graphy

“Trend”-y Tweet Classification!

😋

Motivation

We were wondering, “What in the world do people tweet about?!?” 🤔

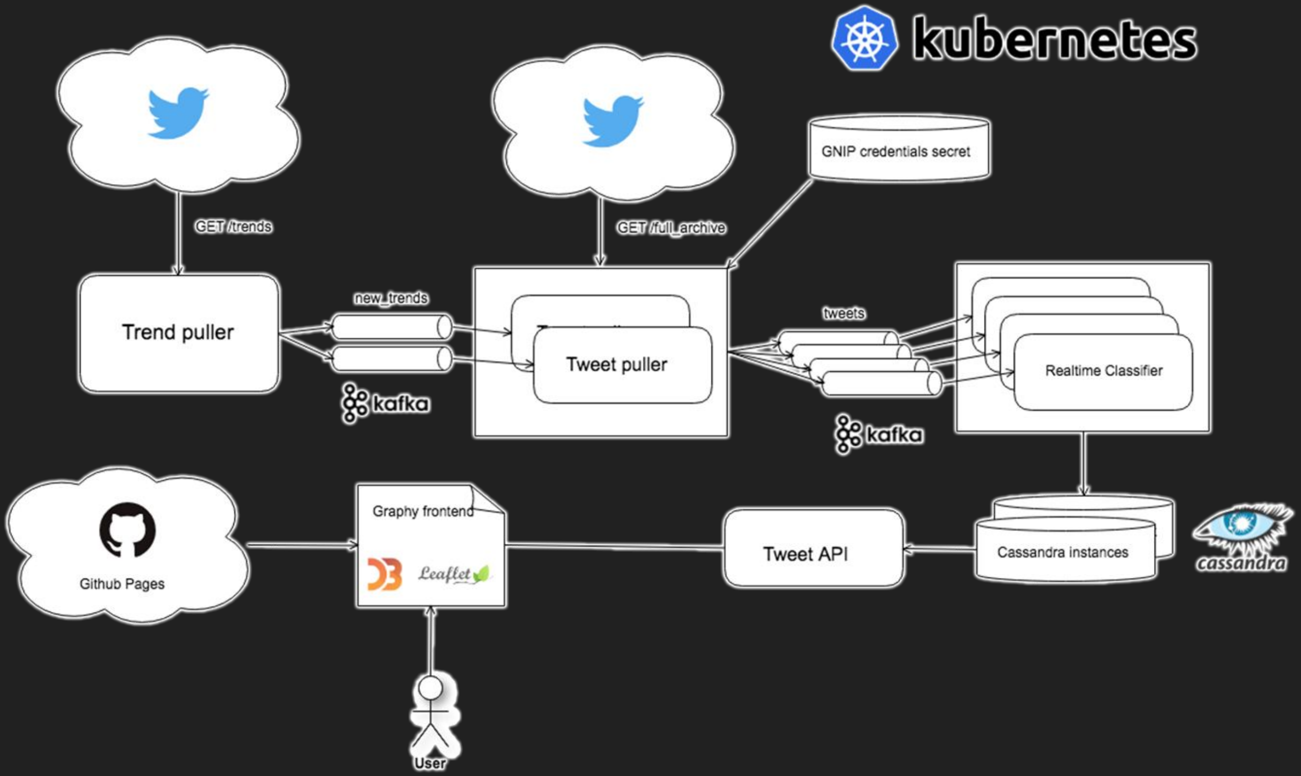
Spoiler: Mostly Entertainment and Sports… 😛

