submitted

January 25, 2022

1 Load train dataset

1. Using google files from colab, upload the training file, train.csv

[1]:

```
from google.colab import files
upload = files.upload()
```

```
<IPython.core.display.HTML object>


Saving train.csv to train (1).csv
```

2 Converting dataset from csv to pandas dataframe

1. Convert the uplaoded dataset into a pandas dataframe.

The variable df looks like

id	keyword	location	text	target
1	NaN	City A	That is a really bad cracker	0
4	Flood	City B	Floods after the rain RT #Flood @WaterBurst	1
5	Collapse	NaN	Big collapse in the half constructed building https://build	1
6	Fire	City C	Team is really fired up!	0
7	NaN	City D	Rocked	0

[2]:

```
import io
import pandas as pd
df = pd.read_csv(io.BytesIO(upload['train.csv']))
```

3 Clean Dataset

1. Replace NaN values to empty string, ' '.
2. Sort dataset according to target values

id	keyword	location	text	target
1		City A	That is a really bad cracker	0
6	Fire	City C	Team is really fired up!	0
7		City D	Rocked	0
4	Flood	City B	Floods after the rain RT #Flood @WaterBurst	1
5	Collapse		Big collapse in the half constructed building https://build	1

[3]:

```python
import numpy as np
df = df.sort_values('target')
df = df.fillna('')
```

4 Getting relevant columns (INPUT)

1. Get the keyword, text and location column

keyword	location	text
City A	That is a really bad cracker	
Fire	City C	Team is really fired up!
City D	Rocked	
Flood	City B	Floods after the rain RT #Flood @WaterBurst
Collapse		Big collapse in the half constructed building https://build

2. Concate the strings element-wise

tweets
City A That is a really bad cracker
Fire City C Team is really fired up!
City D Rocked
Flood City B Floods after the rain RT #Flood @WaterBurst
Collapse Big collapse in the half constructed building https://build

[4]:

```python
keyword = df['keyword']
location = df['location']
text = df['text']
```

[5]:

```python
tweets = np.add(keyword + ' ', np.add(text, ' ' + location))
```

5 Preprocess tweets function

1. Remove numbers

2. Remove special characters and any following that
3. Remove retweets
4. Remove links
5. Remove hashtags
6. Remove at the rates
7. Remove case to all small, strip handles, reduce the length using tokenizer
8. Remove stopwords
9. Stem the words
10. Remove punctuations
11. Store the clean data

tweets
[‘city’, ‘a’, ‘bad’, ‘cracker’]
[‘fire’, ‘city’, ‘c’, ‘team’, ‘fire’, ‘up’]
[‘city’, ‘d’, ‘rock’]
[‘flood’, ‘city’, ‘b’, ‘flood’, ‘rain’, ‘flood’, ‘waterburst’]
[‘collapse’, ‘big’, ‘collapse’, ‘half’, ‘construct’, ‘build’]

[6]:
```
def process_tweet(tweet):
  tweet = re.sub(r"[0-9]", "", tweet)
  tweet = re.sub(r'\$\w*', '', tweet)
  tweet = re.sub(r'^RT[\s]+', '', tweet)
  tweet = re.sub(r'https?://[^\s\n\r]+', '', tweet)
  tweet = re.sub(r'http?://[^\s\n\r]+', '', tweet)
  tweet = re.sub(r'#', '', tweet)
  tweet = re.sub(r'@', '', tweet)
  tokenizer = TweetTokenizer(preserve_case=False, strip_handles=True,reduce_len=True)
  tweet_tokens = tokenizer.tokenize(tweet)
  tweet_clean = []
  stopwords_english = stopwords.words('english')
  stemmer = PorterStemmer()
  for word in tweet_tokens:
    if word not in stopwords_english and word not in string.punctuation:
      stem_word = stemmer.stem(word)
      tweet_clean.append(stem_word)
  return tweet_clean
```

6 Get labels (ACTUAL OUTPUT)

1. Get the count of 0s and 1s usinf the Counter function.
2. Generate a numpy array of 0s and 1s.

labels
0
0
0
1
1

[7]:

```
from collections import Counter
target = df['target']
target_count = Counter(target)
labels = np.append(np.zeros((target_count[0], 1)), np.ones((target_count[1],␣
 ↪1)), axis = 0)
```

7 Getting frequency of each word function

1. Squeeze the labels to a list
2. Generate a dictionary with a tulple as key, (word, target_value) and increase its count

```
{
  ('city', 0): 4,
  ('at', 0): 1,
  ('bad', 0): 1,
  ('cracker', 0): 1,
  ('fire', 0): 2,
  ('c', 0): 1,
  ('team', 0): 1,
  ('up', 0): 1,
  ('d', 0): 1,
  ('rock', 0): 1,
  ('flood', 1): 3,
  ('b', 1): 1,
  ('rain', 1): 1,
  ('waterburst', 1): 1,
  ('collapse', 1): 2,
  ('big', 1): 1,
  ('half', 1): 1,
  ('construct', 1): 1,
  ('build', 1): 1
}
```

