Predicting the group of attackers based off attack details.

Objective:

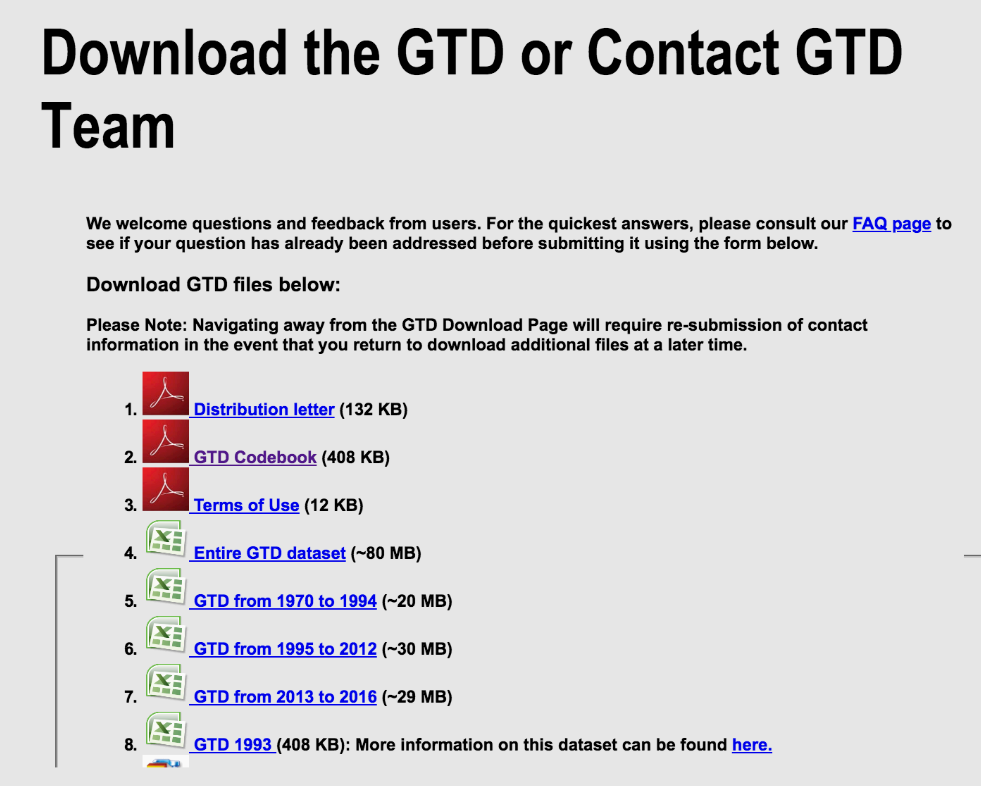
- Currently there are many terrorist attacks in the database that still have no group attributed to them if no group claimed responsibility, we wanted to create a predictive model that can possibly classify these still unknown groups?
- This project is going to be a large multiclass classifier that can use structured data and unstructured data in combination as features, to classify which terrorist group performed which.

Main things to discuss:

- Data cleaning
- Accuracy on just the recent data with structured data
- Accuracy on recent data with structured and unstructured
- Accuracy on data since 1995 with structured and unstructured.
- compare random forest bow/tfidf, logistic regression with bow/tfidf compared to svm bow/tfidf

Reading in the data

- First, we have a large data set, which can be seen by visiting this link: https://www.start.umd.edu/gtd/contact/ but we also



Download the GTD or Contact GTD Team

We welcome questions and feedback from users. For the quickest answers, please consult our FAQ page to see if your question has already been addressed before submitting it using the form below.

Download GTD files below:

Please Note: Navigating away from the GTD Download Page will require re-submission of contact information in the event that you return to download additional files at a later time.