Understanding what’s happening

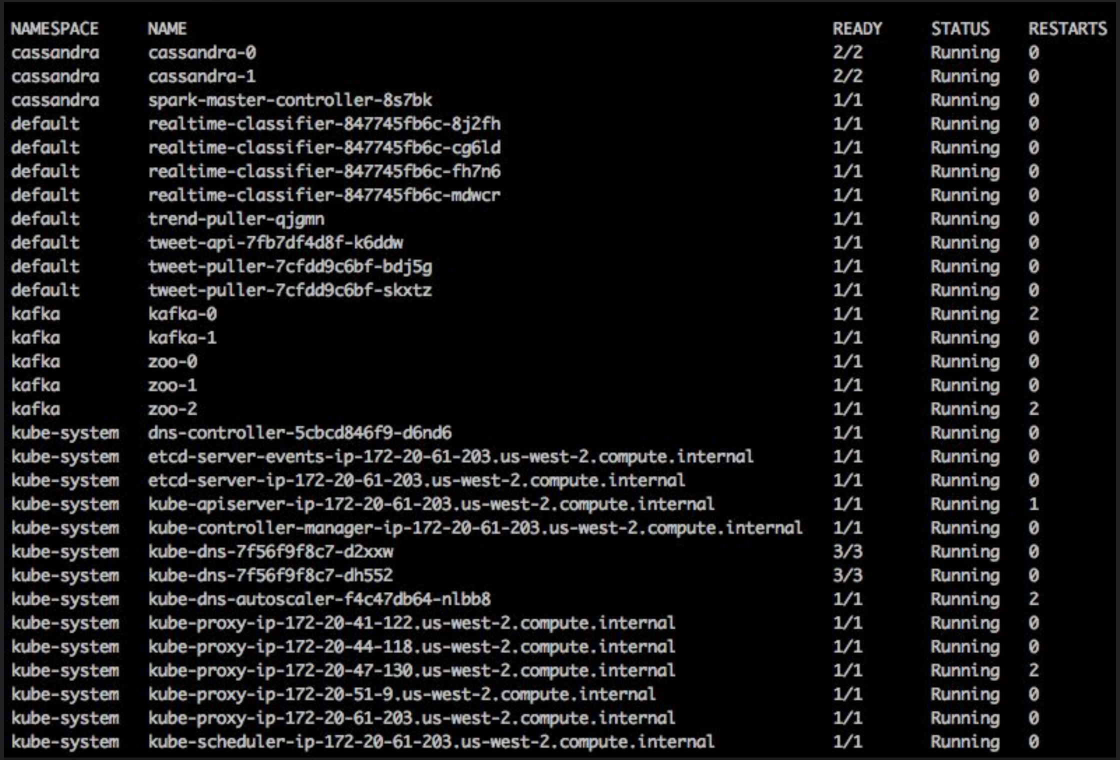
We wanted to analyze this further and find the correlation among the popular content.

Backend + DevOps

System Architecture



System Architecture



```
NAMESPACE     NAME                                                                  READY   STATUS    RESTARTS
cassandra     cassandra-0                                                           2/2     Running   0
cassandra     cassandra-1                                                           2/2     Running   0
cassandra     spark-master-controller-8s7bk                                         1/1     Running   0
default       realtime-classifier-847745fb6c-8j2fh                                  1/1     Running   0
default       realtime-classifier-847745fb6c-cg6ld                                  1/1     Running   0
default       realtime-classifier-847745fb6c-fh7n6                                  1/1     Running   0
default       realtime-classifier-847745fb6c-mdwcr                                  1/1     Running   0
default       trend-puller-qjgmn                                                    1/1     Running   0
default       tweet-api-7fb7df4d8f-k6ddw                                            1/1     Running   0
default       tweet-puller-7cfdd9c6bf-bdj5g                                         1/1     Running   0
default       tweet-puller-7cfdd9c6bf-skxtz                                         1/1     Running   0
kafka         kafka-0                                                               1/1     Running   2
kafka         kafka-1                                                               1/1     Running   0
kafka         zoo-0                                                                 1/1     Running   0
kafka         zoo-1                                                                 1/1     Running   0
kafka         zoo-2                                                                 1/1     Running   2
kube-system   dns-controller-5cbcd846f9-d6nd6                                       1/1     Running   0
kube-system   etcd-server-events-ip-172-20-61-203.us-west-2.compute.internal        1/1     Running   0
kube-system   etcd-server-ip-172-20-61-203.us-west-2.compute.internal               1/1     Running   0
kube-system   kube-apiserver-ip-172-20-61-203.us-west-2.compute.internal            1/1     Running   1
kube-system   kube-controller-manager-ip-172-20-61-203.us-west-2.compute.internal   1/1     Running   0
kube-system   kube-dns-7f56f9f8c7-d2xxw                                             3/3     Running   0
kube-system   kube-dns-7f56f9f8c7-dh552                                             3/3     Running   0
kube-system   kube-dns-autoscaler-f4c47db64-nlbb8                                   1/1     Running   2
kube-system   kube-proxy-ip-172-20-41-122.us-west-2.compute.internal                1/1     Running   0
kube-system   kube-proxy-ip-172-20-44-118.us-west-2.compute.internal                1/1     Running   0
kube-system   kube-proxy-ip-172-20-47-130.us-west-2.compute.internal                1/1     Running   2
kube-system   kube-proxy-ip-172-20-51-9.us-west-2.compute.internal                  1/1     Running   0
kube-system   kube-proxy-ip-172-20-61-203.us-west-2.compute.internal                1/1     Running   0
kube-system   kube-scheduler-ip-172-20-61-203.us-west-2.compute.internal            1/1     Running   0
```

