

2_Quora_Preprocessing

February 28, 2020

0.0.1 1.2.1 : EDA: Advanced Feature Extraction.

```
In [0]: import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from subprocess import check_output
%matplotlib inline
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import os
import gc

import re
from nltk.corpus import stopwords
import distance
from nltk.stem import PorterStemmer
from bs4 import BeautifulSoup
import re
from nltk.corpus import stopwords
# This package is used for finding longest common subsequence between two strings
# you can write your own dp code for this
import distance
from nltk.stem import PorterStemmer
from bs4 import BeautifulSoup
from fuzzywuzzy import fuzz
from sklearn.manifold import TSNE
# Import the Required lib packages for WORD-Cloud generation
# https://stackoverflow.com/questions/45625434/how-to-install-wordcloud-in-python3-6
from wordcloud import WordCloud, STOPWORDS
from os import path
from PIL import Image

In [0]: #https://stackoverflow.com/questions/12468179/unicodedecodeerror-utf8-codec-cant-decod
```

```

if os.path.isfile('df_fe_without_preprocessing_train.csv'):
    df = pd.read_csv("df_fe_without_preprocessing_train.csv",encoding='latin-1')
    df = df.fillna('')
    df.head()
else:
    print("get df_fe_without_preprocessing_train.csv from drive or run the previous no

```

```
In [0]: df.head(2)
```

```

Out[0]:   id  qid1  qid2          question1 \
0    0    1    2  What is the step by step guide to invest in sh...
1    1    3    4  What is the story of Kohinoor (Koh-i-Noor) Dia...

          question2  is_duplicate  freq_qid1 \
0  What is the step by step guide to invest in sh...          0          1
1  What would happen if the Indian government sto...          0          4

    freq_qid2  q1len  q2len  q1_n_words  q2_n_words  word_Common  word_Total \
0           1     66     57           14           12          10.0          23.0
1           1     51     88            8           13           4.0          20.0

    word_share  freq_q1+q2  freq_q1-q2
0    0.434783           2            0
1    0.200000           5            3

```

3.4 Preprocessing of Text

- Preprocessing:

- Removing html tags
- Removing Punctuations
- Performing stemming
- Removing Stopwords
- Expanding contractions etc.

```
In [0]: # To get the results in 4 decemal points
```

```
SAFE_DIV = 0.0001
```

```
STOP_WORDS = stopwords.words("english")
```

```

def preprocess(x):
    x = str(x).lower()
    x = x.replace(",000,000", "m").replace(",000", "k").replace("", "").replace(" ", " ")
    x = x.replace("won't", "will not").replace("cannot", "can not").replace("n't", " not")
    x = x.replace("what's", "what is").replace("'ve", " have").replace("i'm", "i am").replace("re", "re")
    x = x.replace("he's", "he is").replace("she's", "she is").replace("%", " percent ").replace(" ", " rupee ").replace(" ", " ")

```

```

        .replace("", " euro ").replace("'ll", " will")
x = re.sub(r"([0-9]+)000000", r"\1m", x)
x = re.sub(r"([0-9]+)000", r"\1k", x)

porter = PorterStemmer()
pattern = re.compile('\W')

if type(x) == type(''):
    x = re.sub(pattern, ' ', x)

if type(x) == type(''):
    x = porter.stem(x)
    example1 = BeautifulSoup(x)
    x = example1.get_text()

return x

```

- Function to Compute and get the features : With 2 parameters of Question 1 and Question 2

3.5 Advanced Feature Extraction (NLP and Fuzzy Features)

Definition: - **Token**: You get a token by splitting sentence a space - **Stop_Word** : stop words as per NLTK. - **Word** : A token that is not a stop_word

Features: - **cwc_min** : Ratio of common_word_count to min length of word count of Q1 and Q2 $cwc_min = common_word_count / (\min(len(q1_words), len(q2_words)))$ - **cwc_max** : Ratio of common_word_count to max length of word count of Q1 and Q2 $cwc_max = common_word_count / (\max(len(q1_words), len(q2_words)))$ - **csc_min** : Ratio of common_stop_count to min length of stop count of Q1 and Q2 $csc_min = common_stop_count / (\min(len(q1_stops), len(q2_stops)))$ - **csc_max** : Ratio of common_stop_count to max length of stop count of Q1 and Q2 $csc_max = common_stop_count / (\max(len(q1_stops), len(q2_stops)))$ - **ctc_min** : Ratio of common_token_count to min length of token count of Q1 and Q2 $ctc_min = common_token_count / (\min(len(q1_tokens), len(q2_tokens)))$

- **ctc_max** : Ratio of common_token_count to max length of token count of Q1 and Q2 $ctc_max = common_token_count / (\max(len(q1_tokens), len(q2_tokens)))$
- **last_word_eq** : Check if First word of both questions is equal or not $last_word_eq = int(q1_tokens[-1] == q2_tokens[-1])$
- **first_word_eq** : Check if First word of both questions is equal or not $first_word_eq = int(q1_tokens[0] == q2_tokens[0])$
- **abs_len_diff** : Abs. length difference $abs_len_diff = abs(len(q1_tokens) - len(q2_tokens))$
- **mean_len** : Average Token Length of both Questions $mean_len = (len(q1_tokens) + len(q2_tokens))/2$