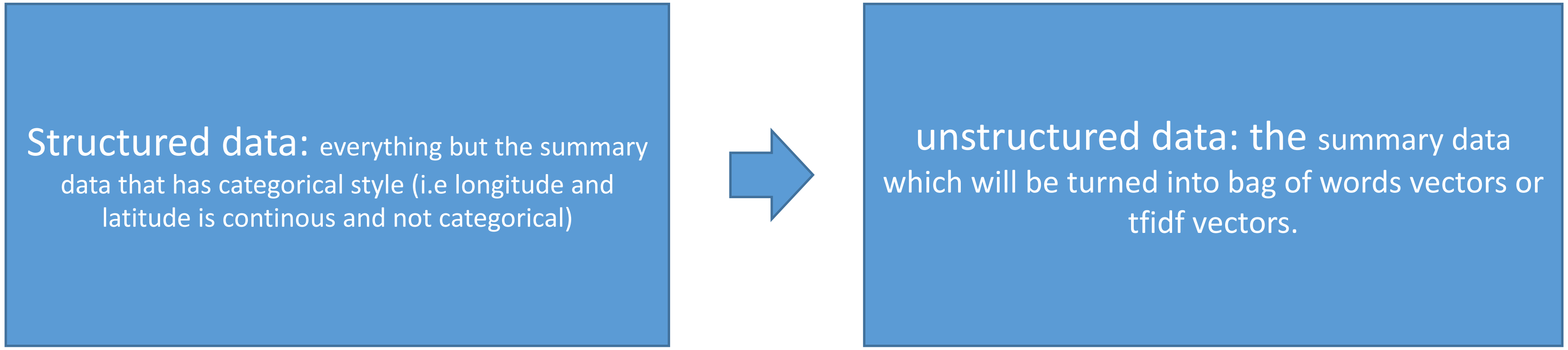
1. Distribution letter (132 KB)
2. GTD Codebook (408 KB)
3. Terms of Use (12 KB)
4. Entire GTD dataset (~80 MB)
5. GTD from 1970 to 1994 (~20 MB)
6. GTD from 1995 to 2012 (~30 MB)
7. GTD from 2013 to 2016 (~29 MB)
8. GTD 1993 (408 KB): More information on this dataset can be found here.

We see the available files to us are all the excel files, and we need to decide how many we are loading in. With just the most recent, we have fairly high accuracy. With the two most recent, we have slightly lower accuracy but not too bad.

The metrics we obtained were

Data:

- the Data came from the GTD otherwise known as the global terrorism databank which can be found in this link:



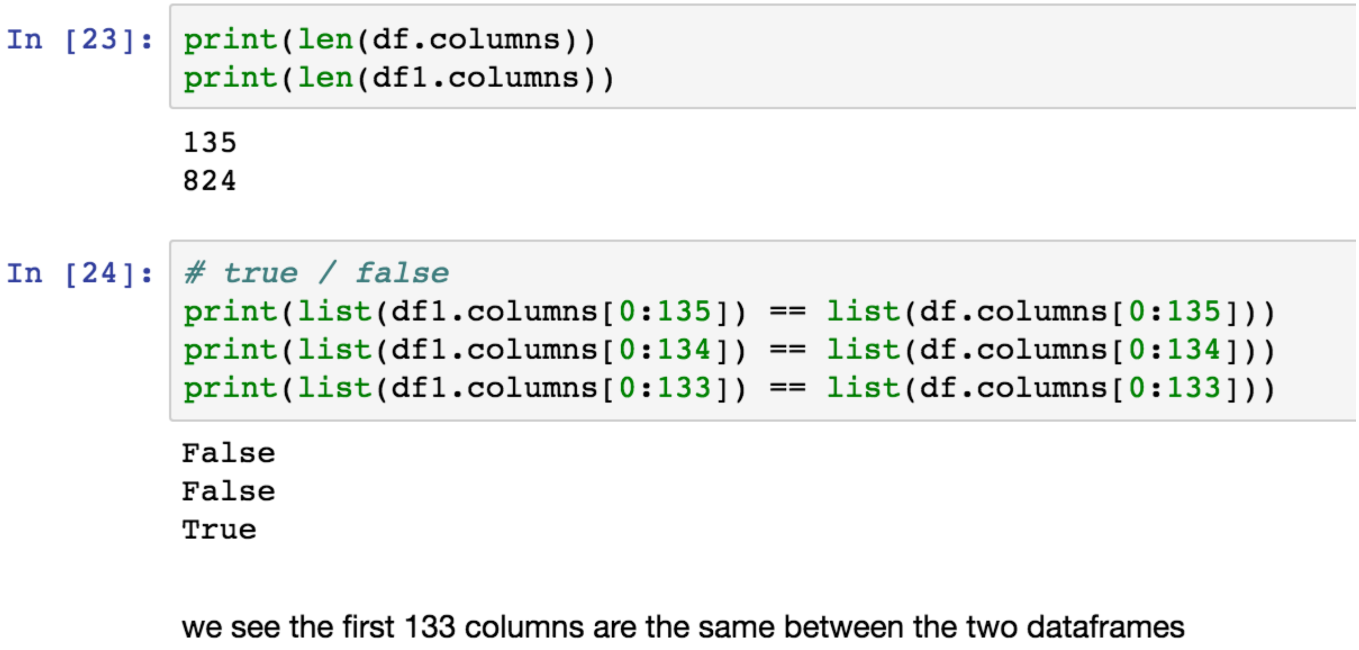
Structured data: everything but the summary data that has categorical style (i.e longitude and latitude is continous and not categorical)