Focus on scalability

System Architecture details

- System details:
    - Kubernetes v1.8.7
    - Deployed with Kops for AWS
- Kubernetes Nodes
    - **Master**: 1 EC2 t2.medium instance
    - **Nodes**: 4 EC2 t2.small instances
- AWS:
    - 2x 100GB EBS (for Cassandra Storage)
    - Elastic Load Balancer to balance if needed to add more APIs servers
- Classifies ~**6K tweets/min**
- **GitHub pages CDN auto-deployment**

Machine Learning

Classification

Topic Classification

Classifying tweets into the following categories:

- Sports
- Politics
- Technology
- Mood
- Entertainment

Sentiment Classification - Joy, Sadness, Anger, Neutral.

Workflow

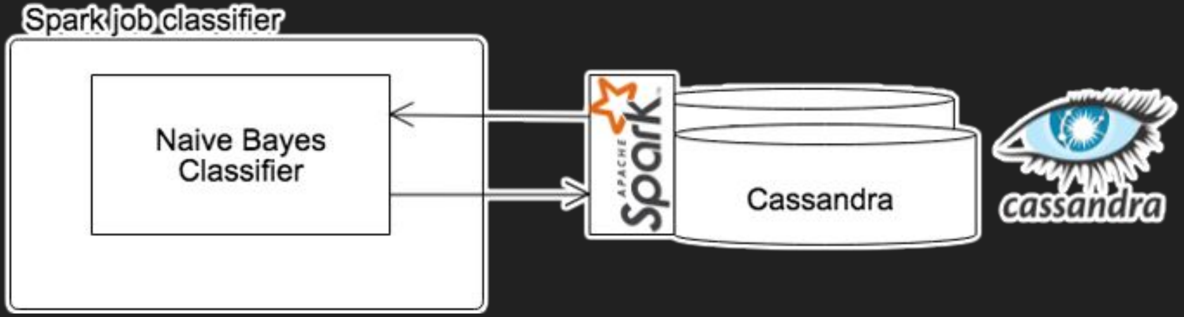
1. Pre-processing
   a. Cleaning Data with regex
   b. Removing twitter specific stop words
   c. Monkey trend labeling
2. Model
   a. Naive Bayes - bag of words model
   b. Suitable for real time pipeline
3. Trend Classification