- **fuzz_ratio** : <https://github.com/seatgeek/fuzzywuzzy#usage>
<http://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/>
- **fuzz_partial_ratio** : <https://github.com/seatgeek/fuzzywuzzy#usage>
<http://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/>
- **token_sort_ratio** : <https://github.com/seatgeek/fuzzywuzzy#usage>
<http://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/>
- **token_set_ratio** : <https://github.com/seatgeek/fuzzywuzzy#usage>
<http://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/>
- **longest_substr_ratio** : Ratio of length longest common substring to min length of token count of Q1 and Q2
 $\text{longest_substr_ratio} = \text{len}(\text{longest common substring}) / (\min(\text{len}(q1_tokens), \text{len}(q2_tokens)))$

```
In [0]: def get_token_features(q1, q2):
    token_features = [0.0]*10

    # Converting the Sentence into Tokens:
    q1_tokens = q1.split()
    q2_tokens = q2.split()

    if len(q1_tokens) == 0 or len(q2_tokens) == 0:
        return token_features

    # Get the non-stopwords in Questions
    q1_words = set([word for word in q1_tokens if word not in STOP_WORDS])
    q2_words = set([word for word in q2_tokens if word not in STOP_WORDS])

    #Get the stopwords in Questions
    q1_stops = set([word for word in q1_tokens if word in STOP_WORDS])
    q2_stops = set([word for word in q2_tokens if word in STOP_WORDS])

    # Get the common non-stopwords from Question pair
    common_word_count = len(q1_words.intersection(q2_words))

    # Get the common stopwords from Question pair
    common_stop_count = len(q1_stops.intersection(q2_stops))

    # Get the common Tokens from Question pair
    common_token_count = len(set(q1_tokens).intersection(set(q2_tokens)))

    token_features[0] = common_word_count / (min(len(q1_words), len(q2_words)) + SAFE_I
    token_features[1] = common_word_count / (max(len(q1_words), len(q2_words)) + SAFE_I
    token_features[2] = common_stop_count / (min(len(q1_stops), len(q2_stops)) + SAFE_I
    token_features[3] = common_stop_count / (max(len(q1_stops), len(q2_stops)) + SAFE_I
    token_features[4] = common_token_count / (min(len(q1_tokens), len(q2_tokens)) + SAFE_I
    token_features[5] = common_token_count / (max(len(q1_tokens), len(q2_tokens)) + SAFE_I
```

```

# Last word of both question is same or not
token_features[6] = int(q1_tokens[-1] == q2_tokens[-1])

# First word of both question is same or not
token_features[7] = int(q1_tokens[0] == q2_tokens[0])

token_features[8] = abs(len(q1_tokens) - len(q2_tokens))

#Average Token Length of both Questions
token_features[9] = (len(q1_tokens) + len(q2_tokens))/2
return token_features

# get the Longest Common sub string

def get_longest_substr_ratio(a, b):
    strs = list(distance.lcs substrings(a, b))
    if len(strs) == 0:
        return 0
    else:
        return len(strs[0]) / (min(len(a), len(b)) + 1)

def extract_features(df):
    # preprocessing each question
    df["question1"] = df["question1"].fillna("").apply(preprocess)
    df["question2"] = df["question2"].fillna("").apply(preprocess)

    print("token features...")

    # Merging Features with dataset

    token_features = df.apply(lambda x: get_token_features(x["question1"], x["question2"]), axis=1)

    df["cwc_min"] = list(map(lambda x: x[0], token_features))
    df["cwc_max"] = list(map(lambda x: x[1], token_features))
    df["csc_min"] = list(map(lambda x: x[2], token_features))
    df["csc_max"] = list(map(lambda x: x[3], token_features))
    df["ctc_min"] = list(map(lambda x: x[4], token_features))
    df["ctc_max"] = list(map(lambda x: x[5], token_features))
    df["last_word_eq"] = list(map(lambda x: x[6], token_features))
    df["first_word_eq"] = list(map(lambda x: x[7], token_features))
    df["abs_len_diff"] = list(map(lambda x: x[8], token_features))
    df["mean_len"] = list(map(lambda x: x[9], token_features))

    #Computing Fuzzy Features and Merging with Dataset

    # do read this blog: http://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching/
    # https://stackoverflow.com/questions/31806695/when-to-use-which-fuzz-function-to-