unstructured data: the summary data which will be turned into bag of words vectors or tfidf vectors.

- First we wanted to look at our unstructured data and see what types of data we had of what. Our structured data model can take in any categorical data. Our unstructured data, like a summary in text form, can be turned into bag of words or tfidf features. We can then feed that into logistic regression or support vector machines as the classifier.

Appending the two dataframes:

- When using the two most recent (1995 -2013) and (2013-2017) We check out the two dataframes and make sure their columns are in the same in order to merge them.



```
In [23]: print(len(df.columns))
         print(len(df1.columns))

135
824
```

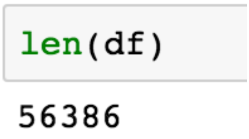
```
In [24]: # true / false
         print(list(df1.columns[0:135]) == list(df.columns[0:135]))
         print(list(df1.columns[0:134]) == list(df.columns[0:134]))
         print(list(df1.columns[0:133]) == list(df.columns[0:133]))

False
False
True
```

we see the first 133 columns are the same between the two dataframes

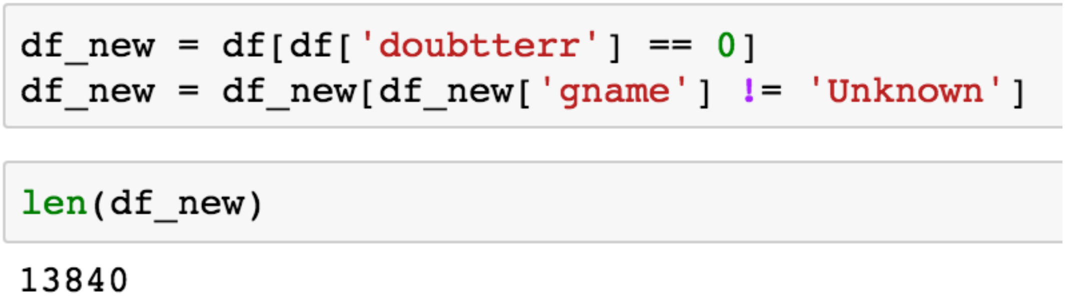
Data Cleaning and feature selection

- doubtterr == 0 : we filter data frame based off of this.



```
len(df)
```

```
56386
```



