A Network Tour of Data Science

Authors:
Khalil Elleuch
Mariem Belhaj Ali
Mouadh Hamdi
Nour Ghalia Abassi

Ecole Polytechnique Federale de Lausanne

Lausanne, 16 January 2019

Introduction

This report covers the project we have done in the scoop of the course Network Tour of Data Science. During the semester we worked with the terrorist attack database in which we had covered almost all the network aspects. This project aim was to summarize all the techniques and tools seen in lectures specially the graph and network data aspects. To do so, we decided to work on a fresh data called Global Terrorism Database which contains more entries and features than the data we used during milestones. However, we constructed a graph from this dataset similar to the one already provided in order to stay consistent.

Overview

Our aim is to anaylse the data set in order to find terrorist communities. By terrorist communities, we mean that we want to find relationships between the attacks and the terrorist groups even if they don't belong to the same group. In other word, we want to find clusters of terrorist groups.

Motivation

The fight against those who would purposefully hurt or kill innocent civilians has been always at the forefront. Acts of terrorism across the globe have increased markedly in recent decades, in most parts of the world. It continues to be a relatively rare event and is instead focused in particular countries or regions of instability. Our aim is to better understand those attacks, the relationships between them, the involved countries and the cibled targets.

Data Description

In this section we will explain our motives of using the Global Terrorism Database. During the milestones and while working with the provided data set, we noticed that our graph is highly disconnected. We had multiple disconnected components. So, we looked for an other data to enrich our dataset in order have a more connected graph and found the Global Terrorism Database (GTD).

The Global Terrorism Database is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.

Data Wrangling

Data Cleaning

We pre-process the dataset to obtain a cleaned one with the columns we care about and easy-to-use names.The dataset contains a huge number of "Unknown"s in Group Name (Group Name designed the groups involved in the terrorist attacks). We deleted about 43% of the data due to two reasons: They have a huge proportion in dataset and learning them doesn't make any sense.

Data Clustering:

In order to get an idea about how the data can be modeled we tried to run k-means clustering. At this part we didn't construct any graph we are using the data as it is except some data cleaning.

- We selected some features that we thought that will be the most relevant to us (longitude, latitude, nwound, nkill,natlty1_txt, targtype1_txt, targsubtype1_txt, weaptype1_txt, attacktype1_txt).
- We handled categorical data.
- we removed outliers in nwounds and nkills.

After going threw the preprocessing above, we trained our data on kmeans model. The main idea was to find clusters, try to construct the world map using longitude and latitude attributes and see if our kmean can cluster the data this way. Below, you can find 4 different clusters depending on the number K of clusters.

