Tweet Sentiment Extraction

Jia Huang

Xi'an Shiyou University

Chinese Academy of Sciences

January 15, 2021

Overview

Project Overview

Data Pre-Processing

Feature Analysis

Feature Selection

Project Overview

- Project Background And Purpose

Data Pre-Processing

- Date Processing
- Feature Item
- Features Item
- Features Item

Feature Analysis

- Feature Analysis
- Dates & Day of the week
- Category & Police District
- X & Y

Feature Selection

- Feature Engineering

Project Overview

Project Background And Purpose

Data Pre-Processing

Feature Analysis

Feature Selection

Project Overview

Project Background And Purpose

Project Overview
Project Background And Purpose
Data Pre-Processing
Feature Analysis
Feature Selection

Defn

- Background

  With all of the tweets circulating every second it is hard to tell whether the sentiment behind a specific tweet will impact a company, or a person's, brand for being viral (positive), or devastate profit because it strikes a negative tone. Capturing sentiment in language is important in these times where decisions and reactions are created and updated in seconds. But, which words actually lead to the sentiment description?

- Purpose

  In this competition we've extracted support phrases from Figure Eight's Data for Everyone platform. The dataset is titled Sentiment Analysis: Emotion in Text tweets with existing sentiment labels, used here under creative commons attribution 4.0. international licence. Your objective in this competition is to construct a model that can do the same - look at the labeled sentiment for a given tweet and figure out what word or phrase best supports it.

Project Overview

Data Pre-Processing

Date Processing

Feature Item

Features Item

Features Item

Feature Analysis

Feature Selection

Data Pre-Processing

Date Processing

Project Overview

Data Pre-Processing

Date Processing

Feature Item

Features Item

Features Item

Feature Analysis

Feature Selection

Two data sets were given during the match: train.csv and test.csv.Train.csv data were used to construct the model and to predict test.csv data. Now let's take a look at the training data provided by this competition, do some exploratory data analysis work, and learn more about the content of this training set.
The data sets train.csv and test.csv are respectively 27408 and 3534 pieces of data.The data is described in terms of four attributes.

	textID	text	selected_text	sentiment
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative
2	088c60f138	my boss is bullying me...	bullying me	negative
3	9642c003ef	what interview! leave me alone	leave me alone	negative
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative

Figure 1

Feature Item

Project Overview

Data Pre-Processing

Date Processing

Feature Item

Features Item

Features Item

Feature Analysis

Feature Selection

As you can see from the table above, the four feature items in the training set are textID, Text, and selected_text sentiment.

- textID
  - Write the ID of the comment.
- Text
  - The specific content of the comment.
- Selected_text
  - Are the keywords that we have chosen to judge the emotional state of the comment.
- Sentiment
  - The emotional polarity of the sentence.

Features Item

Project Overview

Data Pre-Processing

Date Processing

Feature Item

Features Item

Features Item

Feature Analysis

Feature Selection

Out[21]:

	sentiment	text
1	neutral	11117
2	positive	8582
0	negative	7781

There are three types of sentence emotional polarity.Neurtal indicates that the sentence is emotionally neutral, positive means the comment is positive, and negative means the comment is negative.

Features Item

Project Overview

Data Pre-Processing

Date Processing

Feature Item

Features Item

Features Item

Feature Analysis

Feature Selection

- Address - the approximate street address of the crime incident
- X - Longitude
- Y - Latitude

Project Overview

Data Pre-Processing

Feature Analysis

Feature Analysis

Dates & Day of the week

Category & Police District

X & Y

Feature Selection

Constructing Dataset Generator

L

Project Overview

Data Pre-Processing

Feature Analysis

Feature Analysis

Dates & Day of the week

Category & Police District

X & Y

Feature Selection

ook at the Jaccard similarity between Text and selected_text.

	textID	text	selected_text	sentiment	jaccard_score
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral	1.000000
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative	0.200000
2	088c60f138	my boss is bullying me...	bullying me	negative	0.166667
3	9642c003ef	what interview! leave me alone	leave me alone	negative	0.600000
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative	0.214286