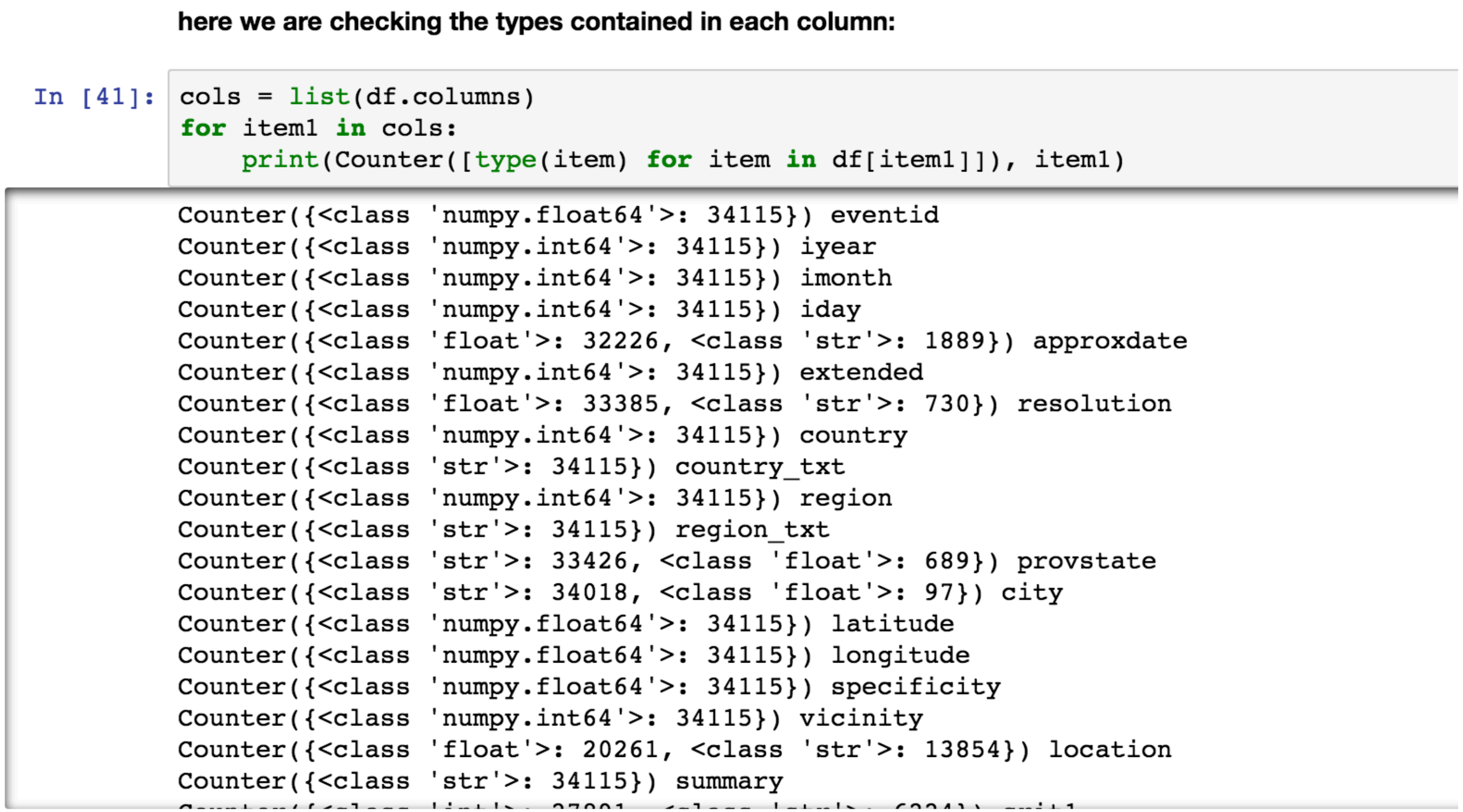
```
df_new = df[df['doubtterr'] == 0]
df_new = df_new[df_new['gname'] != 'Unknown']
```

```
len(df_new)
```

```
13840
```

Feature selection:

