

---

ISyE 2028 – Basic Statistical Methods - Fall 2015  
Bonus Project: “Big” Data Analytics  
Proposal (or Final Report)

---

James Moriarty

For my bonus project, I would like to utilize the statistical tools we are discussing in class to examine data I am working with for the GT Big Data Club.

I first began attending meetings of the BDC at the beginning of this semester. The club is currently focused on a reactive news application; Retina News. Retina News is an aggregator that combs major news sites like Reuters, BBC, etc. for articles and organizes them to present interesting and up-to-date visualizations of the news. Recently I began looking for ways to identify highlights, or trending news, and find out what keywords and phrases would be most representative of what is currently trending.

This is the problem that I will address with my project. Every article in our database contains about 2 – 6 words which are “keywords” of the article. 400 – 500 articles are added to the database daily and we have accumulated more than 100,000 keywords. The challenge will be to periodically evaluate what keywords are “trending”, or appearing more frequently than they normally do.

I plan to calculate how often each keyword appears at discrete time intervals, like once a day. From this I will calculate the mean and variance of the occurrences of each keyword each day. For example, I predict the keyword ‘china’ will appear more frequently than the keyword ‘earthquake’, but the keyword ‘earthquake’ will have a higher variance on a daily basis. (‘china’ is actually the most popular keyword in our dataset)

Once I have built functions to calculate the mean and stdev of each keyword automatically, I will build confidence intervals to predict the amount of times a keyword will appear on any given day. Everyday I will also be able to calculate which keywords have appeared an unusually high number of times and even build a ‘trending’ list ranking of which keywords have been mentioned the most unusually high number of times in a day. This would also allow me to identify which articles feature the day’s trendiest keywords.