[8]:

```
def build_freqs(tweets, ys):
    yslist = np.squeeze(ys).tolist()
    freqs = {}
    for y, tweet in zip(yslist, tweets):
        for word in process_tweet(tweet):
            pair = (word, y)
            if pair in freqs:
                freqs[pair] += 1
            else:
                freqs[pair] = 1
    return freqs
```

8 Using the above functions

[9]:

```
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import TweetTokenizer
import string
import nltk
nltk.download('stopwords')
freqs = build_freqs(tweets, labels)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

9 Sigmoid function

1. Sigmoid is

$$sig(x) = \frac{1}{1 + e^{-z}}$$

[10]:

```
def sigmoid(z):
  return 1 / (1 + np.exp(-1 * z))
```

10 Gradient descent on cost function

Dimensions of X are (5, 3).

Dimensions of are (3, 1).

$$z = X$$

Dimensions of z are (5, 1).

$$h = \frac{1}{1 + e^{-z}}$$

Dimensions of h are (5, 1).

$$J = \frac{-1}{m}(Y^T \cdot log(h)) \cdot ((1 - Y)^T \cdot log(1 - h))$$

Dimensions of Y are (5, 1)

$$\theta := \theta - \frac{\alpha}{m}(X^T \cdot (h - y))$$

```
[11]: def gradientDescent(x, y, theta, alpha, num_iters):
          m = x.shape[0]
          for i in range(0, num_iters):
              z = np.dot(x, theta)
              h = sigmoid(z)
              J = -1./m * (np.dot(y.transpose(), np.log(h)) + np.dot((1 - y).
       ↪transpose(), np.log(1 - h)))
              theta = theta - (alpha / m) * (np.dot(x.transpose(), (h - y)))
          J = float(J)
          return J, theta
```

11 Extract features

Extracting features is basically the number of words in fake and real tweets.

The first feature is the bias which is 1.

There are 14 words belonging to fake catagory

There are 12 words belonging to real catagory

The features are

$$X = [1, 12, 14]$$

```
[12]: def extract_features(tweet, freqs, process_tweet=process_tweet):
          word_l = process_tweet(tweet)
          x = np.zeros((1, 3))
          x[0,0] = 1
          for word in word_l:
```

```
        x[0,1] += freqs.get((word, 1.0), 0)
        x[0,2] += freqs.get((word, 0.0), 0)
    assert(x.shape == (1, 3))
    return x
```

12 Applying gradiant descent

[13]:
```
train_x = np.copy(tweets)
X = np.zeros((len(train_x), 3))
for i in range(len(train_x)):
    X[i, :]= extract_features(train_x[i], freqs)
Y = labels
J, theta = gradientDescent(X, Y, np.zeros((3, 1)), 1e-9, 1500)
print(f"The cost after training is {J:.8f}.")
print(f"The resulting vector of weights is {[t for t in np.squeeze(theta)]}")
```

```
The cost after training is 0.68755206.
The resulting vector of weights is [-1.053516501295299e-07,
6.399116253535195e-05, -6.55973910760322e-05]
```

13 Test data

Getting the weights and applying to the test data is to get the final value of the logistic regression.

$$\hat{y} = \begin{cases} 1 & y_{pred} \geq 0.5 \\ 0 & y_{pred} < 0.5 \end{cases} \tag{1}$$

Predicted value is after the weights value on the test data.

[14]:
```
from google.colab import files
upload = files.upload()
```

```
<IPython.core.display.HTML object>


Saving test.csv to test (1).csv
```

[15]:
```
import io
import pandas as pd
df = pd.read_csv(io.BytesIO(upload['test.csv']))
df = df.fillna('')
```

[16]:
```
id = df['id']
keyword = df['keyword']
```

```
location = df['location']
text = df['text']
```

[17]:

```
test_x = np.add(keyword + ' ', np.add(text, ' ' + location))
```

[18]:

```
def predict_tweet(tweet, freqs, theta):
    x = extract_features(tweet, freqs)
    y_pred = sigmoid(np.dot(x, theta))
    return y_pred
```

[19]:

```
def test_logistic_regression(test_x, freqs, theta, id,␣
 ↪predict_tweet=predict_tweet):
    y_hat = []
    for tweet in test_x:
        y_pred = predict_tweet(tweet, freqs, theta)
        if y_pred > 0.5:
            y_hat.append(1)
        else:
            y_hat.append(0)
    id = list(id)
    import csv
    with open('submit.csv', 'w') as f:
        writer = csv.writer(f)
        writer.writerows(zip(['id'], ['target']))
        writer.writerows(zip(id, y_hat))
```

[20]:

```
test_logistic_regression(test_x, freqs, theta, id)
```