- We manually scrolled through , to check which data types were consistent across the board, which were all strings, which were all ints, etc. this way we will be able to feed them in easily. For feature importance



here we are checking the types contained in each column:

```
In [41]: cols = list(df.columns)
         for item1 in cols:
             print(Counter([type(item) for item in df[item1]]), item1)
```

```
Counter({<class 'numpy.float64'>: 34115}) eventid
Counter({<class 'numpy.int64'>: 34115}) iyear
Counter({<class 'numpy.int64'>: 34115}) imonth
Counter({<class 'numpy.int64'>: 34115}) iday
Counter({<class 'float'>: 32226, <class 'str'>: 1889}) approxdate
Counter({<class 'numpy.int64'>: 34115}) extended
Counter({<class 'float'>: 33385, <class 'str'>: 730}) resolution
Counter({<class 'numpy.int64'>: 34115}) country
Counter({<class 'str'>: 34115}) country_txt
Counter({<class 'numpy.int64'>: 34115}) region
Counter({<class 'str'>: 34115}) region_txt
Counter({<class 'str'>: 33426, <class 'float'>: 689}) provstate
Counter({<class 'str'>: 34018, <class 'float'>: 97}) city
Counter({<class 'numpy.float64'>: 34115}) latitude
Counter({<class 'numpy.float64'>: 34115}) longitude
Counter({<class 'numpy.float64'>: 34115}) specificity
Counter({<class 'numpy.int64'>: 34115}) vicinity
Counter({<class 'float'>: 20261, <class 'str'>: 13854}) location
Counter({<class 'str'>: 34115}) summary
Counter({<class 'int'>: 27891, <class 'str'>: 6224}) crit1
```

Methods for model training:

- Models trained on large dataset, data from 1995 to 2013 and then from 2013 to 2017, the same model trained on that data went down in accuracy but had a larger multi-class classification capacity.
- How model work on unstructured data like bag of words and tfidf vector features.
- In the end, we were able to combine the two features, being bag of words and the structured data

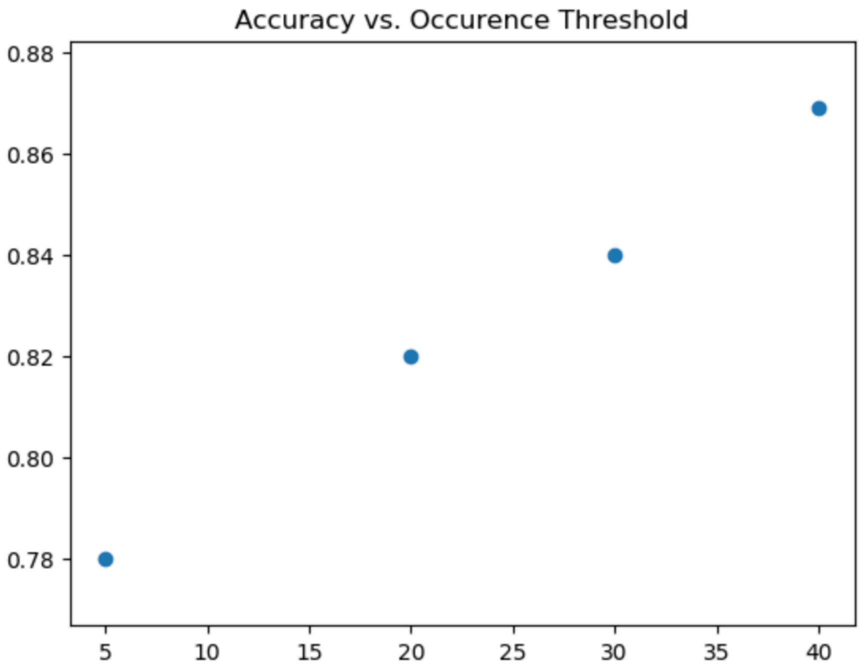
Threshold implications on our multiclass classifier list.

- When applying the threshold on the larger combined data set, we resulted in 121 labels that have sufficient training data
- When applying the threshold on the most recent dataset we resulted in 50 labels.