```

```
# https://github.com/seatgeek/fuzzywuzzy
print("fuzzy features..")

df["token_set_ratio"] = df.apply(lambda x: fuzz.token_set_ratio(x["question1"], x["question2"]), axis=1)
# The token sort approach involves tokenizing the string in question, sorting the tokens,
# then joining them back into a string. We then compare the transformed strings with the other question.
df["token_sort_ratio"] = df.apply(lambda x: fuzz.token_sort_ratio(x["question1"], x["question2"]), axis=1)
df["fuzz_ratio"] = df.apply(lambda x: fuzz.QRatio(x["question1"], x["question2"]), axis=1)
df["fuzz_partial_ratio"] = df.apply(lambda x: fuzz.partial_ratio(x["question1"], x["question2"]), axis=1)
df["longest_substr_ratio"] = df.apply(lambda x: fuzz.longest_common_substring_ratio(x["question1"], x["question2"]), axis=1)
return df
```

```
In [0]: if os.path.isfile('nlp_features_train.csv'):
df = pd.read_csv("nlp_features_train.csv", encoding='latin-1')
df.fillna('')
else:
print("Extracting features for train:")
df = pd.read_csv("train.csv")
df = extract_features(df)
df.to_csv("nlp_features_train.csv", index=False)
df.head(2)
```

```
Out[0]:
```

	id	qid1	qid2	question1	question2	is_duplicate	cwc_min	cwc_max	csc_min	csc_max	ctc_max	last_word_eq	first_word_eq	abs_len_diff	mean_len	token_set_ratio	token_sort_ratio	fuzz_ratio	fuzz_partial_ratio	longest_substr_ratio
0	0	1	2	what is the step by step guide to invest in sh...	what is the step by step guide to invest in sh...	0	0.999980	0.833319	0.999983	0.999983	0.785709	0.0	1.0	2.0	13.0	100	93	93	100	0.982759
1	1	3	4	what is the story of kohinoor koh i noor dia...	what would happen if the indian government sto...	0	0.799984	0.399996	0.749981	0.599988	0.466664	0.0	1.0	5.0	12.5	86	63	66	75	0.596154

[2 rows x 21 columns]

3.5.1 Analysis of extracted features

3.5.1.1 Plotting Word clouds

- Creating Word Cloud of Duplicates and Non-Duplicates Question pairs
- We can observe the most frequent occurring words

```

In [0]: df_duplicate = df[df['is_duplicate'] == 1]
        dfp_nonduplicate = df[df['is_duplicate'] == 0]

        # Converting 2d array of q1 and q2 and flatten the array: like {{1,2},{3,4}} to {1,2,3,4}
        p = np.dstack([df_duplicate["question1"], df_duplicate["question2"]]).flatten()
        n = np.dstack([dfp_nonduplicate["question1"], dfp_nonduplicate["question2"]]).flatten()

        print ("Number of data points in class 1 (duplicate pairs) :",len(p))
        print ("Number of data points in class 0 (non duplicate pairs) :",len(n))

        #Saving the np array into a text file
        np.savetxt('train_p.txt', p, delimiter=' ', fmt='%s')
        np.savetxt('train_n.txt', n, delimiter=' ', fmt='%s')

```

Number of data points in class 1 (duplicate pairs) : 298526
Number of data points in class 0 (non duplicate pairs) : 510054

```

In [0]: # reading the text files and removing the Stop Words:
        d = path.dirname('.')

        textp_w = open(path.join(d, 'train_p.txt')).read()
        textn_w = open(path.join(d, 'train_n.txt')).read()
        stopwords = set(STOPWORDS)
        stopwords.add("said")
        stopwords.add("br")
        stopwords.add(" ")
        stopwords.remove("not")

        stopwords.remove("no")
        #stopwords.remove("good")
        #stopwords.remove("love")
        stopwords.remove("like")
        #stopwords.remove("best")
        #stopwords.remove("!")
        print ("Total number of words in duplicate pair questions :",len(textp_w))
        print ("Total number of words in non duplicate pair questions :",len(textn_w))

```

Total number of words in duplicate pair questions : 16109886
Total number of words in non duplicate pair questions : 33193130

__ Word Clouds generated from duplicate pair question's text __

```

In [0]: wc = WordCloud(background_color="white", max_words=len(textp_w), stopwords=stopwords)
        wc.generate(textp_w)
        print ("Word Cloud for Duplicate Question pairs")
        plt.imshow(wc, interpolation='bilinear')
        plt.axis("off")
        plt.show()

