THE UNIVERSITY OF TEXAS AT DALLAS

Real time traffic accident analysis based on tweets

Name:

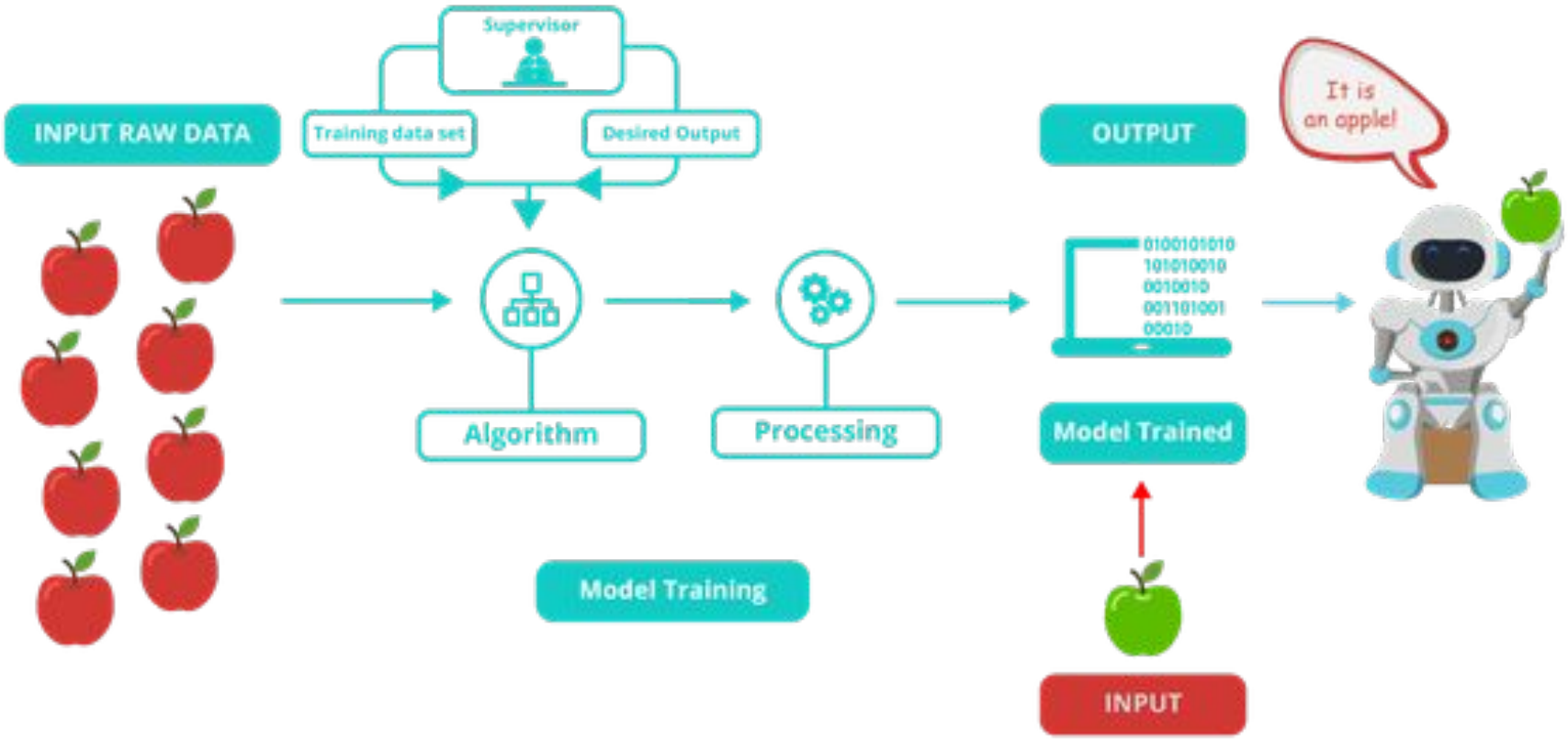
Introduction

- We used tweepy library to query real time tweet with given parameter (location/ hashtag)
- Tweepy is a library monitoring for tweets. We made a custom StreamListener object, which monitors tweets in real time and catches them and publish the these tweet to kafka in real time. With different given parameter such as location and hashtag. For each given location, we can filter the query result with their corresponding geo location.

