Detecting Trends in Twitter Health News

Nasib Mansour and Burcum Eken

Abstract

Twitter has become a major source for data in recent years and it provides great opportunities for data scientists and analysts to develop models to try and solve whatever problem they have in mind since everything is being shared. One of the important areas is health and twitter has a lot of information from news agencies. By considering health tweets, we can get a lot of insights on the health problems the world faced during the time if the collected data, and we can know what problems were popular or not. We will try and use a MapReduce approach to build a model that can detect trends in Twitter data and produce graphs that can help visualize the results.

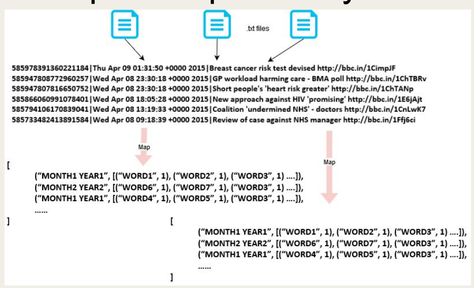
Introduction

For our project, our goal is to find a way to detect trends of health issues. More specifically the questions we are trying to answer are: When does flu become an important issue each year (for the years the data was collected), and can we find another major health issue that has recurring patterns and be able to apply the same model to it? This problem is sure to be done before but not in the same way we are presenting it, we are not sure how it is being done but we can assume that there are some tools that have been developed so that medical professionals always are in the know for trends in health issues.

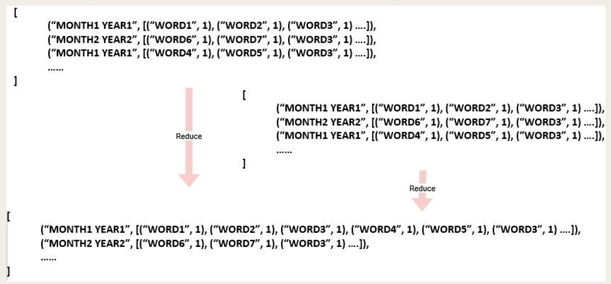
Our contributions include building a model that when given a set of tweets and their corresponding time (month and year preferably) and specify a disease, can be able to produce a monthly trend graph for each year of the data collected. This would make it very convenient for anyone who is curious to get helpful information from only a set of tweets. It is hard to validate the system accuracy since there is nothing to compare it against, but one approach might be using our previous knowledge on when flu for example tends to start and end as a validator. Our goal also is to have a model that is not slow and can finish in a reasonable amount of time.

Methods

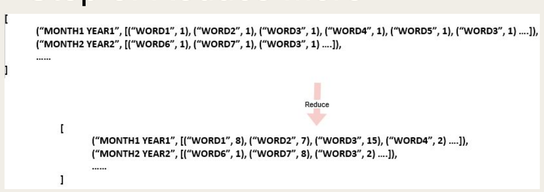
- Step 1: Map to array



- Step 2: Reduce to unique dates



- Step 3: Reduce More



- MapReduce programming model approach
- Pyspark (Spark Python API) for parallel processing
- Run in Jetstream allocation

Results

Fig 1: Trends for **FLU** between 6/2011 and 4/2015

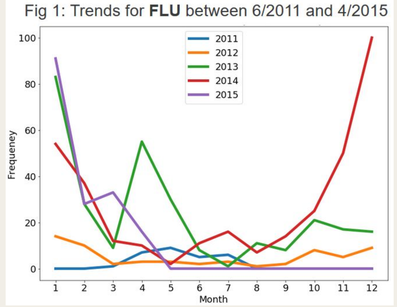


Fig 1 indicates rise of tweets with "flu" at the end of each year and beginning of next year, while flattening in the middle of the year.

Fig 2: Trends for **EBOLA** between 6/2011 and 4/2015

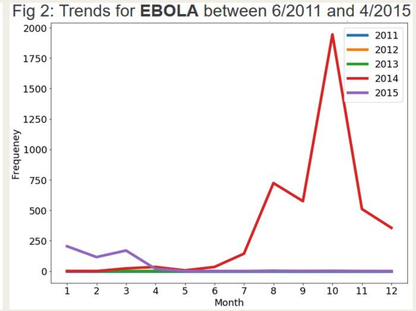


Fig 2 indicates sharp rise in "Ebola" in tweets mid 2014. There was an Ebola epidemic outbreak between 2014-2016 according to CDC[1].

Fig 3: Trends for **HEALTHCARE** between 6/2011 and 4/2015

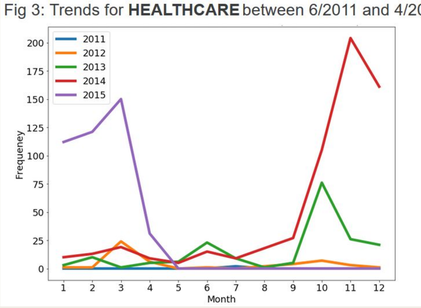
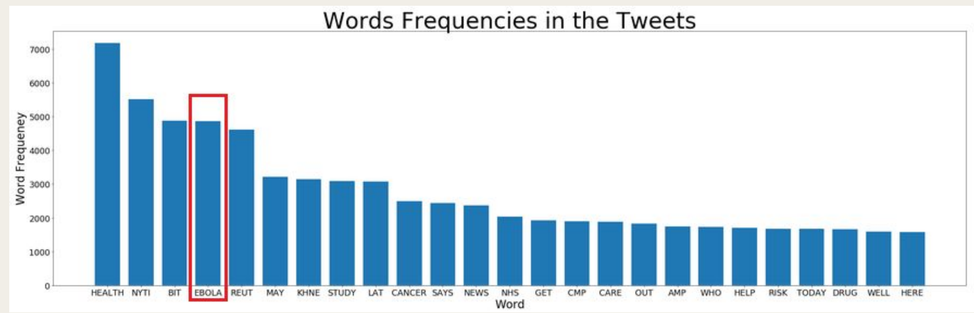


Fig 3 proves that we can use the model for any keyword and produce nice graphs.



- Histogram shows the top 25 word with highest word counts.
- Ebola is the disease with most tweets

Conclusions

- **Advantages/Qualities:**
  - Model Can be applied to any field not just health given relative set of tweets
  - Dynamically detect start and end dates of tweets
  - Produces clean graphs that can be easily understood
- **Lessons Learned:**
  - Importance of understanding the data and cleaning it.
  - Understand how to apply MapReduce Model.
- **Areas of Improvement:**
  - Make use of synonyms and plurals
  - Make more efficient (~ 30 seconds for ~60k tweets)
  - Graphs can be more user friendly configurable
- **Future Work:**
  - Work on suggested improvements.

Acknowledgements

We would like to thank Dr. Michela Taufer and Our GTA Nigel Tan and expert Mike Wyatt for their help and support to get this work done

References

[1] source:https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html

THE UNIVERSITY OF TENNESSEE KNOXVILLE