```

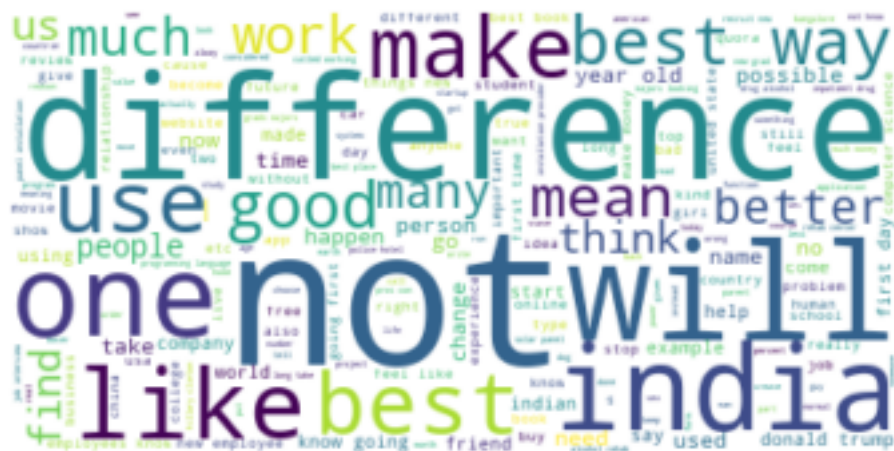
Word Cloud for Duplicate Question pairs



__ Word Clouds generated from non duplicate pair question's text __

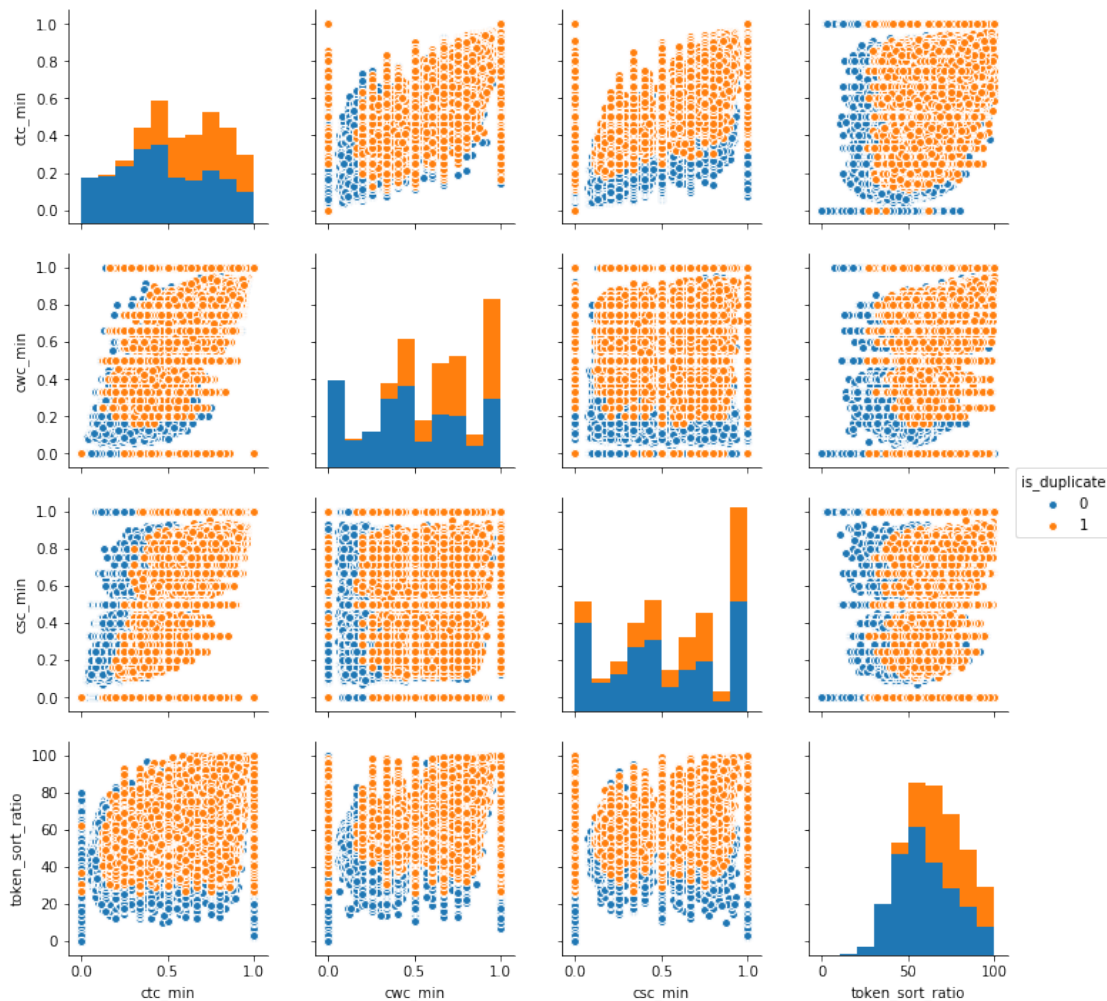
```
In [0]: wc = WordCloud(background_color="white", max_words=len(textn_w), stopwords=stopwords)
# generate word cloud
wc.generate(textn_w)
print ("Word Cloud for non-Duplicate Question pairs:")
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Word Cloud for non-Duplicate Question pairs:



3.5.1.2 Pair plot of features ['ctc_min', 'cwc_min', 'csc_min', 'token_sort_ratio']