Feature Analysis

Project Overview

Data Pre-Processing

Feature Analysis

Feature Analysis

Dates & Day of the week

Category & Police District

X & Y

Feature Selection

- The data set contains nine eigenvalues, the data types are as follows, and we can see that the data set contains' object 'variables (also known as strings) that we need to encode.

```
Out[7]: Dates         datetime64[ns]
        Category              object
        Descript              object
        DayOfWeek             object
        PdDistrict            object
        Resolution            object
        Address               object
        X                    float64
        Y                    float64
        dtype: object

The dataset contains a lot of 'object' variables (aka strings) that we will need to encode.
```

Figure 2: Data Type

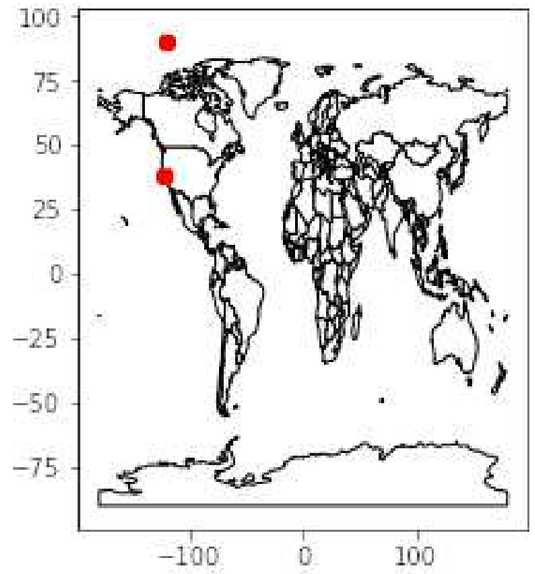

Figure 3: The Total Number of Copies

The 2,323 copies are duplicates and we need to delete them. We will also evaluate the position of the data points using the coordinates.

Feature Analysis

Project Overview

Data Pre-Processing

Feature Analysis

Feature Analysis

Dates & Day of the week

Category & Police District

X & Y

Feature Selection

There are also some wrong locations in the data set. After analysis, we found a total of 67 wrong messages, which also means that we cannot use these 67 wrong messages.

67

Out[8]:

	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y	Coordinates
673114	2005-10-23 18:11:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Sunday	TARAVAL	ARREST, BOOKED	STCHARLES AV / 19TH AV	-120.5	90.0	POINT (-120.50000 90.00000)
688950	2005-08-09 23:15:00	OTHER OFFENSES	DRIVERS LICENSE, SUSPENDED OR REVOKED	Tuesday	TARAVAL	ARREST, CITED	GENEVA AV / INTERSTATE280 HY	-120.5	90.0	POINT (-120.50000 90.00000)
823378	2003-09-21 13:00:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Sunday	BAYVIEW	NONE	GILMAN AV / FITCH ST	-120.5	90.0	POINT (-120.50000 90.00000)
661106	2005-12-29 00:07:00	NON-CRIMINAL	AIDED CASE, MENTAL DISTURBED	Thursday	TENDERLOIN	PSYCHOPATHIC CASE	5THSTNORTH ST / EDDY ST	-120.5	90.0	POINT (-120.50000 90.00000)
679643	2005-09-23 23:00:00	BURGLARY	BURGLARY, ATTEMPTED FORCIBLE ENTRY	Friday	BAYVIEW	NONE	3RD ST / ISLAISCREEK ST	-120.5	90.0	POINT (-120.50000 90.00000)

Dates & Day of the week

Project Overview

Data Pre-Processing

Feature Analysis

Feature Analysis

Dates & Day of the week

Category & Police District

X & Y

Feature Selection

These variables are distributed uniformly between 1/1/2003 to 5/13/2015 (and Monday to Sunday) and split between the training and the testing dataset as mentioned before. We did not notice any anomalies on these variables. The median frequency of incidents is 389 per day with a standard deviation of 48.51. Also, there is no significant deviation of incidents frequency throughout the week. Thus we do not expect this variable to play a significant role in the prediction.

