In [2]:
```
import numpy as np
import pandas as pd

pd.set_option('display.max_columns', None)
```

In [3]:
```
path = 'datasets/'

df1 = pd.read_csv(path + '1-train_extracted_features.csv')
df2 = pd.read_csv(path + '2-train_nlp_features.csv')
#df3 = pd.read_csv(path + '3-train_vectors.csv')
df3_q1 = pd.read_csv(path + '3.1-train_tfidf_weighted_word2vec_96_Q1.csv')
df3_q2 = pd.read_csv(path + '3.1-train_tfidf_weighted_word2vec_96_Q1.csv')
```

In [4]:
```
df3_q1.head(2)
```

Out[4]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	
0	-6.179507	37.450731	-67.929894	32.224274	143.348826	135.374574	17.865208	54.562352	81.618936	232.909839	27.167002	-6.187220	41.996069	-103.5
1	9.236668	-80.371416	-45.785907	78.291656	183.568221	100.894077	74.344804	48.360802	127.297421	112.987302	73.449294	-47.164479	31.560610	-77.9

In [5]:
```
df3_q2.head(2)
```

Out[5]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	
0	-6.179507	37.450731	-67.929894	32.224274	143.348826	135.374574	17.865208	54.562352	81.618936	232.909839	27.167002	-6.187220	41.996069	-103.5
1	9.236668	-80.371416	-45.785907	78.291656	183.568221	100.894077	74.344804	48.360802	127.297421	112.987302	73.449294	-47.164479	31.560610	-77.9

In [7]:
```
df1.head(1)
```

Out[7]:

	id	qid1	qid2	question1	question2	is_duplicate	freq_qid1	freq_qid2	q1_len	q2_len	num_words_q1	num_words_q2	common_word_q12	total_word_q12	s
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	1	1	66	57	14	12	10	23	

In [10]:
```
st = 'Number of features in'
print(f'{st} preprocessed dataframe  ==> {df1.shape[1]}')
print(f'{st} nlp dataframe  ==> {df2.shape[1]}')
print(f'{st} question1 w2v  dataframe  ==> {df3_q1.shape[1]}')
print(f'{st} question2 w2v  dataframe  ==> {df3_q2.shape[1]}')
print(f'{st} final dataframe ==> {df1.shape[1] + df2.shape[1] + df3_q1.shape[1] + df3_q2.shape[1]}')
```

```
Number of features in preprocessed dataframe  ==> 12
Number of features in nlp dataframe  ==> 28
Number of features in question1 w2v  dataframe  ==> 96
Number of features in question2 w2v  dataframe  ==> 96
Number of features in final dataframe ==> 232
```

In [11]:
```
df3_q1['id'] = df2['id']
df3_q2['id'] = df2['id']

df2 = df2.merge(df1, on='id',how='left')
df1 = df3_q1.merge(df3_q2, on='id',how='left')

result = df2.merge(df1, on='id',how='left')
```

In [12]:
```
result.head(2)
```

Out[12]:

	id	is_duplicate	freq_qid1_x	freq_qid2_x	q1_len_x	q2_len_x	num_words_q1_x	num_words_q2_x	common_word_q12_x	total_word_q12_x	shared_words_q12_x
0	0	0	1	1	66	57	14	12	10	23	0.434783
1	1	0	4	1	51	88	8	13	4	20	0.200000

In [15]:
```
result.is_duplicate.unique()
```

Out[15]:
```
array([0, 1])
```

In [ ]:
```
#storing final features to csv file
result.to_csv(path+'4-train_final_96features.csv',index=False)
```

In [ ]:
```

```