- Accuracy in our model went up with a higher threshold as expected

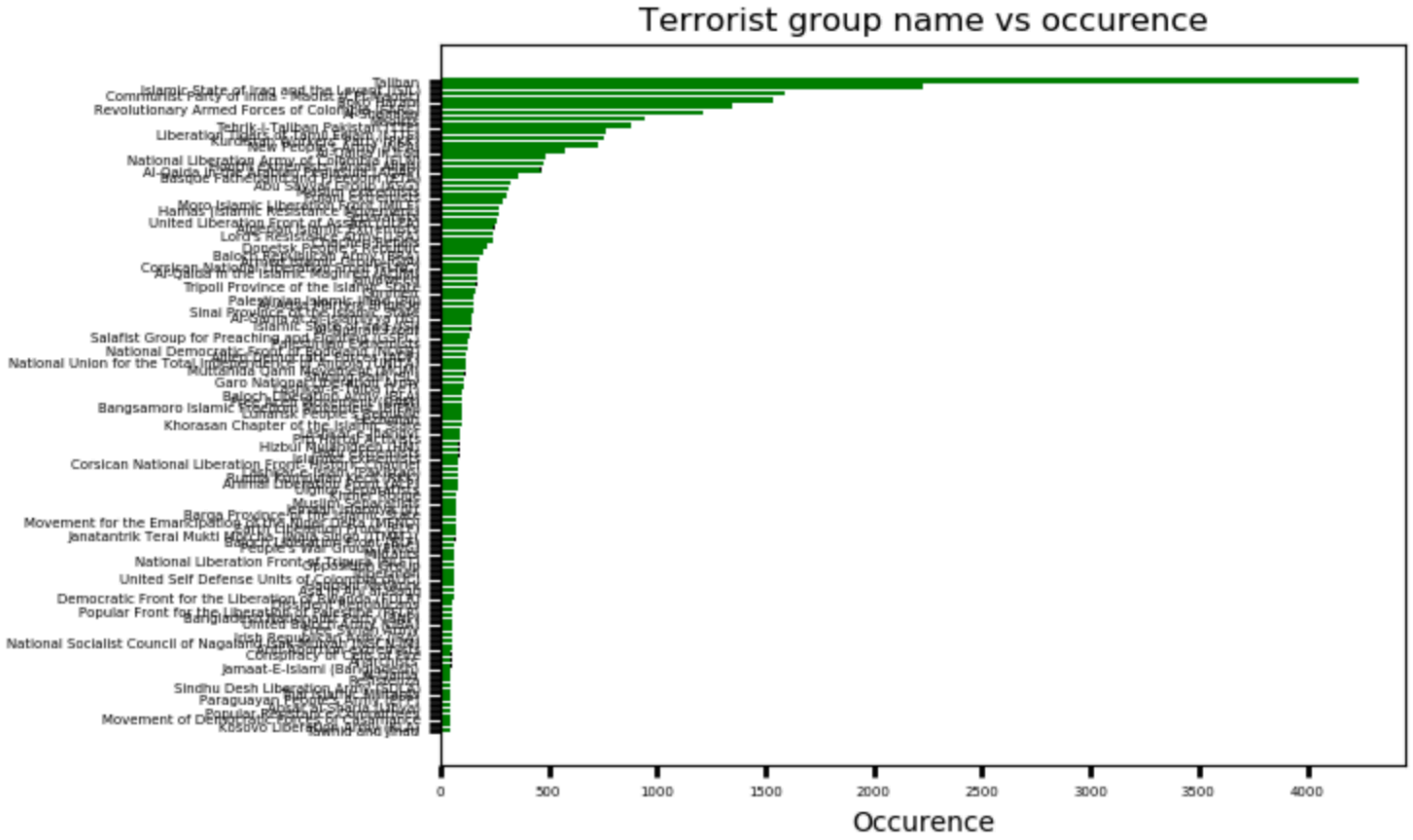
Threshold

- We arbitrarily chose 30 to be the threshold number where we will discard any group names from our test-train set that only have 30 instances or more.

