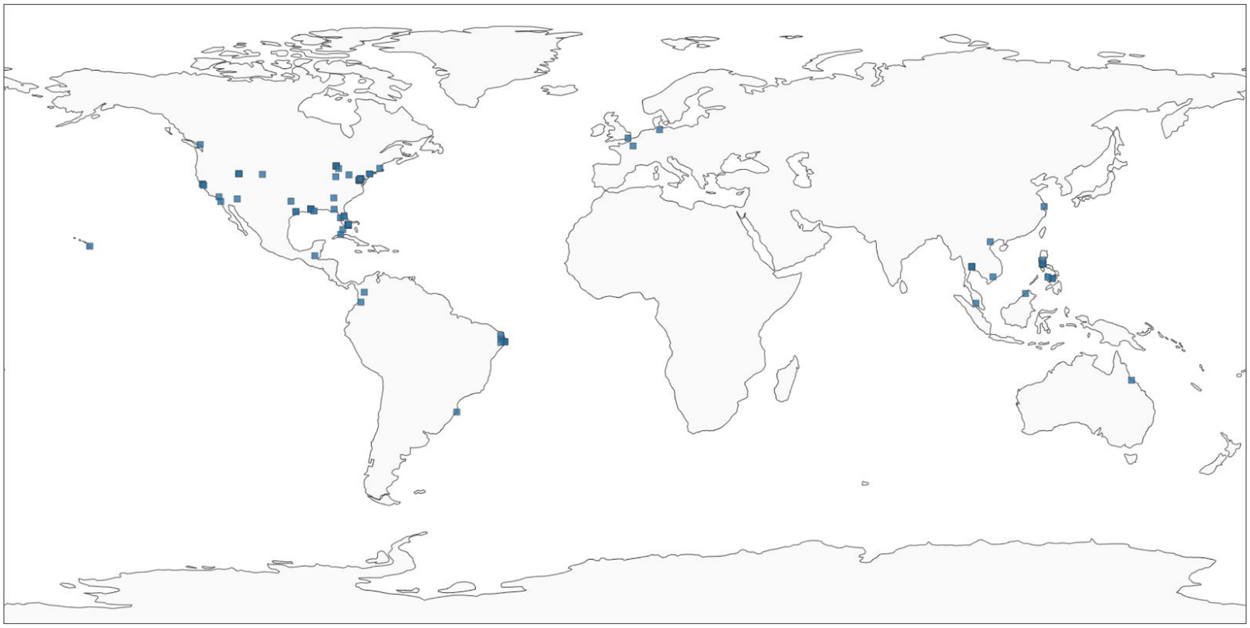
Clustering Articles Report

An analysis of news articles using the k-means clustering algorithm



David Hirschberg

June 6th 2017

INTRODUCTION

Given a tab separated file of news article titles, along with metadata including latitude, longitude, date published, and a generated topic, perform an exploratory analysis on the data and the hotspots found.

HYPOTHESIS

Using the data provided, we can identify unique events happening in hotspots around the world. Specifically, we can identify news articles keywords that are locally more significant than the rest of the world. Using a sliding window of single week, use tf-idf on ngrams of news articles published to determine a hotspot's top ngrams that are locally more significant than the rest of the world.

Results are expected to work moderately, but not very well, due to usage of k-means as the core algorithm, with not a very large dataset. We can make a slight improvement using the k-means++ initialization. Connectivity-based clustering, also known as hierarchical clustering may provide a good alternative.

