DEEP LEARNING CASE STUDY - *Quora Question Pairs similarity using S-BERT*

> Parameshwari S - CB.SC.I5DAS18026

1. Importing all necessary libraries

In [1]:

```
import numpy as np
import pandas as pd
import pandas_profiling
import string
import random
import math
import time
from sklearn.utils import resample
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
plt.style.use('fivethirtyeight')
import seaborn as sns
sns.set_style('darkgrid')
import os
from os import listdir
import itertools
import collections
import scipy.stats
import nltk
import torch
import zipfile
from nltk import word_tokenize
from nltk.stem import WordNetLemmatizer
from gensim.models.doc2vec import Doc2Vec
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
nltk.download('punkt')
import gensim
from gensim.models import Word2Vec
from gensim.scripts.glove2word2vec import glove2word2vec
from collections import Counter, defaultdict
from tqdm import tqdm
from sklearn import utils
from sklearn import metrics
!pip install -U sentence-transformers
from sentence_transformers import SentenceTransformer
!pip install transformers
from transformers import AutoTokenizer, AutoModel
import warnings
warnings.filterwarnings("ignore")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
Collecting sentence-transformers
  Downloading sentence-transformers-2.1.0.tar.gz (78 kB)
     |████████████████████████████████| 78 kB 3.3 MB/s
Collecting transformers<5.0.0,>=4.6.0
  Downloading transformers-4.12.5-py3-none-any.whl (3.1 MB)
     |████████████████████████████████| 3.1 MB 18.9 MB/s
Collecting tokenizers>=0.10.3
  Downloading tokenizers-0.10.3-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.ma
nylinux2010_x86_64.whl (3.3 MB)
     |████████████████████████████████| 3.3 MB 25.3 MB/s
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from sentence-transformers)
(4.62.3)
Requirement already satisfied: torch>=1.6.0 in /usr/local/lib/python3.7/dist-packages (from sentence-trans
formers) (1.10.0+cu111)
Requirement already satisfied: torchvision in /usr/local/lib/python3.7/dist-packages (from sentence-transf
ormers) (0.11.1+cu111)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from sentence-transformers
) (1.19.5)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from sentence-trans
formers) (1.0.1)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from sentence-transformers
) (1.4.1)
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-packages (from sentence-transformers)
(3.2.5)
Collecting sentencepiece
  Downloading sentencepiece-0.1.96-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.2 MB)
     |████████████████████████████████| 1.2 MB 43.9 MB/s
Collecting huggingface-hub
  Downloading huggingface_hub-0.1.2-py3-none-any.whl (59 kB)
     |████████████████████████████████| 59 kB 6.0 MB/s
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from torch>=1.
6.0->sentence-transformers) (3.10.0.2)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from transformers<5.0.0
,>=4.6.0->sentence-transformers) (2.23.0)
Collecting pyyaml>=5.1
  Downloading PyYAML-6.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux
2010_x86_64.whl (596 kB)
     |████████████████████████████████| 596 kB 43.5 MB/s
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-packages (from transformer
s<5.0.0,>=4.6.0->sentence-transformers) (21.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from transformers<5.0.0
,>=4.6.0->sentence-transformers) (3.3.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transform
ers<5.0.0,>=4.6.0->sentence-transformers) (2019.12.20)
Collecting sacremoses
  Downloading sacremoses-0.0.46-py3-none-any.whl (895 kB)
     |████████████████████████████████| 895 kB 52.2 MB/s
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transfor
mers<5.0.0,>=4.6.0->sentence-transformers) (4.8.2)
Requirement already satisfied: pyparsing<3,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packagi
ng>=20.0->transformers<5.0.0,>=4.6.0->sentence-transformers) (2.4.7)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadat