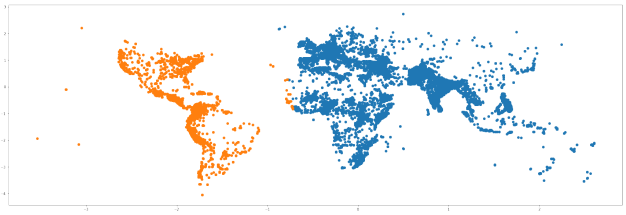

Figure 0.1: K=2

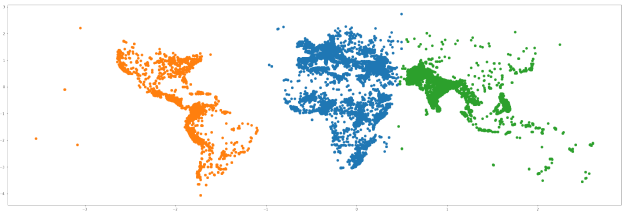

Figure 0.2: K=3

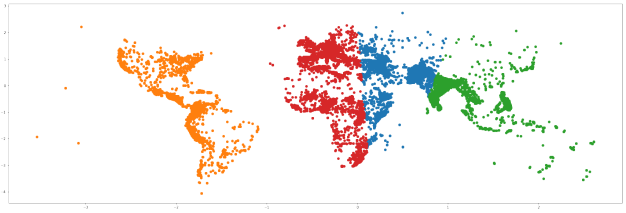

Figure 0.3: K=4

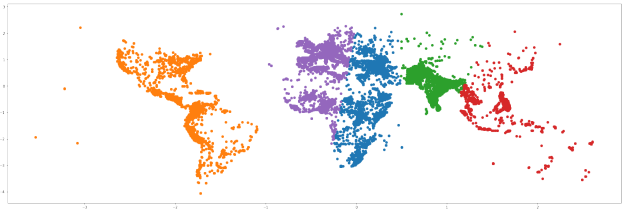

Figure 0.4: K=5

Data Visualisation

In order to visualize some variation of the data features we decides to highlights the evolution of the attacks over the years as you can see in the plot below.

We can see from the graph that independent of the country/ region the number of the terrorism attacks get increased exponentially in the past 20 years.

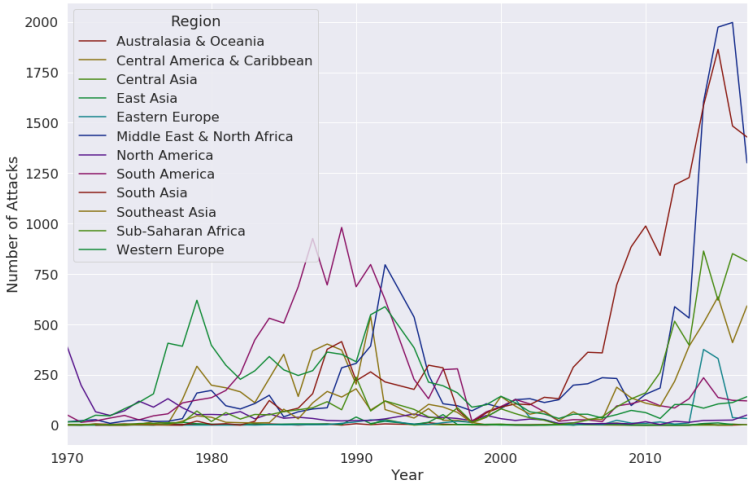

Figure 0.5: Numbers of kills

Concerning the number of attacks and kills, we plotted 2 graphs. The first one present the number of kills in a world map using folium. We can see that all people around the world were murdered because of terrorism and the number of kills differ slightly. The second graph represents the percentage of attacks in each country in 2016. We can see that Iraq, Afghanistan and India are the most involved countries in terrorism.

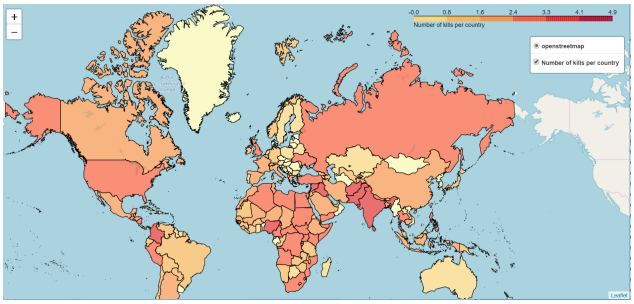

Figure 0.6: Number of kills

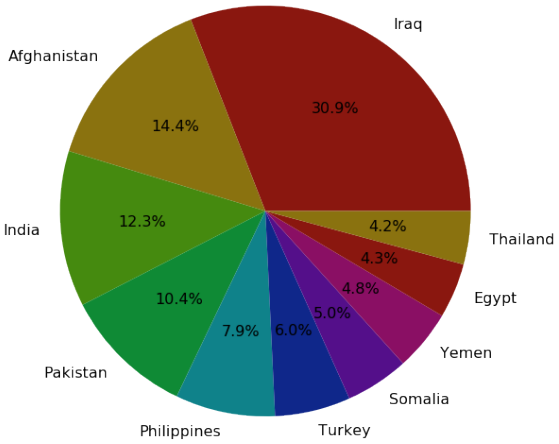

Figure 0.7: Percentage of attacks in each country in 2016

Graph Construction:

In this section we will explain how we constructed our graph from GTD. The main idea was to create a bipartite graph out of the organization name given in gname feature.

