

---

# ISyE 2028 – Basic Statistical Methods - Fall 2015

## Bonus Project: “Big” Data Analytics

### Final Report – James Moriarty

---

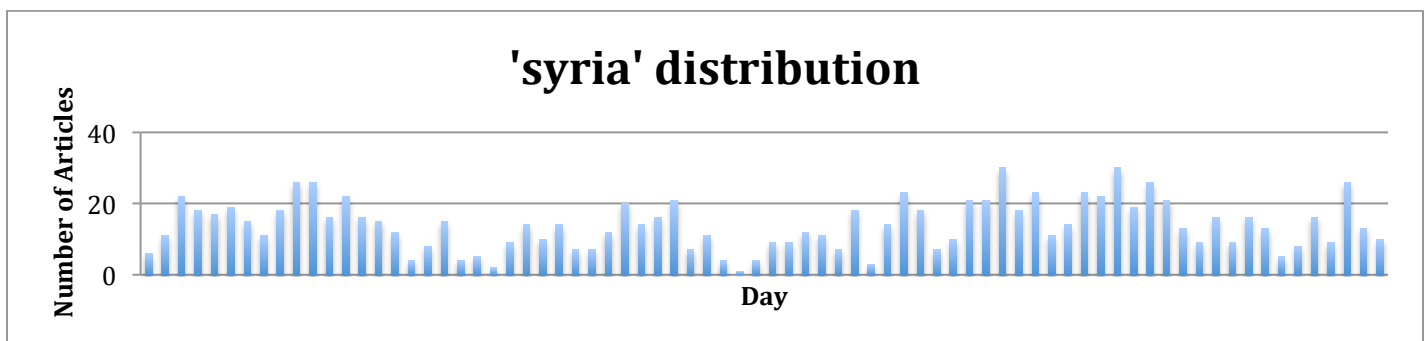
For my bonus project, I utilized the statistical tools we discussed in class to examine data I was working with for the GT Big Data Club.

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.

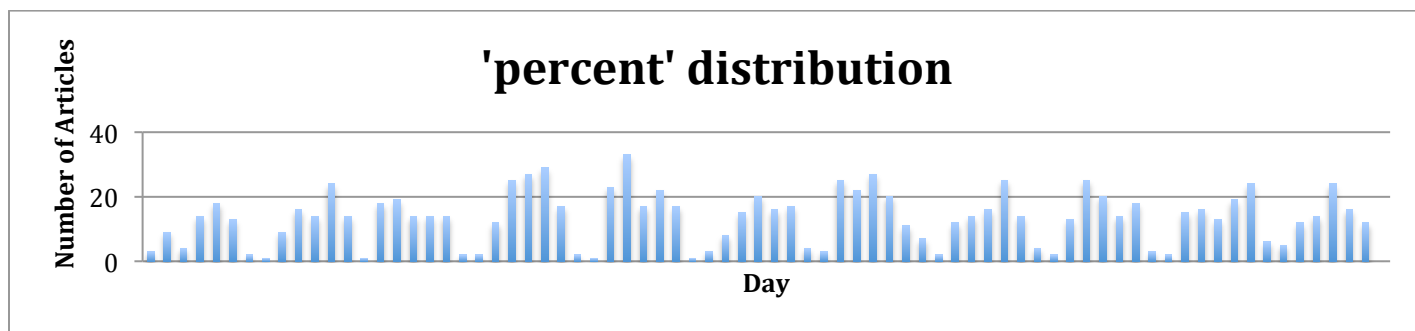
This is the problem that I will address with my project. Every article in our database contains about 2–6 words that are “keywords” of the article. 400–500 articles are added to the database daily and we have accumulated more than 300,000 keywords. In the past 76 days, the most common keywords based on how many articles cited them per day were:

1 syria 13.97368421	6 paris 10.65789474	11 climate 7.986842105
2 percent 13.27631579	7 digital 9.789473684	12 refugees 7.960526316
3 police 12.30263158	8 russia 9.210526316	13 migrants 7.789473684
4 china 11.51315789	9 google 8.342105263	14 eu 7.684210526
5 russian 10.92105263	10 india 8.302631579	15 company 7.605263158

The most popular keywords are a mix of countries, people groups, companies, and other global concepts. All 300,000 keywords have a distribution of how many articles tag them each day. For example, here is the distribution of Syria’s daily mentions:

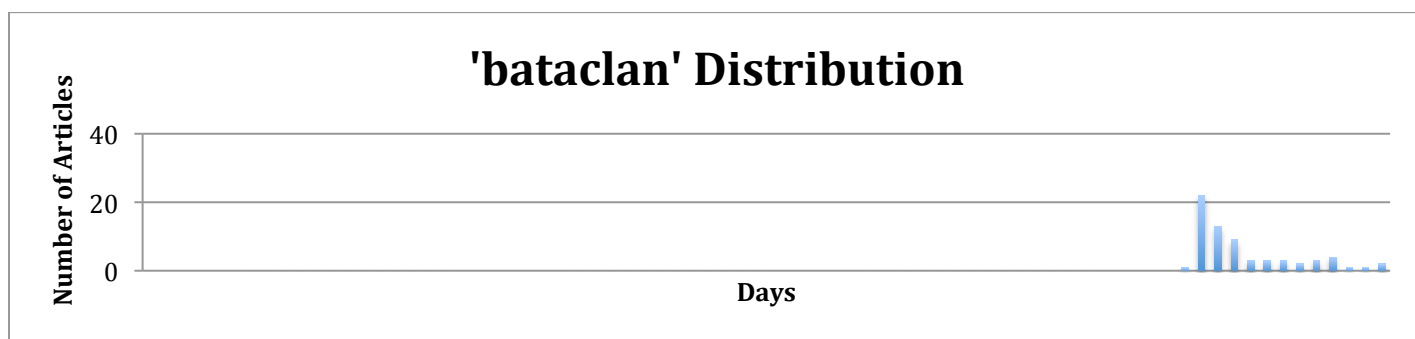


I wanted to use this data to identify the current “trendiest” words, however the keywords that appear the most frequently are not necessarily the trendiest. There were a lot of interesting distributions among the keywords. Here’s the cyclical distribution for ‘percent’:



This keyword sees a substantial decrease in mentions on the weekends, most likely due to the stock market being closed, so there are no articles regarding changes in the companies’ prices.

Here is the distribution of ‘bataclan’, the theatre that was targeted in the Paris attacks:



Trendiness is a vague term, but on the day of the Paris attack, most people would say that ‘bataclan’ was a more relevant term than ‘percent’, even though they were mentioned in similar amounts.

The solution was to quantify trendiness based on p-values. I wrote a function to calculate the standard deviation of each keyword based on its article counts over the past 76 days. The p-value measured the probability a keyword was mentioned x times in the past 24 hours, given its floating average and standard deviation. Because all the keywords have equally sized data sets, this ranking essentially measures how many standard deviations a keyword’s daily count is above its floating average. When I tested this, the top results were very obscure keywords and typos, so I decided to only consider keywords whose average daily mentions exceeded a certain threshold. With some tweaking I decided that a threshold 0.4 mentions a day yielded the best list.

Here were the 12 trendiest words on Sunday November 29<sup>th</sup> over the default 0.4 mention threshold:

1 la 6.419224372	4 black friday 4.784263848	7 abortion 3.893129701	10 body 3.490348607
2 sydney 5.362109137	5 boxing 4.651236675	8 weekend 3.693882334	11 colorado 3.453456649
3 bangui 5.213671857	6 protesters 3.934529879	9 compaore 3.681219671	12 peacekeepers 3.206338544

Here were the 12 trendiest words on Sunday November 29<sup>th</sup> over a threshold of 0.8 mentions per day. These keywords are tagged more often, but are not trending as much.

1 protesters 3.934529879	4 jet 2.481944082	7 suicide 2.395234162	10 rosberg 2.037432244
2 prix 3.076574933	5 murray 2.435865017	8 conflict 2.277493137	11 agreement 1.965264993
3 shopping 2.657122564	6 incident 2.412128349	9 climate 2.259113593	12 mali 1.851961473

There are a couple directions I could work towards to improve the accuracy of the "trending list". One would be to increase the distribution size beyond 76 days. For some reason, I began seeing a sharp decline in performance time when I requested more than 76 days of data. Another improvement would be to experiment with what time size yielded the most accurate results. Maybe comparing the frequency of keywords in 6-hour or 1-week intervals would improve the trend list. The problem with this is that the stream of articles entering the database is too weak to yield precise data at smaller time ranges.

Another issue with the database is that the articles come from a small concentration of media sources. Here are the numbers of articles by source in the last 24 hours:

'france24', 116  
'reuters', 79  
'business\_insider', 43  
'guardian', 37  
'cnn', 27  
'bbc', 16  
'techcrunch', 15  
'venture\_beat', 11  
'aljazeera', 6

Implementing the same methods on tweets rather than articles would substantially improve the quantity and variety of keywords, but would present other challenges.