Emmanuel Valdez

Stat 426 Project Proposal

I am interested in finding the main topics that are being discussed in a variety of twitter queries/hashtags.

Machine learning method: unsupervised learning and/or topic modeling

Data:

I have a list of about 30 queries that I will pull tweets from. I am also working on getting a list of hashtags to pull tweets from. I’ve looked into the tweepy package for twitter scraping and have pulled some data already using that. I will need to do some more research into the package to better understand how to use it though. I’m not sure what/if there are limitations to the amount of data I can pull from twitter. I think it would be beneficial to have access to a month’s worth or more of tweets surrounding a given query/hashtag.

Having a corpus of tweets from each query/hashtag will allow me to perform some unsupervised learning methods on the data in an effort to find main topics of discussion based on the text in the tweets.