Introduction

- We trained and built the the offline model with Random Forest on the given training data. The model will predict four different accident labels. No, Fire, Shooting and Crash.
- The spark streaming process used the pipelined model which consists of StringIndexer, RegexTokenizer, StopWordsRemover, HashingTF, IDF and RandomForestClassifier processes to make prediction on incoming tweets from kafka.
- Then the real time incoming tweet are predicted with this trained pipelined model, and the prediction and location are saved at the elasticsearch, which later on be visualized via kibana.

Offline Training model



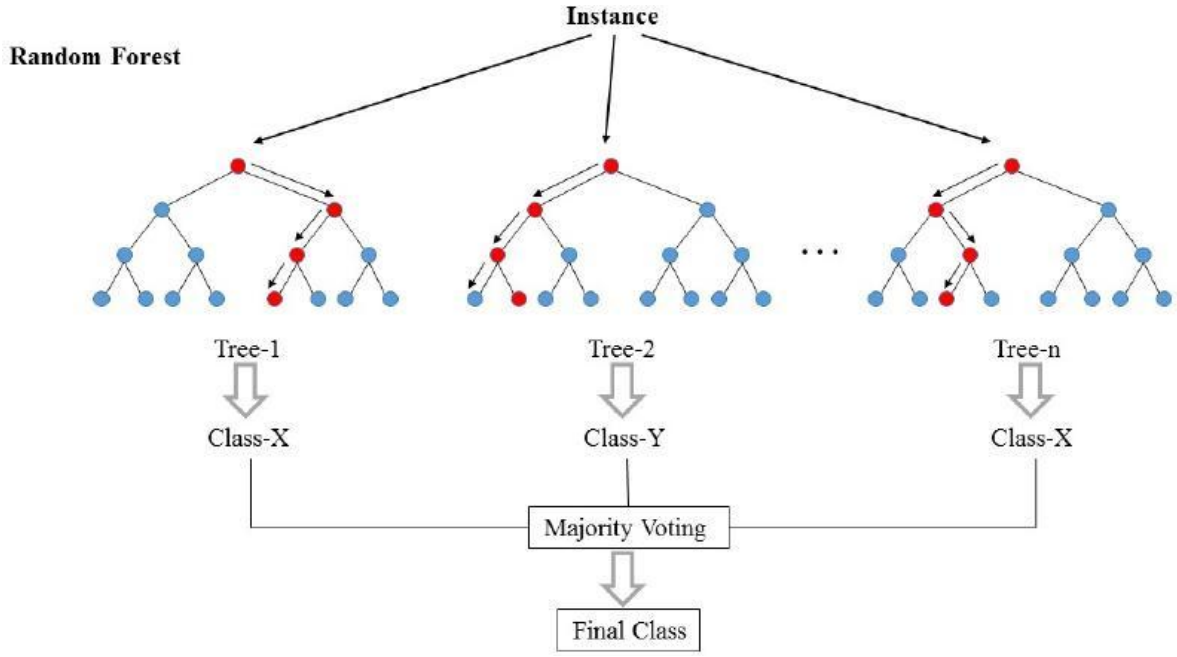
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.i2tutorials.com%2Ftop-machine-learning-interview-questions-and-answers%2Fwhat-are-the-differences-between-supervised-machine-learning-and-unsupervised-machine-learning%2F&psig=AOvVaw0KlqTp--NpSnIaOLP1ykZ6&ust=1575837937456000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLjwwYO0pOYCFQAAAAAdAAAAABAj

Random forest

In short, the random forest is several Decision Trees combined.

The samples are drawn with replacement, known as bootstrapping, which means that some samples will be used multiple times in a single tree.

