Project Proposal: Predicting perpetrators of global terrorism

Supratik Chanda and Joseph Ditton

Description and motivation of what we plan to do

Terrorist attacks are one of the biggest problems for mankind across the world. Stopping global terrorism needs the whole attention of researchers and practitioners to come up with a decisive and actionable plan. Gathering and using historical data about terror events is a complicated task due to the lack of detailed or organized terrorist data. Today, more than ever, predicting and preventing terrorist attacks is of the utmost concern to the intelligence community. Trying to find the 'needle in a haystack' of huge amounts of data is as challenging today as it has ever been. Predicting characteristics of terror attacks based on historical data on terror attacks may help prevent them in the future. This approach captures the intentions of terrorists because terrorist groups are often very careful and deliberate when planning and executing such attacks.

For this research project, we will heavily use Wikipedia to extract terror events. The goal is to find the following information: type of terror (bombing, shooting, stabbing ect), number people killed, a detailed text description of the event, name of the perpetrator or group, and the location and time of the attack. This project aims to predict two aspects:

1) Using the methods and descriptions of terrorists events, predict the potential terror groups or perpetrators that are responsible for attacks. Our dataset will include global terrorism events the year 1970 up to 2019. Accurate predictions may help government agencies predict possible future attacks and who will commit them.
2) Predicting the terror pattern/methods/tactics of various terrorist groups and identifying probable terror groups connected to similar events in future. This can be used to infer probable terror groups' name hiding in an area.

There are few more relevant questions that we can derive:

1) Has the terrorism grown over the years?
2) How the terrorism pattern changed since 1970?
3) Which countries have lost most lives for terrorism? And many more...

The project in a nutshell, identifies key areas of intense research interests. These include analyses of terrorist attack trends of perpetrators.

The possible techniques or approach that we will take:

1) Web scraping for collecting data

2) Data Mining,cleansing and wrangling
3) Natural Language Processing for text analysis
4) Multiclass Classifiers
5) Standardization or Normalization
6) GridSearchCV, RandomizedSearchCV and cross-validation
7) Logistic Regression
8) Pipeline
9) DecisionTree Classifier
10) GaussianNB and MultinomialNB
11) WordCloud

Schedule

Wednesday, April 10th: Proposal due
Friday, April 12th: Have all of the data scraped and cleaned, and complete some basic analysis
Monday, April 15th: Finish analysis
Wednesday, April 17: First draft of report and any tweaks to analysis
Friday, April 19th: Final project done

Division of responsibilities among the team

Together:

1) Evaluate the best algorithm in terms of f1_score and accuracy. Produce a confusion matrix and classification report.
2) Improve the f1_score if possible with the help of GradientBoostingClassifier. This is because of the parameter known as " learning rate".

Joseph:

1) Web scrape the terrorist activities from 1970 to 2019
2) Data preprocessing, cleansing on the web scraped data
3) Use clustering to group related terrorist events
4) Use neural network to perform classification

Supratik:

1) Perform NLP techniques on terror methodologies to extract important words for text analysis
2) Utilize feature extraction techniques to correlate the importance of specific patterns or words with the perpetrator's names.
3) Use Data Mining techniques for pattern recognition after NLP techniques
4) Apply ML algorithms to find the best algorithm fit for predicting of perpetrators.Along with that, we select the optimal model parameters using K-fold cross-validation and GridSearchCV or RandomizedSearchCV whichever take the least execution time.
5) Use "make_pipelines" technique to sequence the standardization, regularization and playing with the hyperparameters of ML algorithms all at one time.

Identification of potential problems that you may encounter and present alternative approaches if necessary

Potential problems that may come up while analysing and predicting:

1) To predict terror groups based on their terror activities, we need distinct terror patterns for every perpetrator. While analyzing trends of terror methods used by various groups, we may land up with common methods or patterns used by different terror groups. This may obscure our prediction.
   Alternative Approach:
   We might use a frequency indicator or a term frequency identifier that will show which terror words are used more often to which terror groups. Linking word frequency with the terror groups can give us a more probable prediction of terror patterns with the related terror gangs
2) Too much terror methods description for each terror group. Unwanted words like grammar are present which might lower the prediction accuracy
   Alternative Approach:
   Remove the stopwords and let the prime words in the sentence.
3) We don't know which machine learning algorithm would be best for the pattern Recognition and eventually prediction of perpetrators.
   Alternative Approach:
   We will use make_pipeline and gridsearchCV/ RandomizedSearchCV along with an execution time to predict with the highest f1_score, accuracy and lowest error rate.

Any preliminary Results:

The URL of each terrorist attack from 1970 to 2019 is currently being scraped from the Wikipedia.

In [63]:
```
webPageUrl =requests.get('https://en.wikipedia.org/wiki/List_of_terrorist_incidents')
allUrlList = soup(webPageUrl.text,'html.parser')
allUrl = allUrlList.find_all('div',attrs={'class':'div-col columns column-width'})
listOfTerroristIncidentsWrtTime=pd.DataFrame()
for each in allUrl:
    j= each.find_all('a',attrs={'href':True})
    for elem in j:
        listOfTerroristIncidentsWrtTime = pd.concat([listOfTerroristIncidentsWrtTime,
                                                     pd.Series('https://en.wikipedia.org/'+elem['href'])],axis=0)
    break
listOfTerroristIncidentsWrtTime.columns=['Year_of_Terrorist_attack']
listOfTerroristIncidentsWrtTime.reset_index(drop=True,inplace=True)
```

In [65]:
```
listOfTerroristIncidentsWrtTime.head()
```

Out[65]:

	Year_of_Terrorist_attack
0	https://en.wikipedia.org//wiki/List_of_terrorist_incidents_in_1970
1	https://en.wikipedia.org//wiki/List_of_terrorist_incidents_in_1971
2	https://en.wikipedia.org//wiki/List_of_terrorist_incidents_in_1972
3	https://en.wikipedia.org//wiki/List_of_terrorist_incidents_in_1973
4	https://en.wikipedia.org//wiki/List_of_terrorist_incidents_in_1974

Link: https://en.wikipedia.org/wiki/List_of_terrorist_incidents

References

a) "***HYBRID CLASSIFICATION ALGORITHMS FOR TERRORISM PREDICTION in Middle East and North Africa***" ---Motaz M. H. Khorshid , Tarek H. M. Abou-El-Enien ,Ghada M. A. Soliman Faculty of Computers & Information, Cairo University, 5 Dr. Ahmed Zoweil St., Orman, Giza 12613, Egypt

b) "***A Conjoint Application of Data Mining Techniques for Analysis of Global Terrorist Attacks.Prevention and Prediction for Combating Terrorism***"---National University of Science and Technology-MiSiS, Moscow, Russian Federation.Innopolis University, Kazan, Russian Federation

c) "***New framework uses patterns to predict terrorist behavior*** "--- Binghamton University

d) "***Can We Predict Where Terrorists Will Strike Next?***"----by Brian Michael Jenkins