   - Based on majority classification of it's tweets

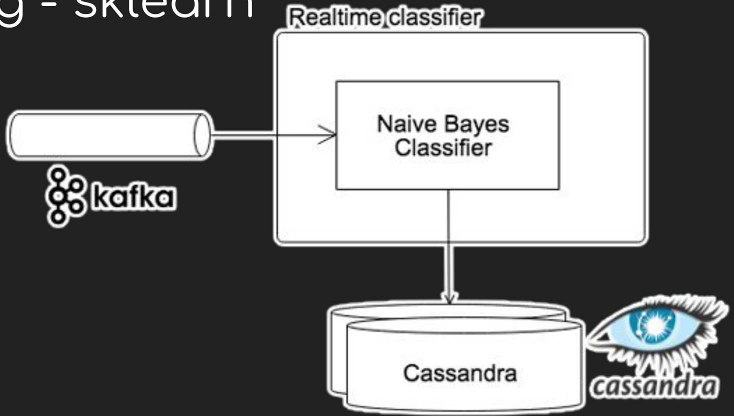
4. Results/Accuracy
   a. Precision and recall ~ 93%
   b. 6k classifications per minute

Infrastructure

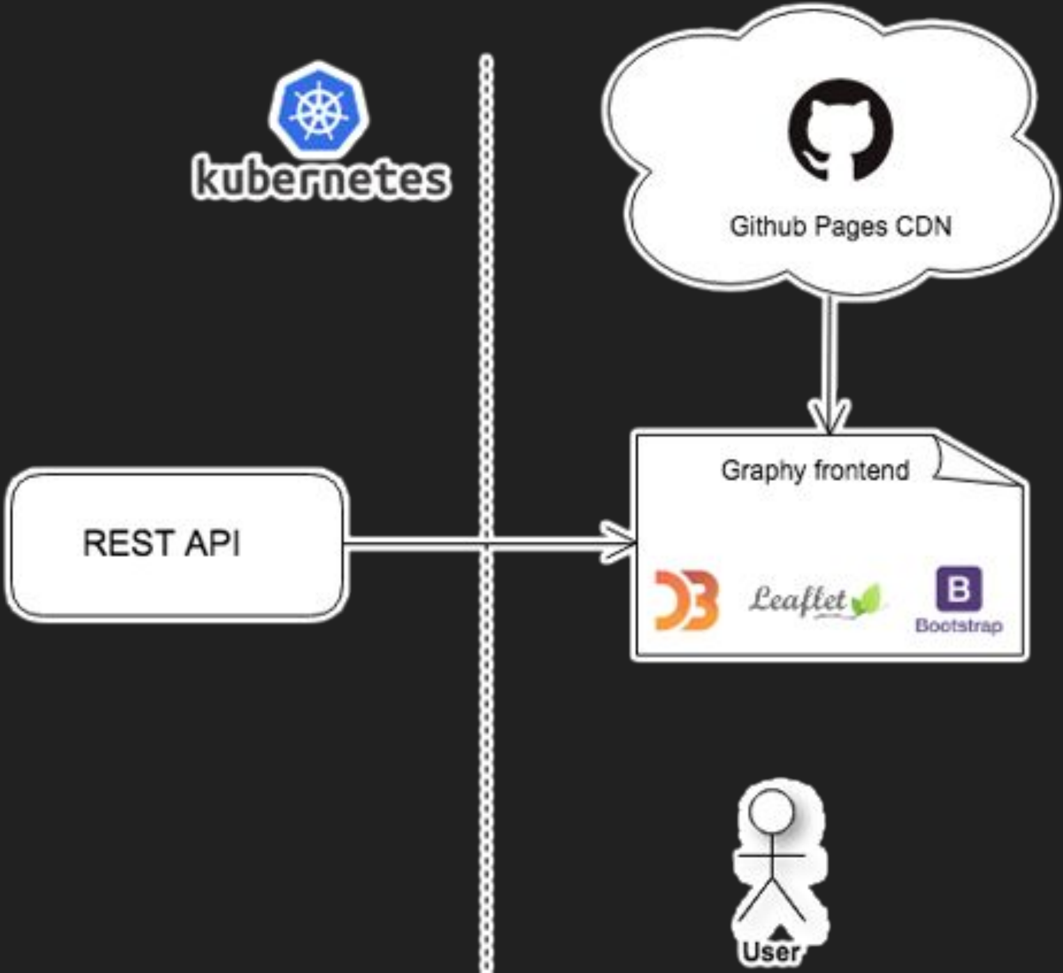
- Batch Processing - Spark-mllib



- Real Time Processing - sklearn

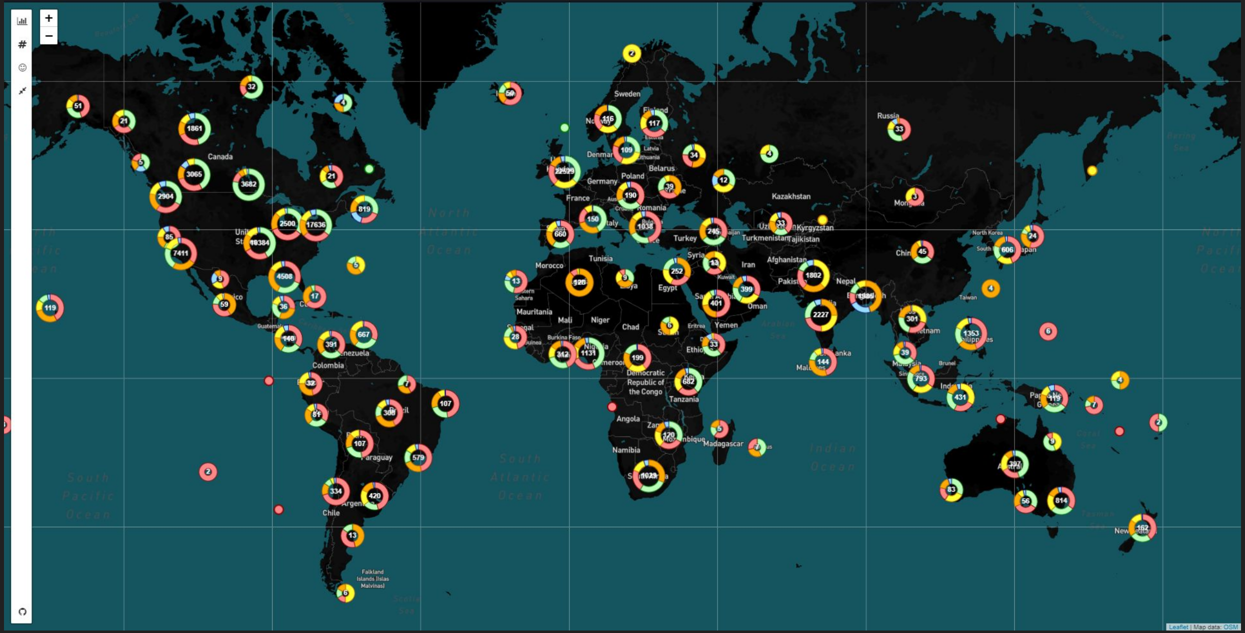


Frontend


kubernetes
Github Pages CDN
REST API
Graphy frontend
D3
Leaflet
B
