Bryan L. Mack

MSDS 600

05-17-2015

In Week 2 of this course, our assignment was to use the given Python scripts to extract data from Twitter. This was my first experience of using Python, and my first experience in creating my own data extract via publically available information.

Setting up Python & the environment in which to run Python (DOS) was relatively simple using Dr. Hart’s provided environment script. Additionally, setting up the Twitter developer account was quite simple and intuitive. I downloaded the twitterstream.py file and modified it using the unique keys provided to me by Twitter. From here, it took a simple execution of my Python script and I was creating a data dump in JSON format from Twitter which we can use to analyze.

The amount of data created in this extraction is massive. I browsed through my “output.txt” file in Sublime editor as well as in Excel (by renaming it to “output.csv”). There are a plethora of fields available to analyze, and I am excited to find useful and ethical ways to use a dataset such as this.

This Twitter extraction ability is very exciting to me; it scratches the surface of what we can do with large datasets and opens the door to many further data analysis techniques. I plan on spending some time this week determining how to extract tweets containing certain words, hashtag references, or those within a specific location. It would be incredibly interesting to extract location-based data and synthesize it with a dataset from Foursquare; analyses could be done to see how busy locations are as well as using text mining of the tweets to determine if the busy locations are receiving positive or negative feedback. That is just a single exploration opportunity I foresee; the possibilities seem endless.