DATA

First 3 rows of data

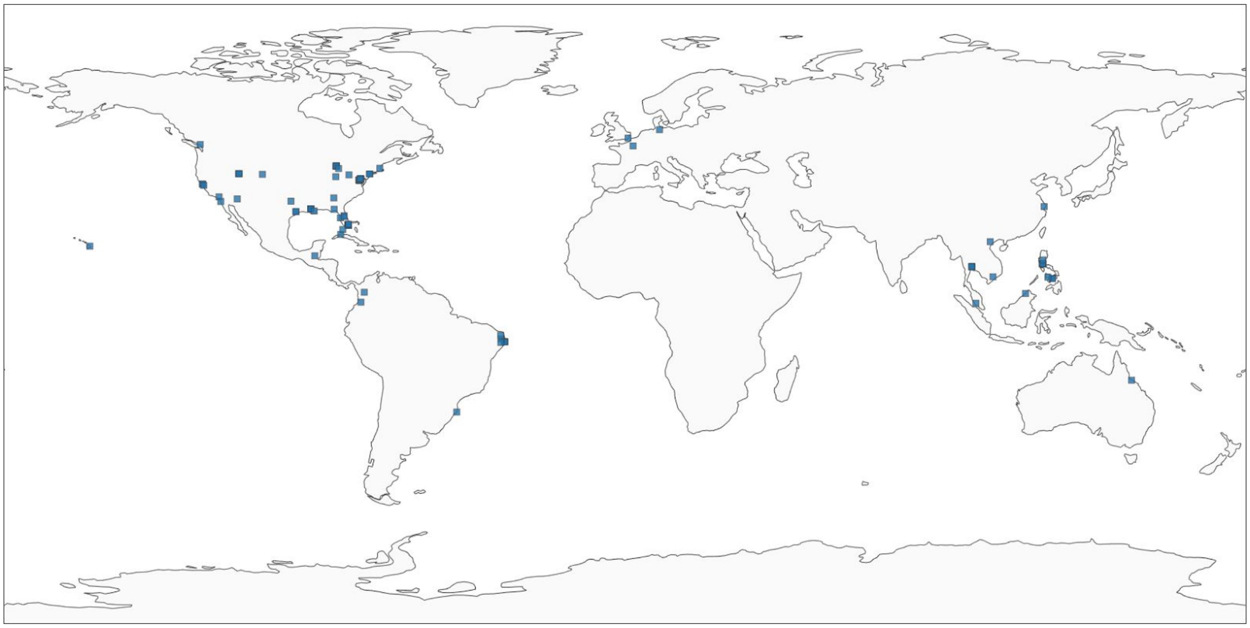
Latitude	Longitude	Title	Date	Topic
25.77427	-80.19366	Fêting Leo DiCaprio’s climate change doc in Miami	2016-10-06	Hurricane Matthew & Standing Water
25.77427	-80.19366	With $1.1 billion in new funding, U.S. health officials outline plan for fighting Zika	2016-10-03	Money & Vaccine Trial
25.77427	-80.19366	Congress approves $1.1 billion in Zika funding	2016-09-28	Planned Parenthood & Senate Republicans

PROCEDURE

1. Plot the locations on a map
2. Cluster the locations into "hotspots" using the k-means clustering algorithm
   a. Write up the clustering results
      i. Describe how well the clustering algorithm worked
      ii. How was K selected
      iii. Describe how the final choice was made on what results were best
   b. If any are thought of, describe other algorithms thought be a good alternative to the k-means clustering algorithm
   c. Plot the hotspots on a map
3. Perform exploratory analysis on the clusters using the provided metadata
   a. Explore tf-idf with ngrams
      i. Stop Words
      ii. Titles and Topics
   b. Explore time when published