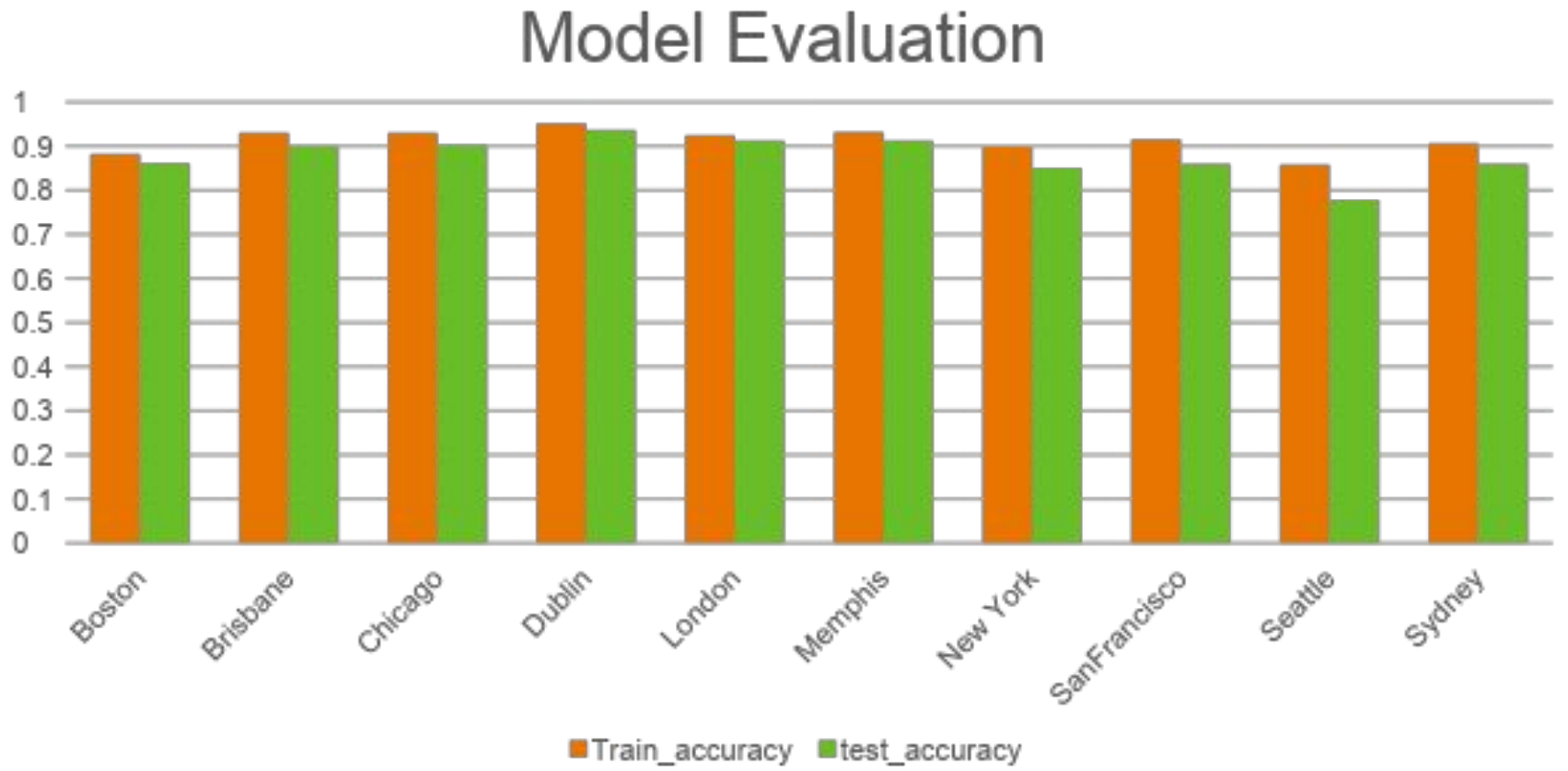
Although each decision tree would have high variances, the overall forest low variance and at the same time keep low bias.