```
In [0]: n = df.shape[0]
sns.pairplot(df[['ctc_min', 'cwc_min', 'csc_min', 'token_sort_ratio', 'is_duplicate']])
plt.show()
```

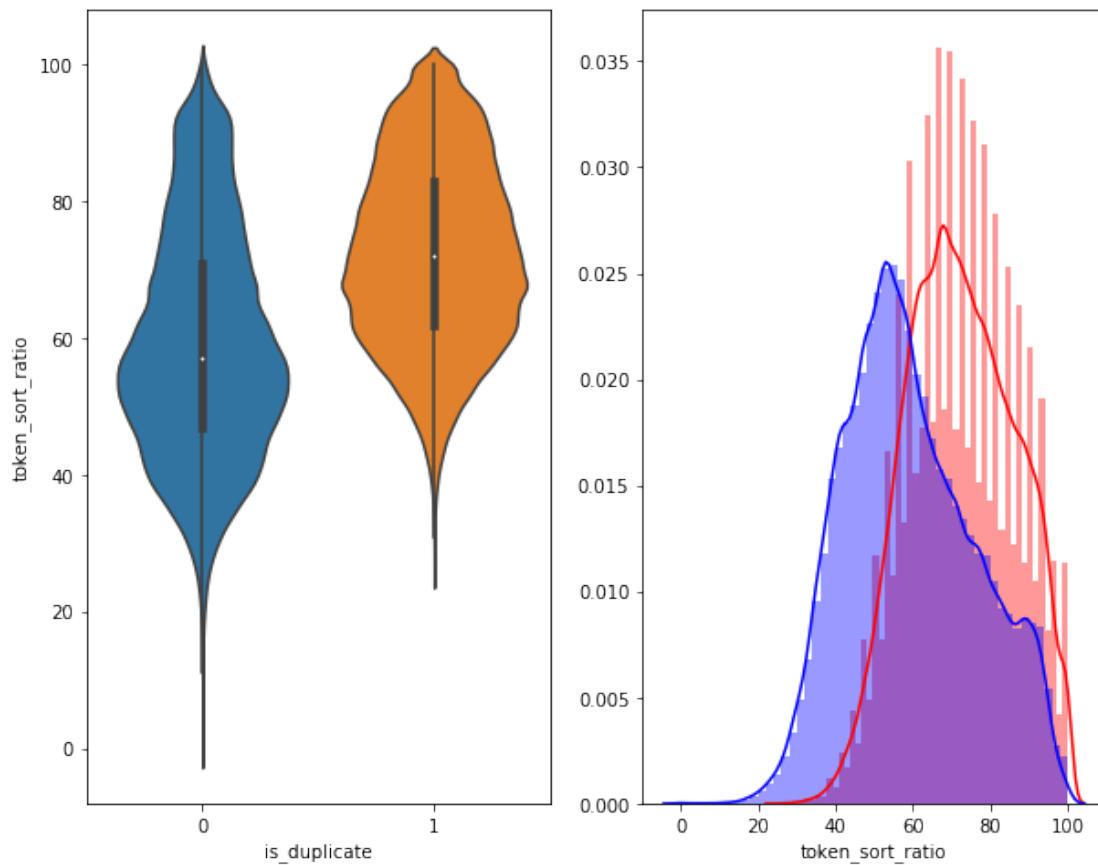


```
In [0]: # Distribution of the token_sort_ratio
plt.figure(figsize=(10, 8))

plt.subplot(1,2,1)
sns.violinplot(x = 'is_duplicate', y = 'token_sort_ratio', data = df[0:] , )