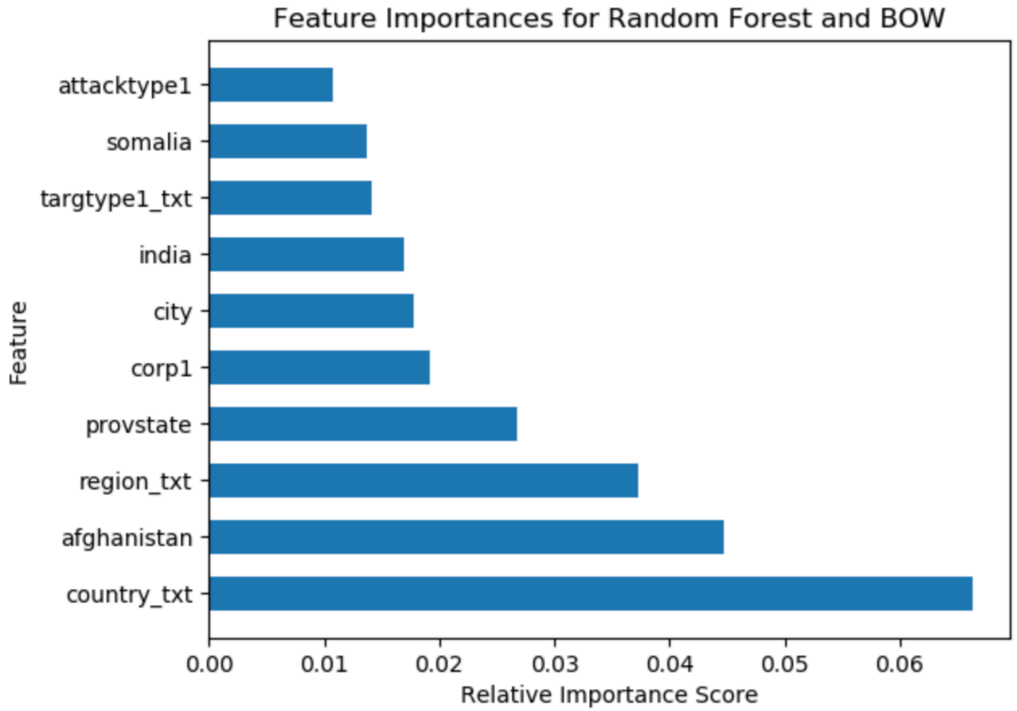


We then run some tests comparing threshold numbers. As you can see, the accuracy increased linearly with the threshold.