https://i.stack.imgur.com/iY55n.jpg

Preprocess

	StringIndexer	RegexTokenizer	StopWordsRemover	HashingTF	Inverse Document Frequency (IDF)
Definition	Translate all the String classification to numeric	extracts tokens by matching the regex (Set Gap= False)	drops all the stop words from the input sequences	takes sets of terms and converts those sets into fixed-length feature vectors	rescale the feature vectors
Example	No:0,Crash:1,Fire:2,Shooting:3				
Input	Crash	This is a text block where you can tab across and add copy	[take, route, 213, to, avoid, traffic, get, stuck, in, traffic, anyway, solution, drive, off, bridge, now, no, more, traffic, ever, again]	[take, route, 213, avoid, traffic, get, stuck, traffic, anyway, solution, drive, bridge, traffic, ever]	(200,[4,23,55,80,103,109,124,148,159,177,181,194],[3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0])
Output	1	[take, route, 213, to, avoid, traffic, get, stuck, in, traffic, anyway, solution, drive, off, bridge, now, no, more, traffic, ever, again]	[take, route, 213, avoid, traffic, get, stuck, traffic, anyway, solution, drive, bridge, traffic, ever]	(200,[4,23,55,80,103,109,124,148,159,177,181,194],[3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0])	(200,[4,23,55,80,103,109,124,148,159,177,181,194],[7.1,3.2,2.7,3.1,2.7,3.3,3.2,2.6,2.5,3.2,3.2,3.0])



We split the dataset to 70% for training and 30% For testing and come out with accuracy for evaluation.

Streaming real-time model

	prediction	Predictlabel
his by accident. This is YOUR confirmation. YOU ARE GOING TO MAKE IT, no matter what it looks like rig…	0.0	NO
fferent people in our lives at different times. It's not an accident. It's not a coincident. There's a r…	0.0	NO
hicles involving known injury - ALEXANDER ST/E MAIN ST, Rochester #roc	1.0	crash
some sights/bars.	2.0	fire

For each RDD,

We predict the result of classification of the Tweeter

How producer run

Producer

```
ster ● python3 producer.py London accident
Coordinates:
[-0.213503, 51.512805, -0.105303, 51.572068]
"1;We fought non stop for 3.5 years - it took over my entire being. Right now th
e bulldozers are in. There doesn\u2019t seem to be anyone that fights the corner
 of these facilities #swimhour  Must say though that has always been really help
ful.;crash"
Message published successfully.
"2;Robert Johnson sold his soul to The Devil in exchange for Guitar Prowess and
died at 27. Every 27 Club Suicide and \u201cAccident\u201d or Murder is a refere
nce to this. They all sold their Soul for Success too, that\u2019s the Legend.;c
rash"
Message published successfully.
"3;After being in a car accident this year that fractured my neck I\u2019m bless
ed to be able to go back on tour!   I thank God tha\u2026;crash"
Message published successfully.
"4;STOP PUTTING YOUR CHILDREN IN THEIR CARSEATS WITH WINTER COATS ON \u2757\ufe0
f\u2757\ufe0f\u2757\ufe0f\u2757\ufe0f YOUD BE SAVING THEIR LIVES IF YOU JUST TOO
K IT OFF\u2026;crash"
Message published successfully.
"5;when good things happen: what a nice accident!  when bad things happen: it\u2
019s all my fault;crash"
Message published successfully.
```

Call tweepy API to get the coordinate of city(southwestern and Northeastern)

Remove all the RT @ , @, and url which are not relevant.