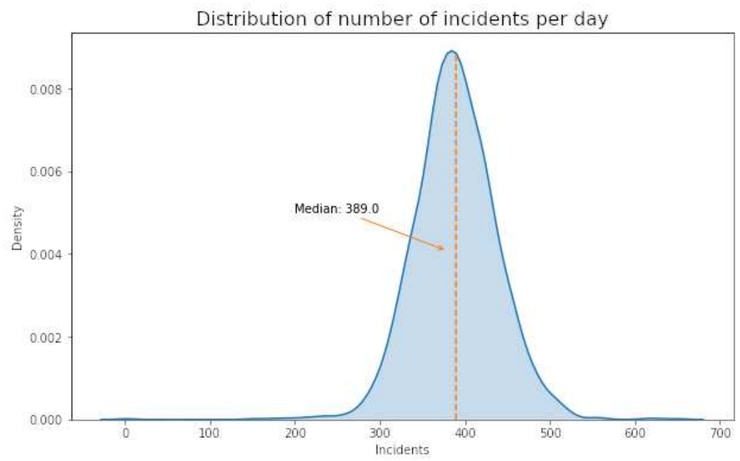

Figure 4: Dates and Day of the Week

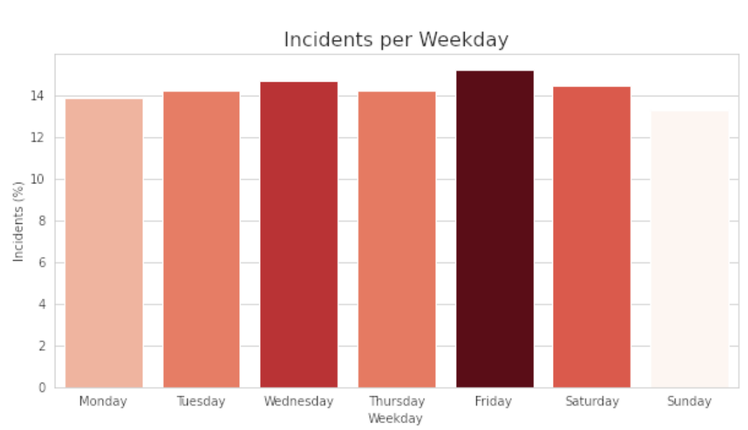

Figure 5: Per Weekday

Category & Police District

Project Overview

Data Pre-Processing

Feature Analysis

Feature Analysis

Dates & Day of the week

Category & Police District

X & Y

Feature Selection

There are 39 discrete categories that the police department file the incidents with the most common being Larceny/Theft (19.91%), Non/Criminal (10.50%), and Assault(8.77%). There are significant differences between the different districts of the City with the Southern district having the most incidents (17.87%) followed by Mission (13.67%) and Northern (12.00%).