RESULTS

1. Plot Locations on a map



2. Cluster the locations into "hotspots" using the k-means clustering algorithm

a. Write up the clustering results

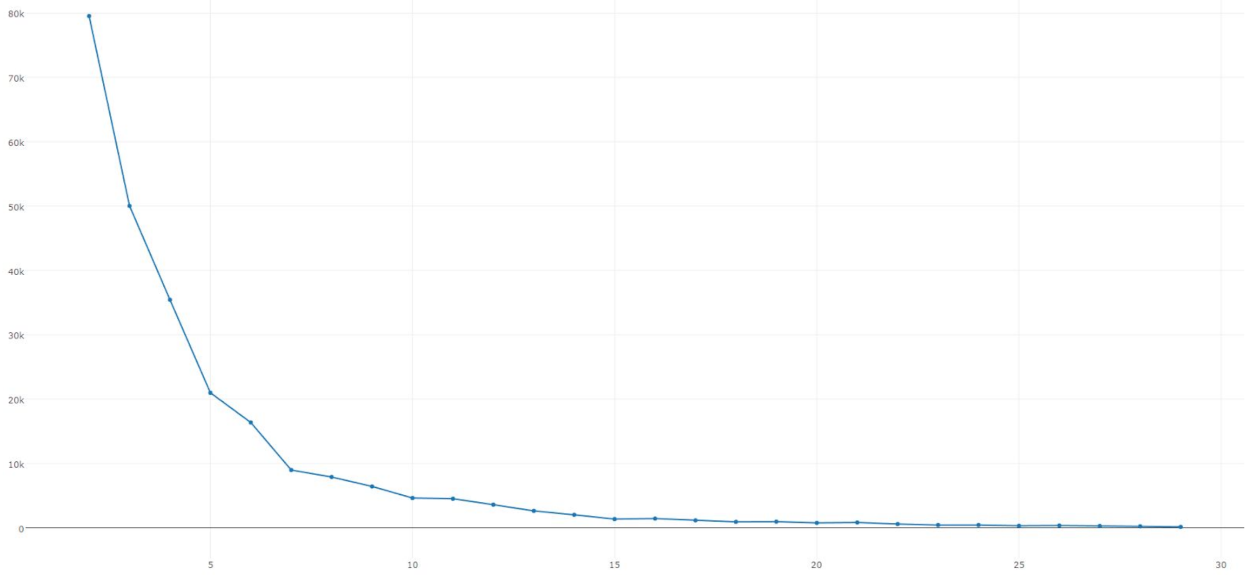
i. Describe how well the clustering algorithm worked

Hotspots are moderately accurate. We can at a glance see where the hotspots are. For the clustering on the entire dataset, we can see hotspots in the Western United States, New England, Southern United States/Florida, Hawaii, Brazil, Europe (hotspot in Belgium), and Philippines.

Visually, we can see that improvements are needed, south east Asia could easily have the single hotspot split into two. The Southern United States could also be split up as well.

ii. How was K selected

Two rules were used. For the complete dataset, the Elbow method was used, that is find the elbow point on the graph for k versus Sum Squared Error (SSE) . For the sliding window, there is a colloquial rule of thumb for k-means to use $\sqrt{n/2}$
K = 7 was chosen from the Elbow method.



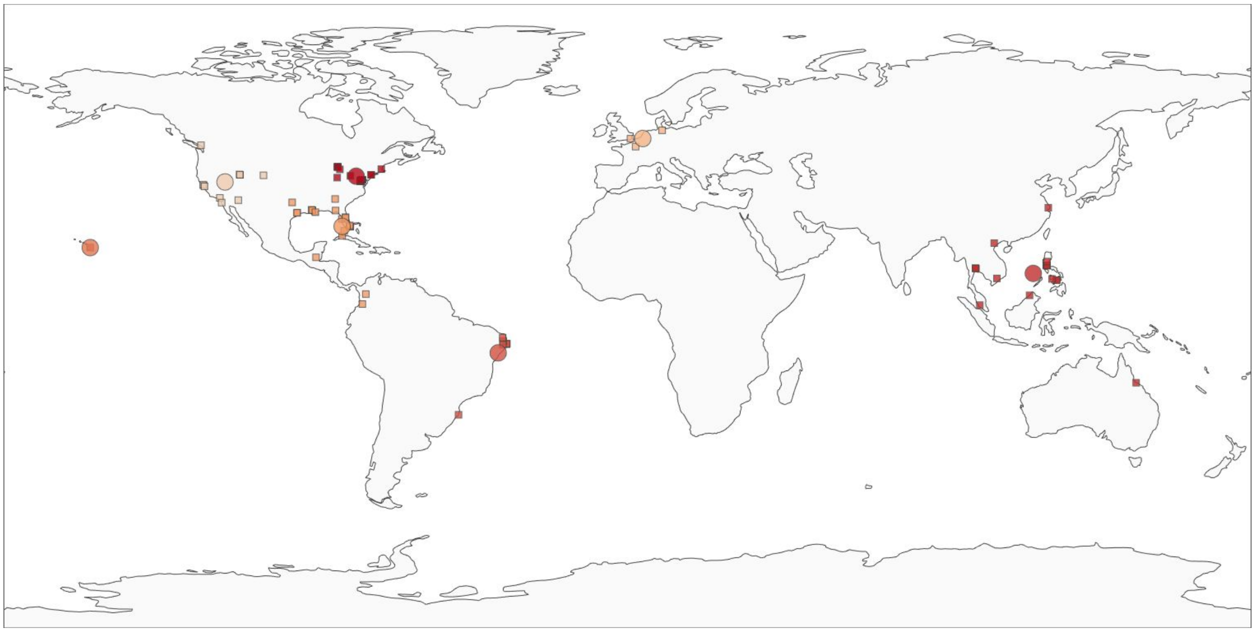
iii. Describe how the final choice was made on what results were best

