Tweeflix: Public Health Surveillance by Streaming Real-Time Twitter Data

Kavin Kumar Arumugam; Shivani Sivasankar

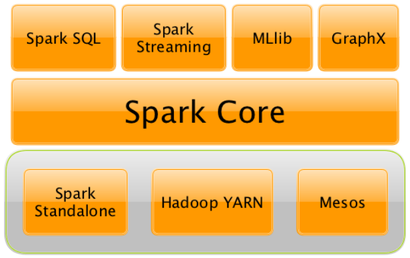
University of Missouri – Kansas City, School of Computing and Engineering

Introduction

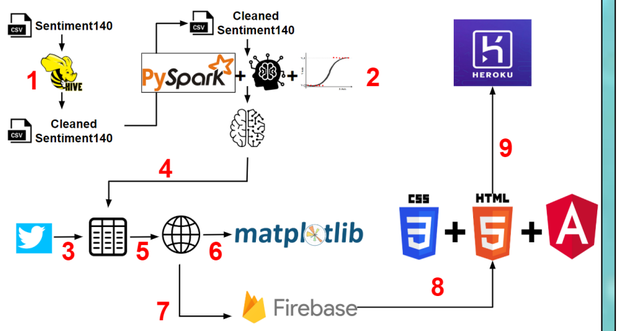
Promoting and protecting the health of communities is the goal of public health. Epidemiology is concerned with the dynamics of health conditions in populations. Rapid response through improved surveillance is important to combat emerging infectious diseases. The goal of this work is to use data-mining techniques such as tweet classification, sentiment analysis and content classification for public health surveillance.

Objective

In this project, we examine the use of information embedded in the Twitter stream to:
1. Track rapidly-evolving public sentiment with respect to commonly reported diseases
2. Track and measure actual disease activity



Methods



Technologies used

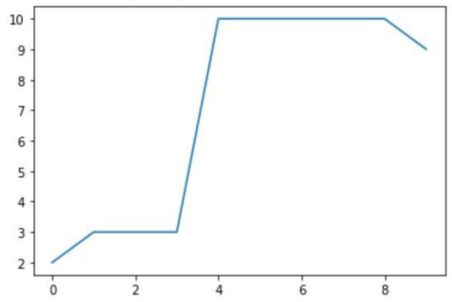
1. **Python 3.7**

2. **pySpark** - SparkContext
3. **pySpark.Streaming** - StreamingContext
4. **pySpark.sql** - SQLContext
5. **pySpark.ml** – Machine Learning
6. **graphframes**
7. **Matplotlib**
8. **Seaborn**
9. **Jupyter-Notebook**

Results

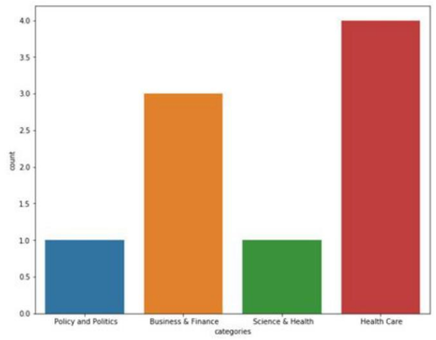
Query 1 - Number of tweets per iteration (every 2 seconds)

The number of tweets streamed every two second (per iteration) were visualized using line graph.



Query 2 - Tweets classified based on the type of content

Streaming tweets classified based on the type of category people are talking about right now

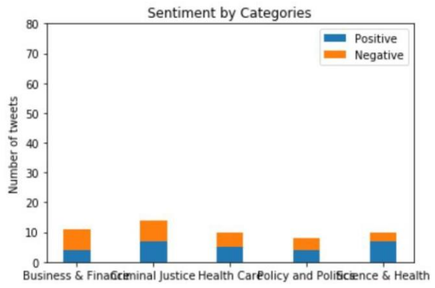


Query 3 -Top Trending Hashtags

Top trending hashtags from the streaming twitter data represented as a word cloud

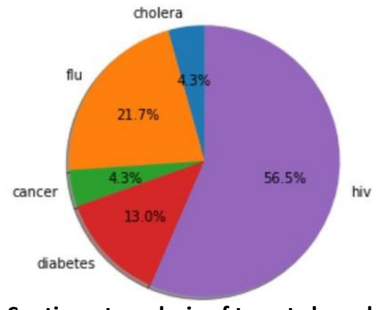
Query 4 - Sentiment analysis of tweets based on content category

The number of tweets which are positive or negative in sentiment are classified based on the content category which is represented by a stacked bar chart.