Bootstrap
User

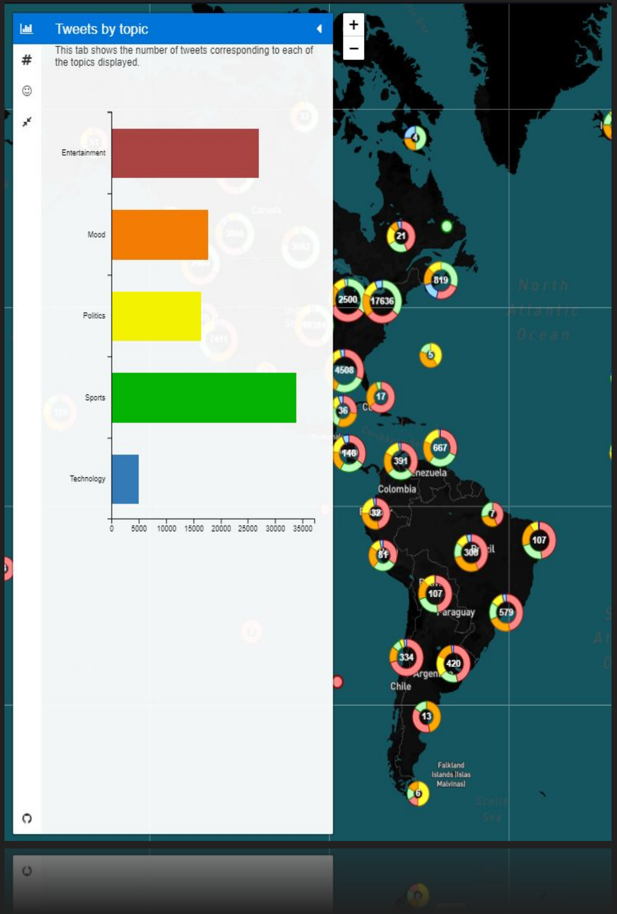
Background Map

Map - created with Mapbox (customized URL) and Leaflet. ClusterPies, Marker Cluster, D3.



Sidebar Features

- **Tweets by Topic: C3 Chart**
- A D3-based reusable chart library
- Present bar chart of most popular tweets

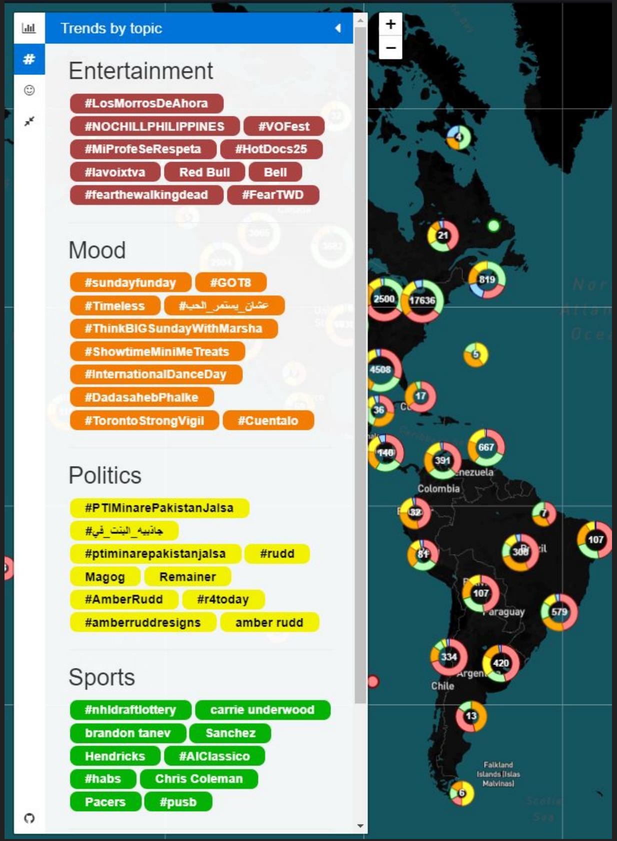


Tweets by topic

This tab shows the number of tweets corresponding to each of the topics displayed.

Sidebar Features

- Trends by Topic: Bootstrap Badges
- Displays hottest tweets by each trend



Trends by topic

Entertainment

#LosMorrosDeAhora #NOCHILLPHILIPPINES #VOFest #MiProfeSeRespeta #HotDocs25 #lavoixtva Red Bull Bell #fearthewalkingdead #FearTWD

Mood

#sundayfunday #GOT8 #Timeless #عشان_يستمر_الحب #ThinkBIGSundayWithMarsha #ShowtimeMiniMeTreats #InternationalDanceDay #DadasahebPhalke #TorontoStrongVigil #Cuentalo