Pass in two parameters : city and hash_tag

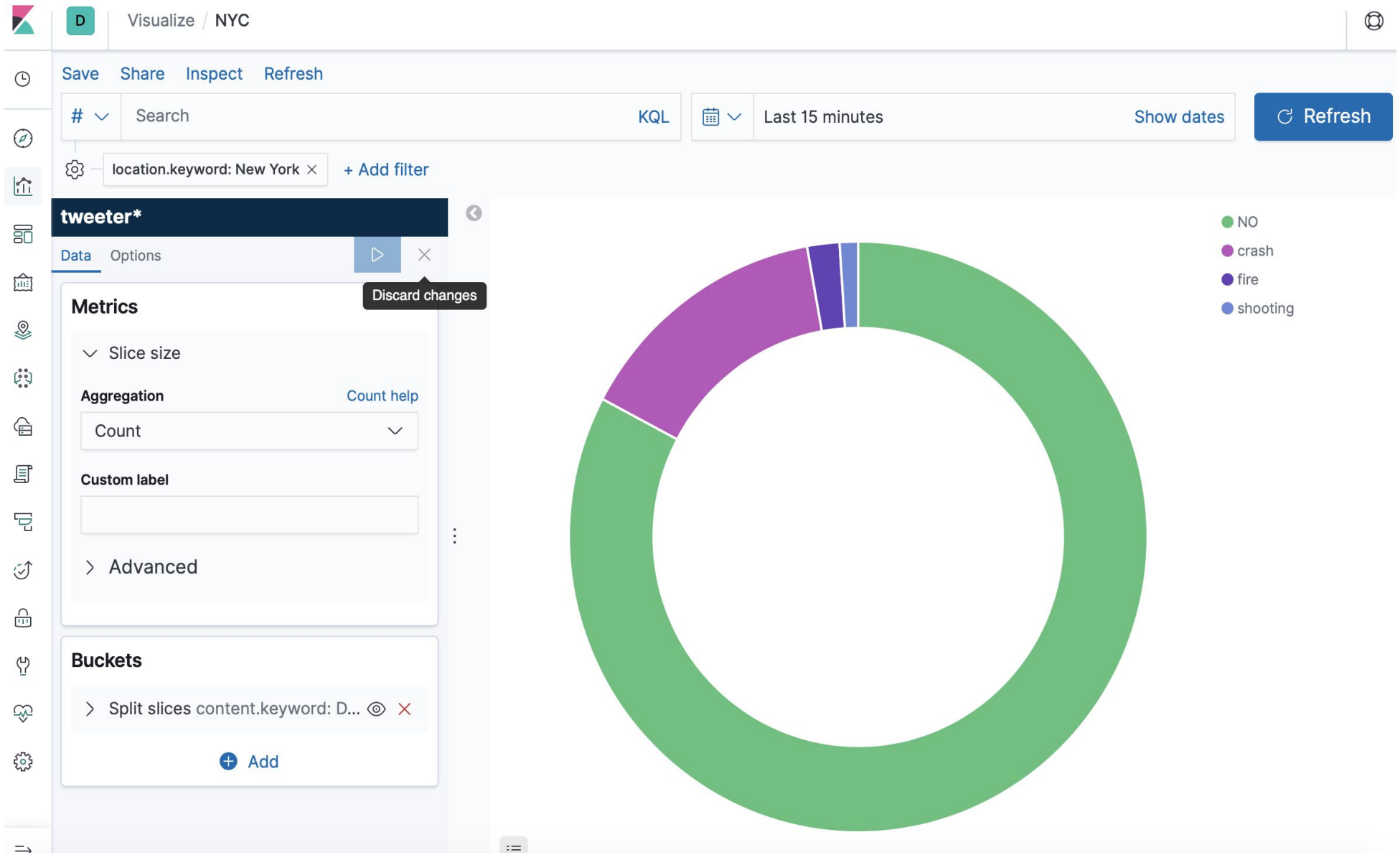
Visualization

The predicted label generated by the ml model and location of the tweets are stored at Elasticsearch, which is a full-text search and analytics engine.