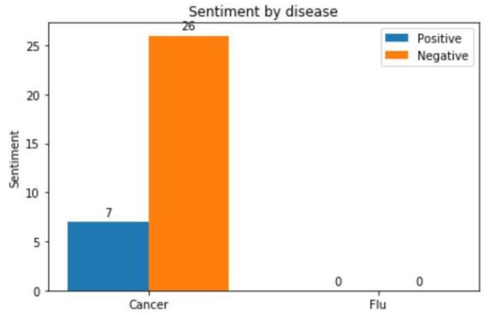
Query 5 -Proportion of streaming tweets classified by popular diseases

To identify the number of people currently talking about the popular diseases (Cholera, influenza, cancer, diabetes, HIV/AIDS) where its proportion is represented in the form of pie chart



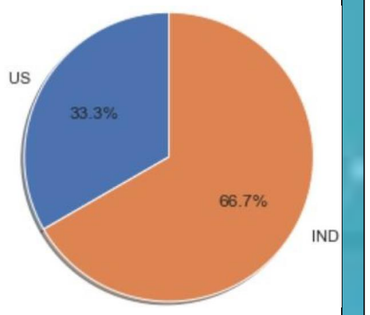
Query 6 - Sentiment analysis of tweets based on disease category

The number of tweets which are positive or negative in sentiment are classified based on the disease category which is represented by a grouped bar chart.

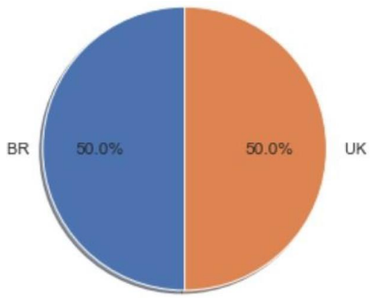


Query 7 -Positive sentiment tweets classified by country for HIV disease

Positive sentiments about the tweets which mentions HIV was counted and its proportion was grouped by country and represented as a pie chart.



Query 8 – Negative sentiment tweets classified by country for HIV disease



Negative sentiments about the tweets which mentions HIV was counted and its proportion was grouped by country and represented as a pie chart.

Query 9 - Top trending Hashtags specific to health care

Analysis of the most frequently mentioned Hashtags in the health care content category is represented by a word cloud

Query 10 - Prevalence of disease classified by country

The number of streaming tweets which mention the five most common diseases (Cholera, influenza, cancer, diabetes, HIV/AIDS) were classified based on the location of the tweets.

	cnt	_4	_9
0	1	AUS	cholera
1	1	BR	flu
2	3	AUS	diabetes
3	2	US	flu
4	1	BR	cancer
5	1	IND	flu
6	1	IND	cholera

Performance

An average of 10 tweets are collected for every iteration (2 seconds). Spark query engine takes about ~4 seconds to execute a query for every iteration of streaming data in a single node cluster

Conclusion

- This real-time public health surveillance system enables online processing capabilities of the raw tweet stream, which is a departure from methods that mostly use curated collections of tweets for their analysis.
- This system can also effectively track daily disease activities and the volume changes of tweets mentioning disease related terms over time.
- All of the output data is visualized as interactive maps, pie charts, and bar graphs.
- Our system is highly scalable and can be easily extended to track other diseases.
- Because the system is completely automated and the output of analysis is updated near real time, it can detect disease outbreaks significantly faster than the traditional disease surveillance system that collects public health data from medical practices.

Reference

https://www.kaggle.com/kazanova/sentiment140
https://data.world/elenadata/vox-articles/workspace/file?filename=dsjVoxArticles.tsv
https://www.toptal.com/apache/apache-spark-streaming-twitter
http://users.eecs.northwestern.edu/~kml649/publication/kdd2013.pdf

Guidance

Dr.Praveen Rao Ph.D.
University of Missouri – Kansas City
School of Computing and Engineering