a->transformers<5.0.0,>=4.6.0->sentence-transformers) (3.6.0)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from nltk->sentence-transfor
mers) (1.15.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-pa
ckages (from requests->transformers<5.0.0,>=4.6.0->sentence-transformers) (1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests-
>transformers<5.0.0,>=4.6.0->sentence-transformers) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->tran
sformers<5.0.0,>=4.6.0->sentence-transformers) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests
->transformers<5.0.0,>=4.6.0->sentence-transformers) (2021.10.8)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from sacremoses->transfor
mers<5.0.0,>=4.6.0->sentence-transformers) (1.1.0)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from sacremoses->transform
ers<5.0.0,>=4.6.0->sentence-transformers) (7.1.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from scikit
-learn->sentence-transformers) (3.0.0)
Requirement already satisfied: pillow!=8.3.0,>=5.3.0 in /usr/local/lib/python3.7/dist-packages (from torch
vision->sentence-transformers) (7.1.2)
Building wheels for collected packages: sentence-transformers
  Building wheel for sentence-transformers (setup.py) ... done
  Created wheel for sentence-transformers: filename=sentence_transformers-2.1.0-py3-none-any.whl size=1210
00 sha256=6430659addd677fd798926a605d7cc0d626988ddb7e1d48fb5344939f0184528
  Stored in directory: /root/.cache/pip/wheels/90/f0/bb/ed1add84da70092ea526466eadc2bfb197c4bcb8d4fa5f7bad
Successfully built sentence-transformers
Installing collected packages: pyyaml, tokenizers, sacremoses, huggingface-hub, transformers, sentencepiec
e, sentence-transformers
  Attempting uninstall: pyyaml
    Found existing installation: PyYAML 3.13
    Uninstalling PyYAML-3.13:
      Successfully uninstalled PyYAML-3.13
Successfully installed huggingface-hub-0.1.2 pyyaml-6.0 sacremoses-0.0.46 sentence-transformers-2.1.0 sent
encepiece-0.1.96 tokenizers-0.10.3 transformers-4.12.5
Requirement already satisfied: transformers in /usr/local/lib/python3.7/dist-packages (4.12.5)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from transformers) (2.2
3.0)
Requirement already satisfied: sacremoses in /usr/local/lib/python3.7/dist-packages (from transformers) (0
.0.46)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from transformers) (3.3
.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transform
ers) (2019.12.20)
Requirement already satisfied: huggingface-hub<1.0,>=0.1.0 in /usr/local/lib/python3.7/dist-packages (from
 transformers) (0.1.2)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.7/dist-packages (from transformers) (4
.62.3)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from transformers) (
1.19.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-packages (from transformer
s) (21.2)
Requirement already satisfied: tokenizers<0.11,>=0.10.1 in /usr/local/lib/python3.7/dist-packages (from tr
ansformers) (0.10.3)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transfo
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transfor
mers) (4.8.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-packages (from transformers) (
6.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (from
huggingface-hub<1.0,>=0.1.0->transformers) (3.10.0.2)
Requirement already satisfied: pyparsing<3,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packagi
ng>=20.0->transformers) (2.4.7)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadat
a->transformers) (3.6.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-pa
ckages (from requests->transformers) (1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests-
>transformers) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests
->transformers) (2021.10.8)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->tran
sformers) (2.10)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from sacremoses->transformer
s) (1.15.0)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from sacremoses->transform
ers) (7.1.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from sacremoses->transfor
mers) (1.1.0)
```