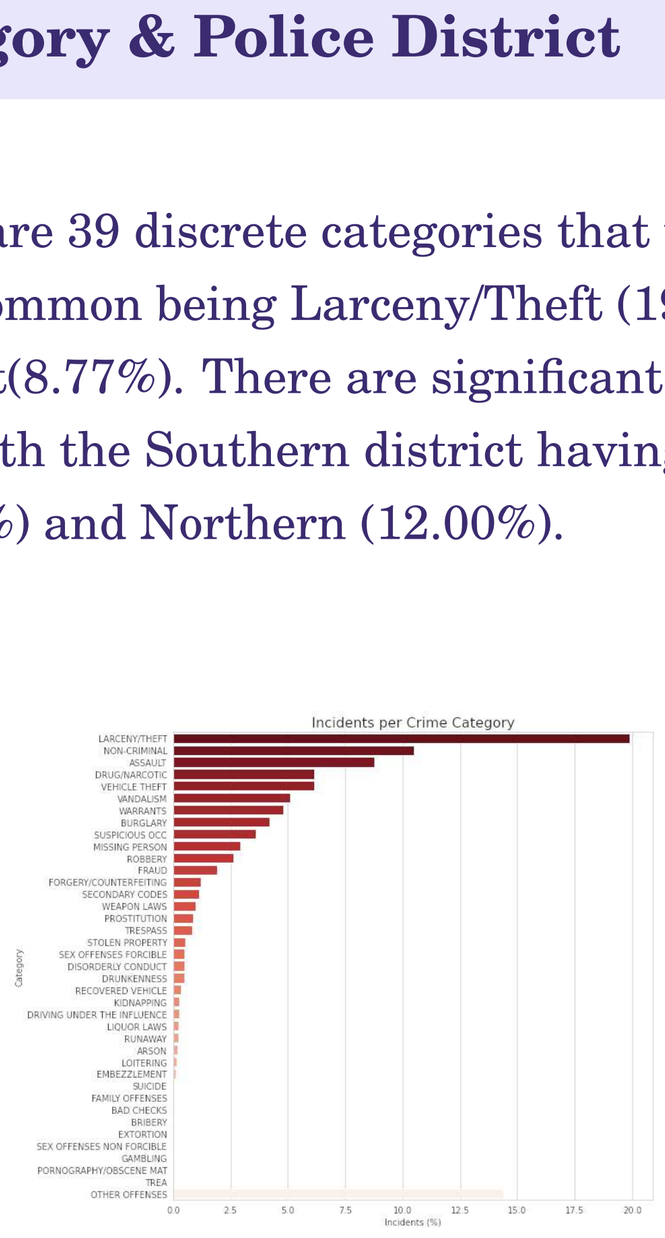

Figure 6: Category

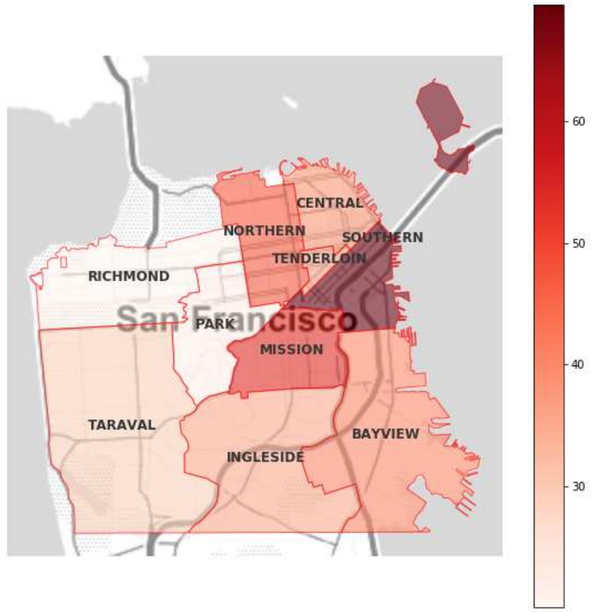

Figure 7: Police District

X & Y

Project Overview

Data Pre-Processing

Feature Analysis

Feature Analysis

Dates & Day of the week

Category & Police District

X & Y

Feature Selection

- Address Address, as a text field, requires advanced techniques to use it for the prediction. Instead in this project, we will use it to extract if the incident has happened on the road or in a building block.
- X - Longitude Y - Latitude We have tested that the coordinates belong inside the boundaries of the city. Although longitude does not contain any outliers, latitude includes some 90o values which correspond to the North Pole.

Project Overview

Data Pre-Processing

Feature Analysis

Feature Selection

Feature Engineering

Feature Selection

Feature Engineering

Project Overview

Data Pre-Processing

Feature Analysis

Feature Selection

Feature Engineering

Then, we created additional features. More specifically:

- From the 'Dates' field, we extracted the Day, the Month, the Year, the Hour, the Minute, the Weekday, and the number of days since the first day in the data.
- From the 'Address' field we extracted if the incident has taken place in a crossroad or on a building block.

Out[19]:

Weight	Feature
0.0579 ± 0.0012	Minute
0.0470 ± 0.0008	Y
0.0355 ± 0.0008	X
0.0179 ± 0.0002	Block
0.0176 ± 0.0008	n_days
0.0138 ± 0.0009	Hour
0.0129 ± 0.0007	PdDistrict
0.0108 ± 0.0004	Year
0.0028 ± 0.0002	Month
0.0017 ± 0.0004	Day
0.0014 ± 0.0003	DayOfWeek