Politics

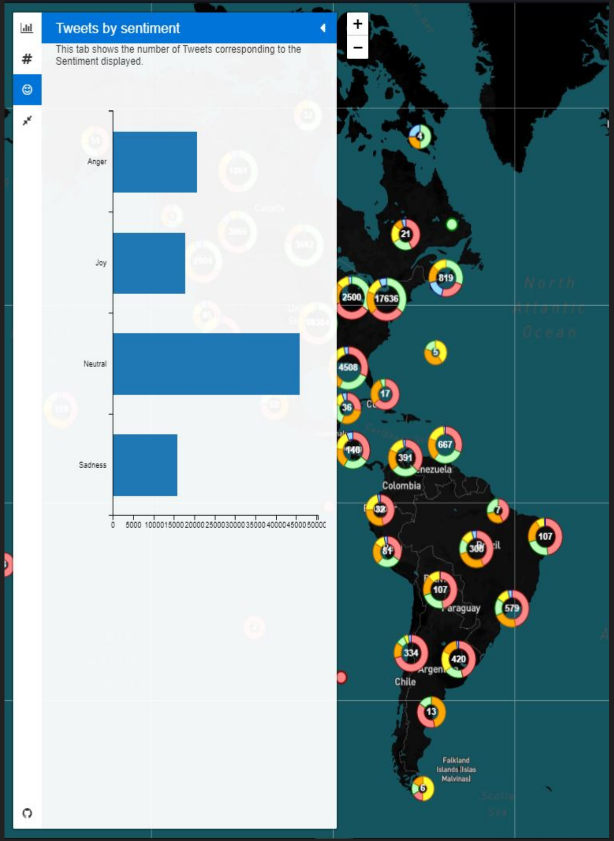
#PTIMinarePakistanJalsa #جاذبيه_البنت_في #ptiminarepakistanjalsa #rudd Magog Remainer #AmberRudd #r4today #amberruddresigns amber rudd

Sports

#nhldraftlottery carrie underwood brandon tanev Sanchez Hendricks #AlClassico #habs Chris Coleman Pacers #pusb

Sidebar Features

- **Sentiment Analysis**
- Emojis for each tweet on map
- Description…if needed…will be added.



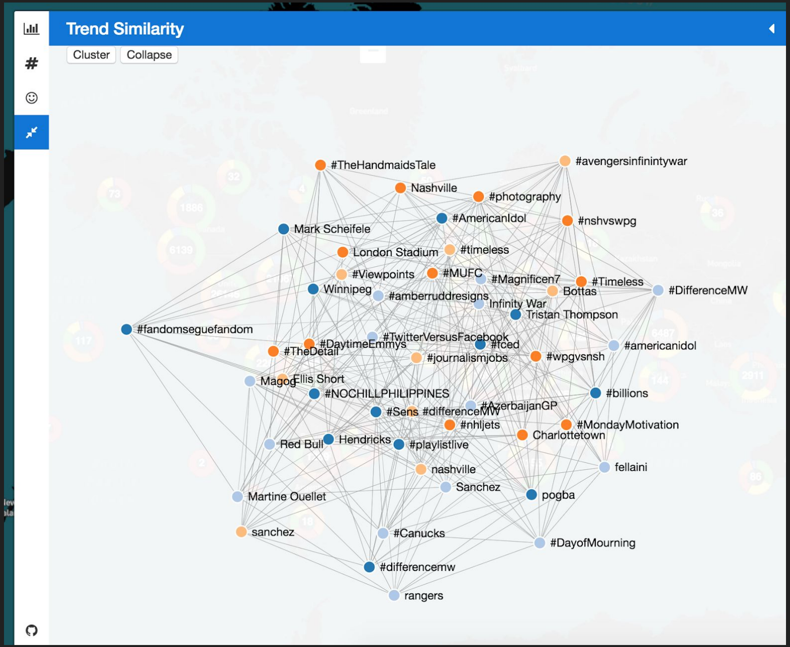
Tweets by sentiment

This tab shows the number of Tweets corresponding to the Sentiment displayed.