plt.subplot(1,2,2)
sns.distplot(df[df['is_duplicate'] == 1.0]['token_sort_ratio'][0:] , label = "1", color
```

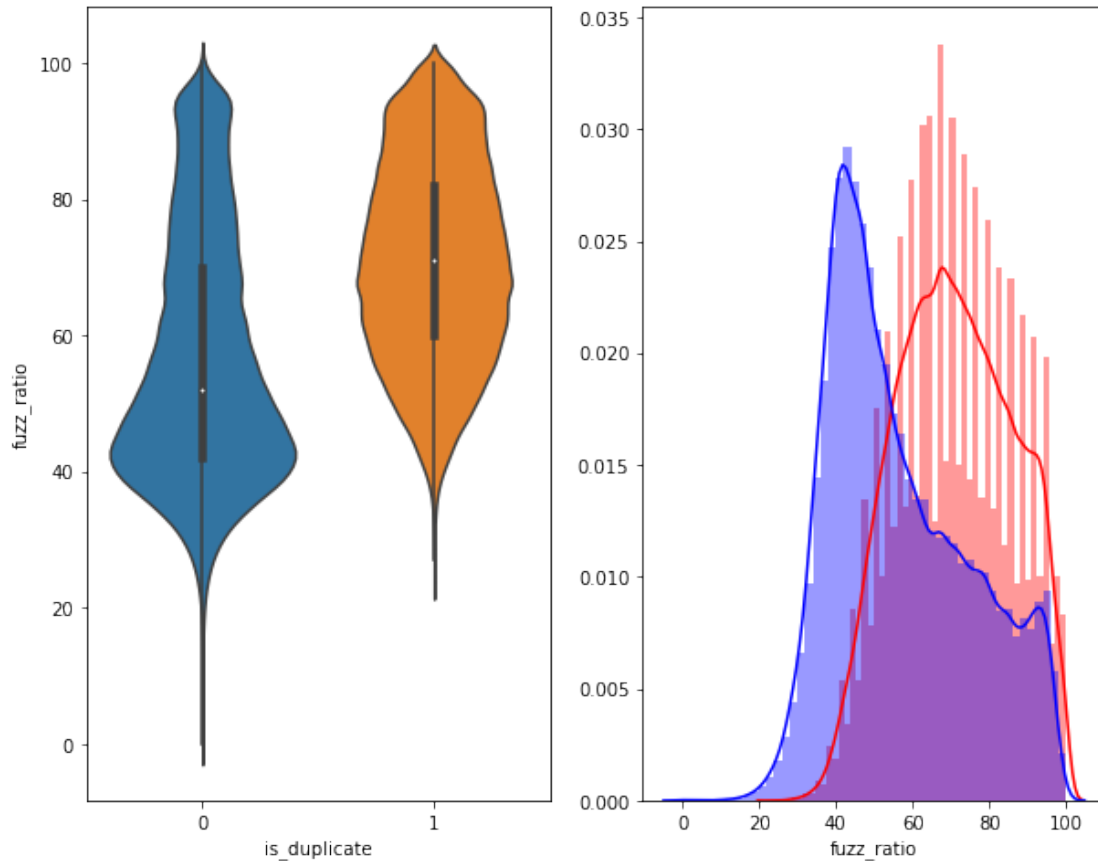
```
sns.distplot(df[df['is_duplicate'] == 0.0]['token_sort_ratio'][0:], label = "0" , color = 'blue')
plt.show()
```



```
In [0]: plt.figure(figsize=(10, 8))
```

```
plt.subplot(1,2,1)
sns.violinplot(x = 'is_duplicate', y = 'fuzz_ratio', data = df[0:] , )
```

```
plt.subplot(1,2,2)
sns.distplot(df[df['is_duplicate'] == 1.0]['fuzz_ratio'][0:] , label = "1", color = 'red')
sns.distplot(df[df['is_duplicate'] == 0.0]['fuzz_ratio'][0:] , label = "0" , color = 'blue')
plt.show()
```



3.5.2 Visualization

```
In [0]: # Using TSNE for Dimentionalitiy reduction for 15 Features(Generated after cleaning the
```

```
from sklearn.preprocessing import MinMaxScaler
```

```
dfp_subsampled = df[0:5000]
```

```
X = MinMaxScaler().fit_transform(dfp_subsampled[['cwc_min', 'cwc_max', 'csc_min', 'csc_max']])
```

```
y = dfp_subsampled['is_duplicate'].values
```

```
In [0]: tsne2d = TSNE(
    n_components=2,
    init='random', # pca
    random_state=101,
    method='barnes_hut',
    n_iter=1000,
    verbose=2,
    angle=0.5
).fit_transform(X)
```

```
[t-SNE] Computing 91 nearest neighbors...
```

```
[t-SNE] Indexed 5000 samples in 0.011s...
```

```

[t-SNE] Computed neighbors for 5000 samples in 0.912s...
[t-SNE] Computed conditional probabilities for sample 1000 / 5000
[t-SNE] Computed conditional probabilities for sample 2000 / 5000
[t-SNE] Computed conditional probabilities for sample 3000 / 5000
[t-SNE] Computed conditional probabilities for sample 4000 / 5000
[t-SNE] Computed conditional probabilities for sample 5000 / 5000
[t-SNE] Mean sigma: 0.116557
[t-SNE] Computed conditional probabilities in 0.433s
[t-SNE] Iteration 50: error = 80.9244080, gradient norm = 0.0428133 (50 iterations in 13.099s)
[t-SNE] Iteration 100: error = 70.3858795, gradient norm = 0.0100968 (50 iterations in 9.067s)
[t-SNE] Iteration 150: error = 68.6138382, gradient norm = 0.0058392 (50 iterations in 9.602s)
[t-SNE] Iteration 200: error = 67.7700119, gradient norm = 0.0036596 (50 iterations in 9.121s)
[t-SNE] Iteration 250: error = 67.2725067, gradient norm = 0.0034962 (50 iterations in 11.305s)
[t-SNE] KL divergence after 250 iterations with early exaggeration: 67.272507
[t-SNE] Iteration 300: error = 1.7737305, gradient norm = 0.0011918 (50 iterations in 8.289s)
[t-SNE] Iteration 350: error = 1.3720417, gradient norm = 0.0004822 (50 iterations in 10.526s)
[t-SNE] Iteration 400: error = 1.2039998, gradient norm = 0.0002768 (50 iterations in 9.600s)
[t-SNE] Iteration 450: error = 1.1133438, gradient norm = 0.0001881 (50 iterations in 11.827s)
[t-SNE] Iteration 500: error = 1.0579143, gradient norm = 0.0001434 (50 iterations in 8.941s)
[t-SNE] Iteration 550: error = 1.0221983, gradient norm = 0.0001164 (50 iterations in 11.092s)
[t-SNE] Iteration 600: error = 0.9987167, gradient norm = 0.0001039 (50 iterations in 11.467s)
[t-SNE] Iteration 650: error = 0.9831534, gradient norm = 0.0000938 (50 iterations in 11.799s)
[t-SNE] Iteration 700: error = 0.9722011, gradient norm = 0.0000858 (50 iterations in 12.028s)
[t-SNE] Iteration 750: error = 0.9643636, gradient norm = 0.0000799 (50 iterations in 12.120s)
[t-SNE] Iteration 800: error = 0.9584482, gradient norm = 0.0000785 (50 iterations in 11.867s)
[t-SNE] Iteration 850: error = 0.9538348, gradient norm = 0.0000739 (50 iterations in 11.461s)
[t-SNE] Iteration 900: error = 0.9496906, gradient norm = 0.0000712 (50 iterations in 11.023s)
[t-SNE] Iteration 950: error = 0.9463405, gradient norm = 0.0000673 (50 iterations in 11.755s)
[t-SNE] Iteration 1000: error = 0.9432716, gradient norm = 0.0000662 (50 iterations in 11.493s)
[t-SNE] Error after 1000 iterations: 0.943272

```

