Global Terrorism

Motivation and Goals:

Terrorism is described as the violent and inhuman act that get perpetuated for political, ideological and religious goals with the aim of creating fear among the neutral citizen. It is not only the physical attack by itself but also the psychological impact it has on a society for many years after. In the past years the attention has risen. Even our database we use contains more than 170,000 cases.

We will use the Global Terrorism Database (GTD) taken from Kaggle, which is „an open-source database including information on terrorist attacks around the world from 1970 through 2016", as quoted from the description. We download the data set once and do not use streaming processing to add any new data sets. It contains more than 170,000 entries with more than 100 attributes on location, tactics, perpetrators, targets and outcomes.

Our goals are to analyze where terrorist attacks are most common, which attacks are most deadly and with which weapons, and to distinguish between the targets of the terrorist attacks.
Since terrorist attacks usually are highly dependent on the political situation, we will sort the data sets by continent and decade, since the political situation depends on time and place.

Then we will analyze what weapons cause the deadliest incidents, whether there is correlation between the time frame of the incident and the deadliness, what targets are usually targeted by what weapons, and also check what other correlations there may be. As work progresses, we might be interested in other directions of analysis or discard some of the old questions, if we find them less intriguing.

Information about the incidents and exploring non-obvious connections might be helpful in determining where the next attacks might be and what characteristics they have. To prevent terror attacks, one needs to recognize what type of weapons may be used against what targets to know what to look out for.

To summarize, our aim is to examine trends about the future terrorism with the analysis from the Global Terrorism Database.

Implementation requirements:

We will be using the Global Terrorism Database (GTD) data set taken from Kaggle, which is „an open-source database including information on terrorist attacks around the world from 1970 through 2016", as quoted from the description. We will download the data set once and do not use streaming processing to add any new data sets. The link for this data set is here: https://www.kaggle.com/START-UMD/gtd/data

We will separate the data by decade and region, then use clustering to separate similar attacks with the k-means method into groups. It will help us figure out, what groups of attacks have the tendency to yield a higher death toll. We will also separate the data in the same way by decade and region, but also separate by weapon type, target type or type of attacker. When we determine the average death toll for each group, we can figure out, if these categories make sense for figuring out the attacks with highest severity, and if the clustering method gives betters results in comparison.

With over 100 attributes for this data set, we will need to filter out most of them. Particularly detailed text descriptions are useless for clustering. Similarly, attributes that give names of the individuals targeted do not assist us in any way. In addition, many attributes are doubled as a number and as a text description, for example the entry for attribute „targtype1_txt“ reads as „Private Citizens & Property“ and corresponds to attribute „targtype1“ entry „14“. Since numeric values like integers occupy less space, it would be advantageous to keep only attributes like „targtype1“ whenever we can. Apart from these measures, we may still need to throw out many more attributes to make the dataset manageable to work with.

For evaluation we will be showing clustering and regression graphs with Python and describe our findings. We will be using the Python libraries numpy, scipy, sklearn and matplotlib, and possibly more.

Architecture:

We will be using the Hadoop infrastructure for organization of data storage, with Hbase used for Online NoSQL. Yarn will be used for workload management. These choices were made, because these are the tools we have been introduced to in class.

We have chosen to use Hive for Batch Processing. Since we are using a static data set, there is no need for stream processing tools. There is the possibility of choosing Impala instead of Hive, but due to familiarity with Hive and since Hive is commonly used for batch processing, we have settled on Hive for now. If we run into too long processing times, we will consider switching to Impala.

Similarly, we will use Hive for Analytical SQL. For clustering and portrayal in graphic format, we will be using Python. Sqoop will be our tool of choice for Data Integration, because it seems to us be a tool best used for batch processing, unlike Flume.