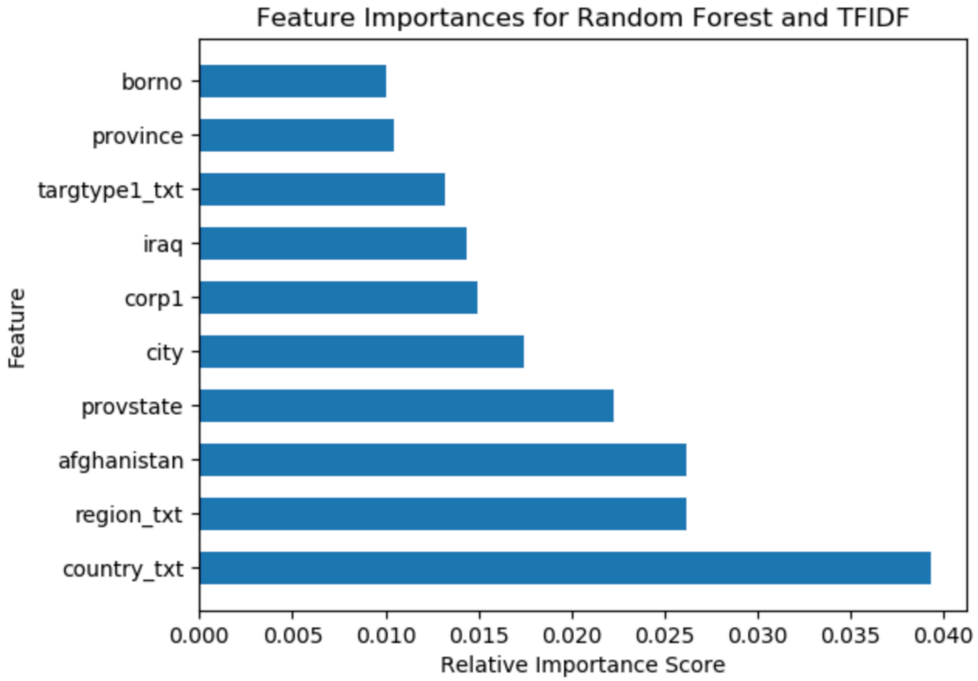
EDA


Terrorist group name vs occurence
Taliban
Islamic State of Iraq and the Levant (ISIL)
Communist Party of India - Maoist (CPI-Maoist)
Boko Haram
Revolutionary Armed Forces of Colombia (FARC)
Al-Shabaab
Tehrik-i-Taliban Pakistan (TTP)
Liberation Tigers of Tamil Eelam (LTTE)
Kurdistan Workers' Party (PKK)
New People's Army (NPA)
National Liberation Army of Colombia (ELN)
Al-Qaida in the Arabian Peninsula (AQAP)
Basque Fatherland and Freedom (ETA)
Abu Sayyaf Group (ASG)
Moro Islamic Liberation Front (MILF)
Hamas (Islamic Resistance Movement)
United Liberation Front of Assam (ULFA)
Lord's Resistance Army (LRA)
Donetsk People's Republic
Baloch Republican Army (BRA)
Corsican National Liberation Front (FLNC)
Al-Qaida in the Islamic Maghreb (AQIM)
Tripoli Province of the Islamic State
Palestinian Islamic Jihad (PIJ)
Sinai Province of the Islamic State
Salafist Group for Preaching and Fighting (GSPC)
National Democratic Front of Bodoland (NDFB)
National Union for the Total Independence of Angola (UNITA)
Muttahida Qami Movement (MQM)
Garo National Liberation Army
Baloch Liberation Army (BLA)
Bangsamoro Islamic Freedom Movement (BIFM)
Luhansk People's Republic
Khorasan Chapter of the Islamic State
Hizbul Mujahideen (HM)
Corsican National Liberation Front- Historic Channel
Animal Liberation Front (ALF)
Barqa Province of the Islamic State
Movement for the Emancipation of the Niger Delta (MEND)
Janatantrik Terai Mukti Morcha- Jwala Singh (JTMM-J)
People's War Group (PWG)
National Liberation Front of Tripura (NLFT)
United Self Defense Units of Colombia (AUC)
Democratic Front for the Liberation of Rwanda (FDLR)
Popular Front for the Liberation of Palestine (PFLP)
United Baloch Army (UBA)
Irish Republican Army (IRA)
National Socialist Council of Nagaland-Isak-Muivah (NSCN-IM)
Conspiracy of Cells of Fire
Jamaat-E-Islami (Bangladesh)
Sindhu Desh Liberation Army (SDLA)
Paraguayan People's Army (EPP)
Movement of Democratic Forces of Casamance
Kosovo Liberation Army (KLA)
Tawhid and Jihad
0
500
1000
1500
2000
2500
3000
3500
4000
Occurence

Top ten features in the model of random forest:



Here we have a model combining structured data and bag of words,



Here we have a model combining structured data and TFIDF

Methods for model training:

- Random Forest Classification: structured + BOW/tfidf
- Logistic Regression: structured + BOW/tfidf
- Support vector machines: structured + BOW/tfidf

- Random Forest uses label encoding for structured data
- Logistic Regression used one-hot encoding for structured data

Results for a large 121 multi-classifier data trained on data from 1994.

- We obtained Accuracy = 0.86 Random Forest Classification: structured + BOW
- We obtained Accuracy = 0.80 Random Forest Classification: structured + tfidf

- We obtained Accuracy = 0.92 for a large 121 multi-classifier data trained on data from 1994. Logistic Regression: structured + BOW
- SVM took too long to train in the end.

We found that if removed the names of the groups from the summary unstructured data, then we can run bag of words on that, because originally when we ran bag of words we were getting suspiciously high accuracy, and it was discovered that the title of group name was usually included, for example, when a group claimed responsibility, (this is also most likely the case because we filtered doubtterr = 0.)