- We created an edge between two organizations if they attacked the same target.
- If the two groups attacked more than one target then the edge is a weighted with the weight equal to the number of common targets.

Using the above technique we will try to find out if there is relations between the terrorist group and the target. We can see that if a group always attacks the same targets, this graph can also capture if there is a collaboration between terrorist group.

Network Analysis:

Node importance:

Page Rank:

PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages. It make a sens that we use it in our study since we can consider targets as web pages.

Degree Centrality:

Degree centrality starts with the assumption that the person with the most connections (edges) is the most important. Rather than returning a count it is the degree of the node divided by the total possible number of edges that the node could have. For the case of the directed graph the degree of the incoming vertices and outgoing vertices would likely be treated separately.

In other words centrality algorithms determine a node's relative importance within a graph by looking at how connected it is to other nodes. It is used for instance to identify key players within organizations.

Clustering:

Community detection:

First, we worked with the Louvain community detection algorithm.It is a method to extract communities from large networks. The method is a greedy optimization method that appears to run in time O(n$\log n$).

Unfortunatelly, we didn't get anaything releavent using this algorithm. There is no real seperation between nodes and no clusters were detected.

Spectral clustering

spectral clustering make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. In our case, we get an interesting result : we found 5 clusters using spectral clustering.

Machine Learning:

Part 1:

In this section we will use some machine learning analysis. Our aim was to predict the 'Unkown' groups which we considered as our missing data. First let's start by running world cloud text analysis from the text contained in the addnotes attributes. We plotted the results of our training in the following picture.

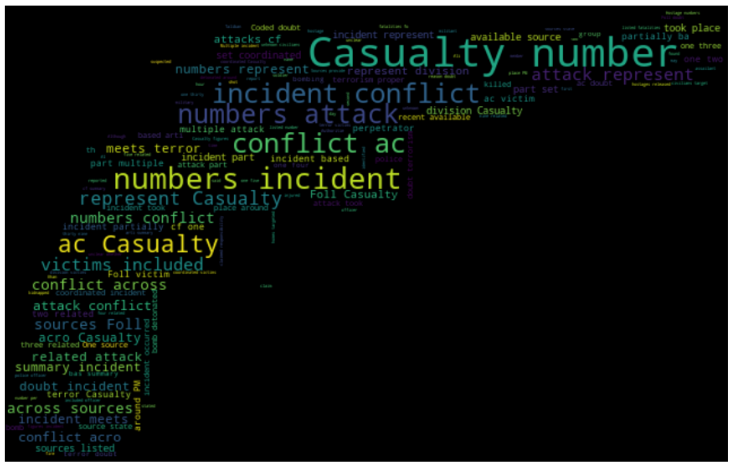

Figure 0.8: Word Cloud

At this step we will predict the 'Unkown' values, we started by trying to predict the missing values out of all gname groups which they where equal to 70385. This gave us an accuracy of 0.62.

Part 2:

Our purpose in this section is to predict the unknown gnames using Machine learning and the spectral clusters made so far: In order to have a better accuracy we will use the construction of graph. Instead of predicting out of all the gname we will try to find out to which cluster this 'Unkown' belong.

We will split our dataset in two groups, we will take off the unknown gname from the dataset, we will train a random forest classifier on the data that we have using the cluster given by the spectral clustering made in the previous section.

Results: 0.88 accuracy to assign each unknown gname to a cluster of gnames that it belongs to.

Conclusion:

As a conclusion we can say that the network analysis is a powerful tool that allowed us to analyse aspect of the data that we couldn't investigate by running some machine learning procedures or by running some stats. As general conclusion about the data set we can say that terrorism is attacking every country in the world so governments need to rise awareness and to invest more in fights against terrorism