Sidebar Features

- Trend Similarity: D3 Network Graph

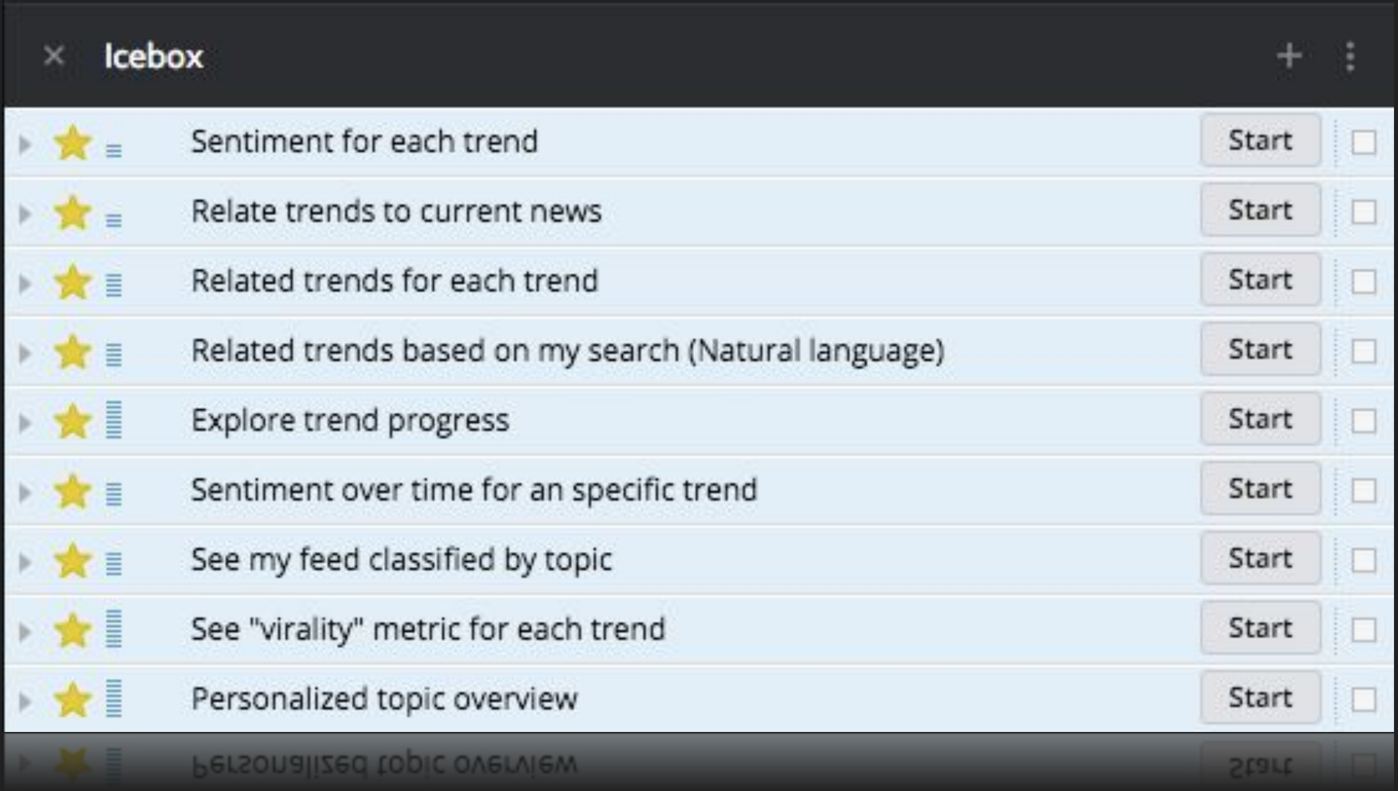
- Clustering Trends with Doc2Vec and **Clauset Algorithm**
  - Professor Aaron Clauset's Algorithm



Issues we struggled upon

- Spark job deployment
- Deploying Kubernetes in AWS is not easy
- D3 learning curve
- Difficulties to label data
    - Non-existent labeled data
    - Tweets are sometimes ambiguous

Future Work



Icebox

- ★ Sentiment for each trend
- ★ Relate trends to current news
- ★ Related trends for each trend
- ★ Related trends based on my search (Natural language)
- ★ Explore trend progress
- ★ Sentiment over time for an specific trend
- ★ See my feed classified by topic
- ★ See "virality" metric for each trend
- ★ Personalized topic overview

Thank You!