TERROROSM DATASET:

==========

==========

1. CONTENT

Information on more than 180,000 Terrorist Attacks

The Global Terrorism Database (GTD) 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

2. ABOUT THIS FILE

Geography: Worldwide.

Time period: 1970-2017, *except 1993.*

Unit of analysis: Attack.

Variables: >100 variables on location, tactics, perpetrators, targets, and outcomes

Sources: Unclassified media articles (Note: Please interpret changes over time with caution. Global patterns are driven by diverse trends in particular regions, and data collection is influenced by fluctuations in access to media coverage over both time and place.)

Definition of terrorism:

"The threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation."

See the GTD Codebook for important details on data collection methodology, definitions, and coding schema.

As we have 135 columns in the data, the 10 sample columns are:

1. Eventid: A 12-digit Event ID system. First 8 numbers – date recorded “yyyymmdd”. Last 4 numbers – sequential.
2. iyear: This field contains the year in which the incident occurred.
3. imonth: This field contains the number of the month in which the incident occurred.
4. day: This field contains the numeric day of the month on which the incident occurred.
5. Approxdate
6. extended:1 = "Yes" The duration of an incident extended more than 24 hours. 0 = "No" The duration of an incident extended less
7. resolution
8. country: This field identifies the country code
9. country_txt: This field identifies the country or location where the incident occurred.
10. region: This field identifies the region code based on 12 regions

3. PROBLEM STATEMENT

The dataset is very valuable as it can be used to answer the following questions:

- A big picture of terrorism around the world and its evolution over the years;
- Countries with most incidents recorded;
- Countries with highest number of victims;
- A dashboard for terrorism analysis in some countries;
- Incidents that lasted more than 24h (extended = 1);
- Major radical groups responsible for terrorist attacks (gname);
- Attacks with the highest number of terrorists (nperps);

4. ROUGH APPROACH

- Extensive EDA and research from GTD.
-Time Series analysis based on the data.
-Beginner level NLP implementation: Regular expressions, e.g.: to replace special characters with white space in the columns, tokenization, Regex.
-Usage of kmeans clustering in order to cluster the terrorist attack based on geographic location, number of victims, Nationality of the perpetrator, target type, weapon type, attack type.
-Feature engineering and profiling each cluster.

5.SOURCES:

KAGGLE:

HTTPS://WWW.KAGGLE.COM/START-UMD/GTD

GTD DATABASE:

HTTPS://START.UMD.EDU/GTD/

GTD CODEBOOK:

https://start.umd.edu/gtd/downloads/Codebook.pdf

PUBLICATIONS:

https://www.start.umd.edu/publications?combine=&author%5B%5D=13781&year%5Bvalue%5D%5Byear%5D=