```

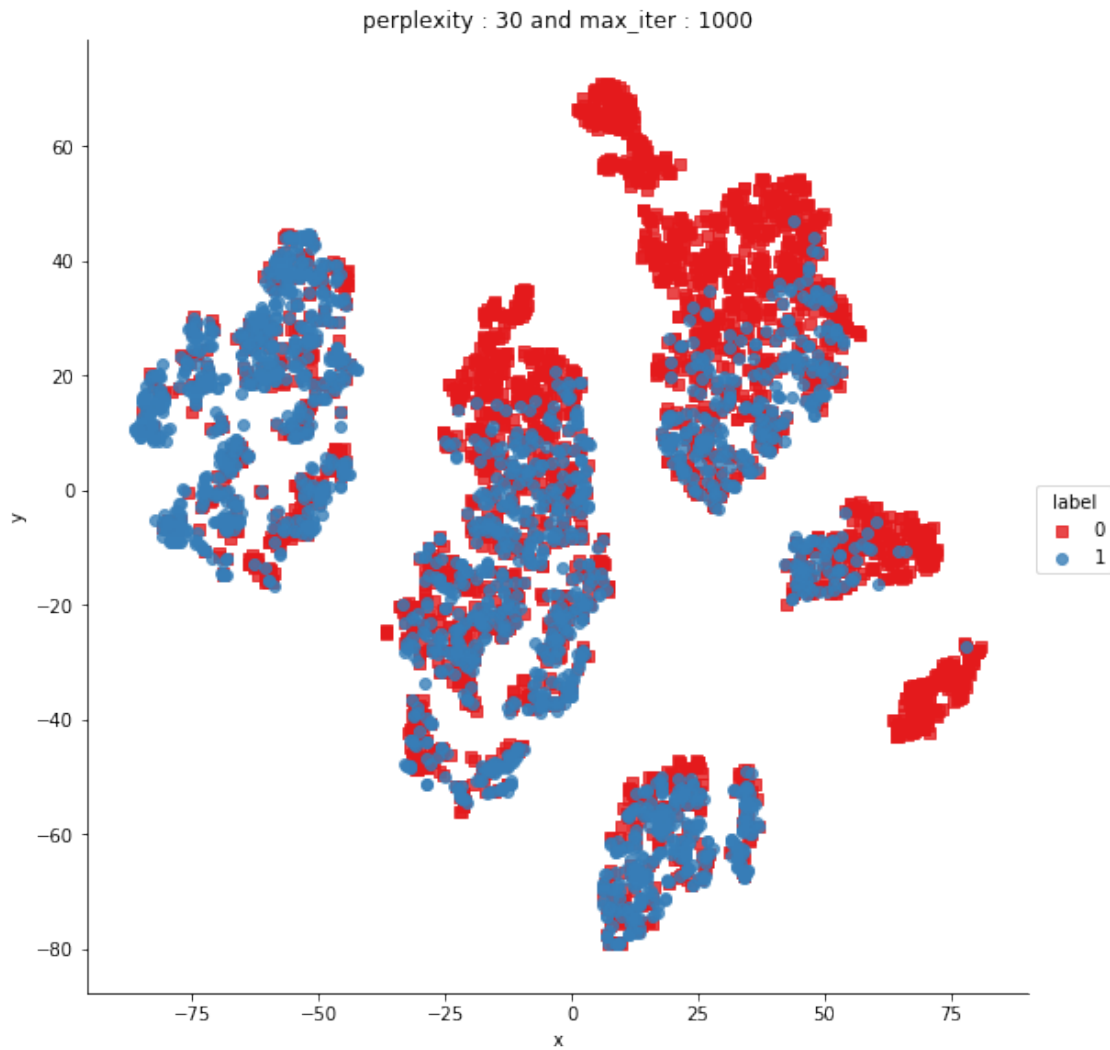
In [0]: df = pd.DataFrame({'x':tsne2d[:,0], 'y':tsne2d[:,1] , 'label':y})

```

