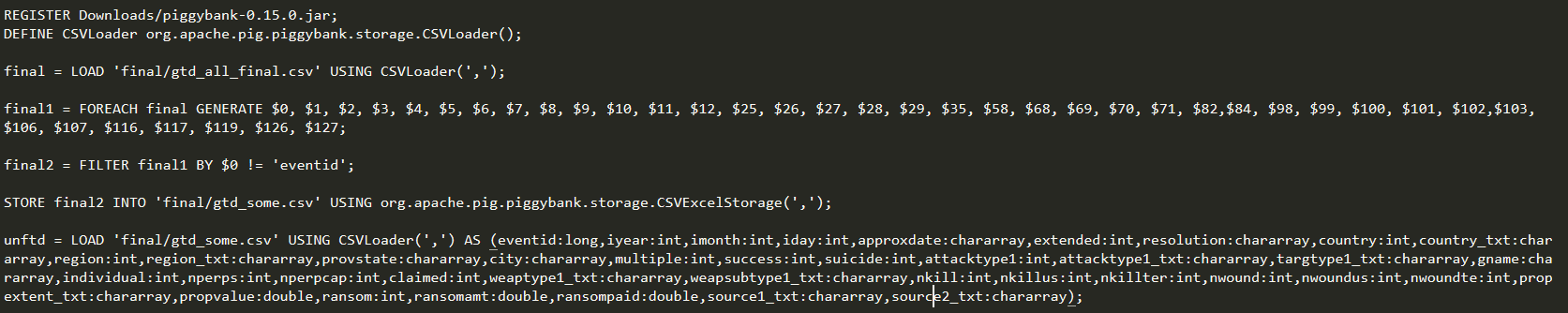
For my Final assignment I used the following versions:

Scala: 2.11.11

JRE System Library: jdk1.8.0\_161

To start with the data was moved from local file system to the Hadoop file system and the schema was defined for the same. Please find the below screenshot:



As per the question, I filtered my data and kept data post year 2002:



In my program I have taken following variables as features:

* Year
* Month
* Country
* Success
* Attack type
* Scite1 (Source1)
* Scite2 (Source2)

My target variable is ‘nkill’ which is the number of people killed. But for my assignment, it was converted into Boolean variable ‘nkill’ = 0 which means there was no casualty ‘1’ otherwise. The transformation was done in “Pig Latin” the code is as follows:



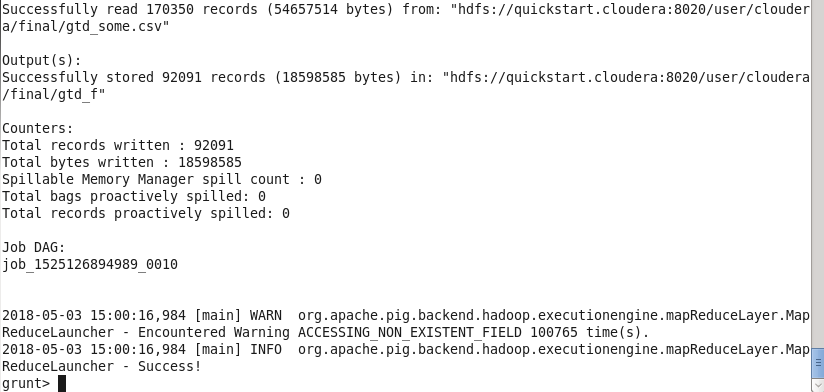
Before doing recoding I decided to take out Null value as I didn’t want to recode null value as it will make my model weak. The filtering was done in Pig Latin. Please refer to the screenshot below:



After doing the preprocessing and manipulation of data I saved it in a folder and named it “gtd\_f”. I will use the path of this folder in my scala code so that it can read data from it:



Output after running the above code



To ingest data into spark framework I have made special case class “TerrorCase” class. Since I am using columns Scite1 and Scite2 as features and since they have comments in them about the source I used “LemmaTuple” to lemmatise the string using “StanfordCoreNLP” which will also split, tokenize each words for that cell as in my model I will pass the array of IDF and Word2Vec to my logistic regression model and to calculate IDF and W2V I need count vectorizer which passes the count of the words present in the vector for every cell.

After the ingesting data we transform the categorical data using “OneHotEncoder” which turns them in binary as result making them continuous variable which can then be fed into my model. Before encoding them, we convert the categorical string value into categorical numerical value using “StringIndexer”

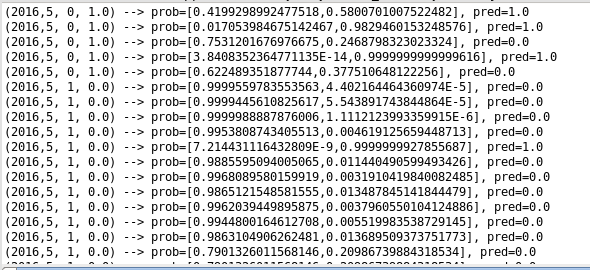
After this we calculate IDF and W2V which will be fed into our model.

Next step is that of defining the set of “features” which will be used to predict the value “nkill”

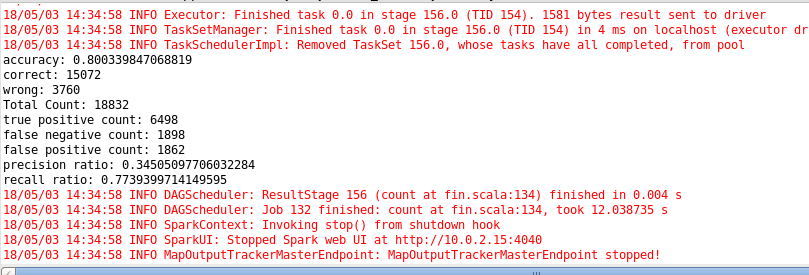
After that we define the “Pipeline” which basically tells the order of execution of different tasks.

After that we split our data into training and test for fitting and testing purpose. Following that we call the inbuilt Logistic Regression function after that we calculate and print various matrices as per the question. Please refer to the screenshot below:

Output of my model:



Different matrices which define the robustness of my model



Below the print statement we have defined the core NLP functions.