Using the Elbow method gave relatively decent results, especially when the k-means++ centroid initialization was used. The combination of these two methods prevented hotspots from appearing in the middle of the Atlantic Ocean

b. Describe other algorithms thought be a good alternative to the k-means algorithm

I would like to explore the usage of connectivity-based clustering, also known as hierarchical clustering. Specifically, I would like to see the results of using Complete-linkage clustering or UPGMA or WPGMA (**U**n/**W**eighted **P**air **G**roup **M**ethod with **A**rithmetic Mean). If we ignore clusters that have less than a threshold of data points, (in our use case, could be 2), we can have clusters that are more geographically relevant.

c. Plot the hotspots on a map



3. Perform exploratory analysis on the clusters using the provided metadata

a. Explore tf-idf with ngrams (For the entire dataset)

i. Stop Words

A stop words list was used
A morphological structure change on tokens was not used

ii. Titles and Topics

Titles and Topics are both used in the ngram and tf-idf calculations

Interesting hotspots:

1. Western United States:
    a. CDC dismisses whistleblower
    b. Lanciotti Lab
2. Brazil
    a. Health problems
    b. Van der linden problems
3. Florida
    a. Hurricane Matthew standing water
    b. Beach mosquitoes
4. New England
    a. Senate approved funding government
    b. Money vaccine trial

More can be seen on interactive graph

b. Explore time when published

Same procedure as above, but applied to a sliding window of a single week length. An interactive graph is available.

Interesting hotspots:

1. 10-3 to 10-10
    a. Western United States
        i. First Case Zika Spreading
        ii. Zika Related Neurological Damage
    b. Brazil
        i. 62 cases zika
        ii. men possible
        iii. zika related neurological

2. 10-12 to 10-19
    a. South East Asia
        i. Zika infections epidemic
        ii. ho chi minh province
        iii. 2 cases zika infection
    b. England
        i. Eggs mosquito capable carrying
        ii. Virus found

CONCLUSION

Using the data provided, we can identify unique events happening in hotspots around the world. Specifically, we can identify news articles keywords that are locally more significant than the rest of the world.

Using a sliding window of single week, using tf-idf on ngrams of news articles published to determine a hotspot's top ngrams that are locally more significant than the rest of the world.

Results are moderately effective, but due to the usage of k-means as the core algorithm, with not a very large dataset, better results are lacking. K-means++ initialization and the Elbow method did somewhat have relatively better results.

We can quickly extract hotspots from around the world and see what makes this hotspot different from others. When combined with a list of global top ngrams, we can see both a global and local picture.

FUTURE WORK

Connectivity-based clustering, also known as hierarchical clustering may provide a good alternative.

NOTES

Plotly used to make graphs

to install "pip install plotly"