Large-Scale Social Media Data Retrieval
Christopher Cahill (EAS 2018)
Advisor: Lynn Wu

My research included helping Wharton Operations, Information, and Decisions Department Professor Lynn Wu and PhD Fujie Jin with their project regarding analyzing start-up success with respect to presence on various social media. The overarching task was to collect data from both consumer-focused social media and investor-focused social media. From there, the relationship between effective start-up techniques and their social media presence could be examined in further detail. The consumer-focused social media namely consisted of Facebook and Twitter, for which posts/tweets were collected, as well as the time, date, and identifiers of all posts/tweets. For Facebook, the number of likes for the company page was collected, as well as the number of likes, shares, and comments for each post. For Twitter, the followers and following accounts were collected. On the business side, LinkedIn, AngelList, and CrunchBase data was collected. AngelList and CrunchBase specifically cater towards start-ups, meaning they provide a wealth of data regarding the company, investors, and the network of people involved. LinkedIn provides a: 1) potential network of current people, 2) information about the people involved and their network with respect to education and past start-ups, and 3) information regarding the skill-set of people involved. All of the data, when taken from the web or an API, came in a raw form, like HTML or XML. In other words, it needed to be refined before use. As a result, I cleaned all of the files so that the end data was easy to manipulate and readable. I primarily used Python when working with Twitter and Facebook (FQL), as well as when I was refining the data. For LinkedIn and scraping, I used Java with Selenium. There was a total of slightly less than 40,000 companies involved, generating dozens of gigabytes of data in total.