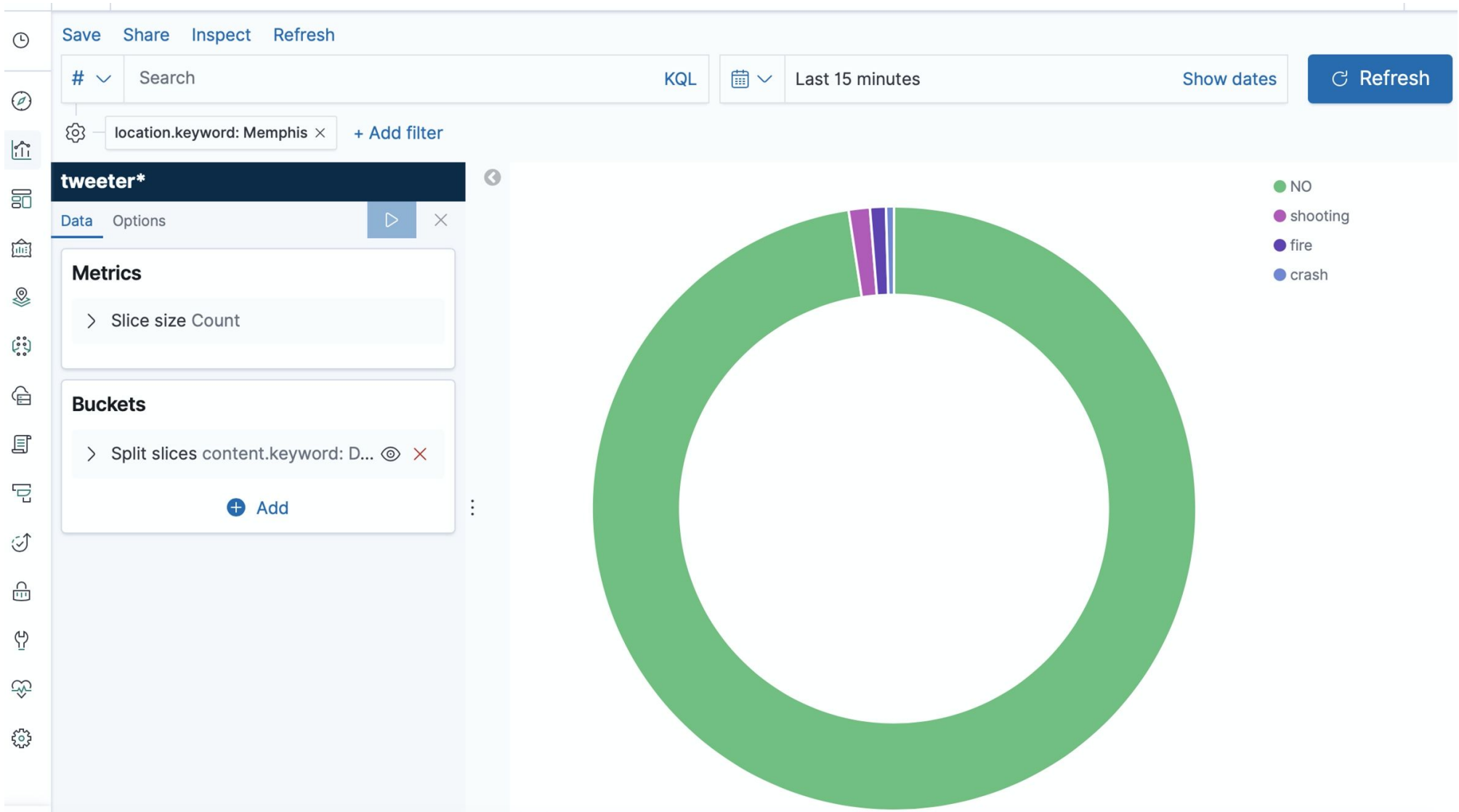
One advantage of using elasticsearch is that it can seamlessly work with Kibana, is an open source data visualization dashboard for Elasticsearch

We created elasticsearch index first, index the data. Then on Kibana, we created corresponding index patterns, then filtered out the tweets for each city and create corresponding pie chart to show the distribution of the different cases for each city.

Sample Pie chart for each city



Sample Pie chart for each city



Thank you