```

# draw the plot in appropriate place in the grid
sns.lmplot(data=df, x='x', y='y', hue='label', fit_reg=False, size=8,palette="Set1",mar=0)
plt.title("perplexity : {} and max_iter : {}".format(30, 1000))
plt.show()

```



```
In [0]: from sklearn.manifold import TSNE
```

```
tsne3d = TSNE(
    n_components=3,
    init='random', # pca
    random_state=101,
    method='barnes_hut',
    n_iter=1000,
    verbose=2,
    angle=0.5
).fit_transform(X)
```

```
[t-SNE] Computing 91 nearest neighbors...
```

```
[t-SNE] Indexed 5000 samples in 0.010s...
```

```
[t-SNE] Computed neighbors for 5000 samples in 0.935s...
```

```
[t-SNE] Computed conditional probabilities for sample 1000 / 5000
```

```

[t-SNE] Computed conditional probabilities for sample 2000 / 5000
[t-SNE] Computed conditional probabilities for sample 3000 / 5000
[t-SNE] Computed conditional probabilities for sample 4000 / 5000
[t-SNE] Computed conditional probabilities for sample 5000 / 5000
[t-SNE] Mean sigma: 0.116557
[t-SNE] Computed conditional probabilities in 0.363s
[t-SNE] Iteration 50: error = 77.7944183, gradient norm = 0.1014017 (50 iterations in 34.931s)
[t-SNE] Iteration 100: error = 69.2682266, gradient norm = 0.0248657 (50 iterations in 15.147s)
[t-SNE] Iteration 150: error = 67.7877655, gradient norm = 0.0150941 (50 iterations in 13.761s)
[t-SNE] Iteration 200: error = 67.1991119, gradient norm = 0.0126559 (50 iterations in 13.425s)
[t-SNE] Iteration 250: error = 66.8560715, gradient norm = 0.0074975 (50 iterations in 12.904s)
[t-SNE] KL divergence after 250 iterations with early exaggeration: 66.856071
[t-SNE] Iteration 300: error = 1.2356015, gradient norm = 0.0007033 (50 iterations in 13.302s)
[t-SNE] Iteration 350: error = 0.9948602, gradient norm = 0.0001997 (50 iterations in 18.898s)
[t-SNE] Iteration 400: error = 0.9168936, gradient norm = 0.0001430 (50 iterations in 13.397s)
[t-SNE] Iteration 450: error = 0.8863022, gradient norm = 0.0000975 (50 iterations in 16.379s)
[t-SNE] Iteration 500: error = 0.8681002, gradient norm = 0.0000854 (50 iterations in 17.791s)
[t-SNE] Iteration 550: error = 0.8564141, gradient norm = 0.0000694 (50 iterations in 17.060s)
[t-SNE] Iteration 600: error = 0.8470711, gradient norm = 0.0000640 (50 iterations in 15.454s)
[t-SNE] Iteration 650: error = 0.8389117, gradient norm = 0.0000561 (50 iterations in 17.562s)
[t-SNE] Iteration 700: error = 0.8325295, gradient norm = 0.0000529 (50 iterations in 13.443s)
[t-SNE] Iteration 750: error = 0.8268463, gradient norm = 0.0000528 (50 iterations in 17.981s)
[t-SNE] Iteration 800: error = 0.8219477, gradient norm = 0.0000477 (50 iterations in 17.448s)
[t-SNE] Iteration 850: error = 0.8180174, gradient norm = 0.0000490 (50 iterations in 18.376s)
[t-SNE] Iteration 900: error = 0.8150476, gradient norm = 0.0000456 (50 iterations in 17.778s)
[t-SNE] Iteration 950: error = 0.8122067, gradient norm = 0.0000472 (50 iterations in 16.983s)
[t-SNE] Iteration 1000: error = 0.8095787, gradient norm = 0.0000489 (50 iterations in 18.581s)
[t-SNE] Error after 1000 iterations: 0.809579

```

```

In [0]: trace1 = go.Scatter3d(
    x=tsne3d[:,0],
    y=tsne3d[:,1],
    z=tsne3d[:,2],
    mode='markers',
    marker=dict(
        sizemode='diameter',
        color = y,
        colorscale = 'Portland',
        colorbar = dict(title = 'duplicate'),
        line=dict(color='rgb(255, 255, 255)'),
        opacity=0.75
    )
)

data=[trace1]
layout=dict(height=800, width=800, title='3d embedding with engineered features')
fig=dict(data=data, layout=layout)

```

```
py.ipplot(fig, filename='3DBubble')
```