inspect it through Amazon Mechanical Turk (Mturk). Lab5 and HW5 will hopefully provide a simple example of (almost) end-to-end data collection and analysis you might face for a data science problem. Our collective goal for Lab5 and HW5 is to understand the level of civility in social media news discussions. To study this problem, we first need to collect data from social media (lab) and then get human judges to read this content and judge them (hw). If you need to, please refresh your memory on how to use APIs from SI506 or SI507!

**What data to collect:** We will randomly assign half of the groups to New York Times Facebook page ([www.facebook.com/nytimes](http://www.facebook.com/nytimes)) and the other half to Wall Street Journal Facebook page ([www.facebook.com/wsj](http://www.facebook.com/wsj)). The group assignments will be posted on canvas. By having different groups working on different news outlet Facebook pages, we hope that the meta analysis we will perform when all homeworks are turned in will provide some interesting insights (e.g. Are there more or less personal attacks on NYTimes or WSJ?)

Step 1

Each team should crawl all comments (comments & nested comments) that have been written under any post from 09/20/2017 to 09/26/2017 on the relevant page (either NYTimes or WSJ). Also, sanitize comments; you need to get rid of any emoji. Create a txt file, si618\_f17\_lab5\_total\_number\_of\_comments\_uniquename1\_uniquename2.txt, and simply write down the total number of comments your team gets.

Step 2

Then, randomly sample 100 comments from the whole comments your team crawled.

Step 3

Save the random sample of 100 comments into a csv file containing four columns. The column names should be pagename, post\_id, comment\_id, comment. pagename is going to be either "wsj" or "nytimes" whichever your team is assigned to. Save it as si618\_f17\_lab5\_random\_sample\_100\_comments\_uniquename1\_uniquename2.csv

All steps can be done with separated codes (parts) but it should be written in one python module and should be able to complete all the steps at a single execution. Save the python module as si618\_f17\_lab5\_uniquename1\_uniquename2.py

**What to submit:**

Submit a zip file named si618\_f17\_lab5\_youruniqname.zip containing your python file as well as the .txt and .csv files.