2. Uploading the data

In [5]:

```
train_df = pd.read_csv('/content/train.csv.zip')
train_df.head(3)
```

Out[5]:

	id	qid1	qid2	question1	question2	is_duplicate
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	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0

In [6]:

```
train_df.shape
```

Out[6]:

```
(404290, 6)
```

3. Data preprocessing

In [7]:

```
train_df.isnull().sum()
```

Out[7]:

```
id              0
qid1            0
qid2            0
question1       1
question2       2
is_duplicate    0
dtype: int64
```

In [8]:

```
train_df[train_df.isnull().any(1)]
```

Out[8]:

	id	qid1	qid2	question1	question2	is_duplicate
105780	105780	174363	174364	How can I develop android app?	NaN	0
201841	201841	303951	174364	How can I create an Android app?	NaN	0
363362	363362	493340	493341	NaN	My Chinese name is Haichao Yu. What English na...	0

In [9]:

```
train_df = train_df.fillna(value="")
train_df.isnull().sum()
```

Out[9]:

```
id              0
qid1            0
qid2            0
question1       0
question2       0
is_duplicate    0
dtype: int64
```

4. Data exploration

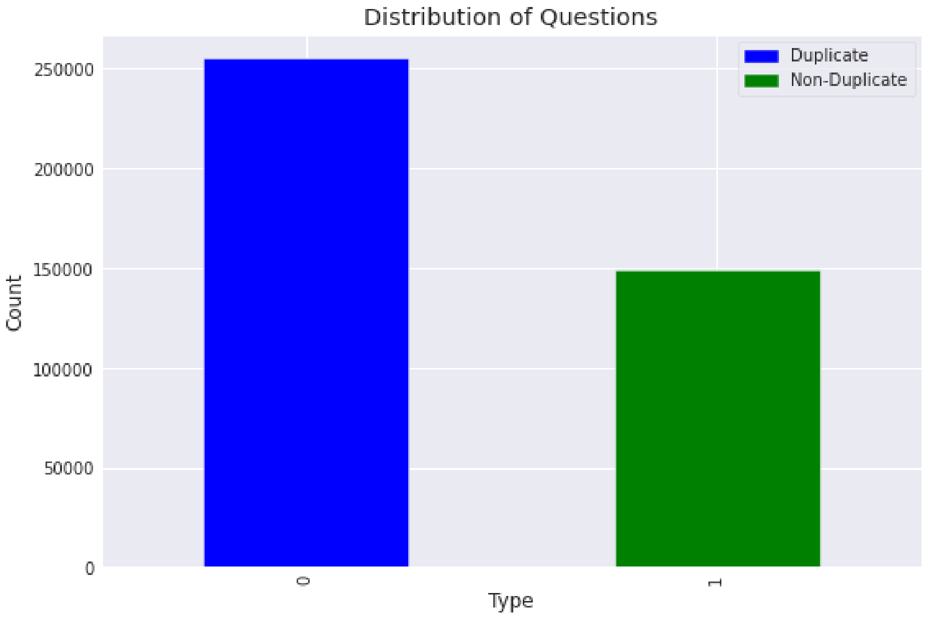
In [ ]:

```
plt.figure(figsize=(8,6))
train_df.is_duplicate.value_counts().plot(kind='bar', color=['b','g'])

D = mpatches.Patch(color='b', label='Duplicate')
ND = mpatches.Patch(color='g', label='Non-Duplicate')

plt.legend(handles=[D,ND], loc='best')

plt.xlabel('Type')
plt.ylabel('Count')
plt.title('Distribution of Questions')
plt.show()
```

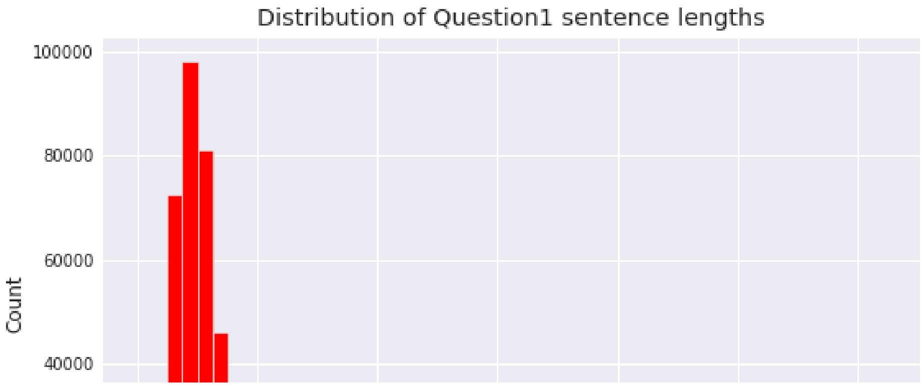


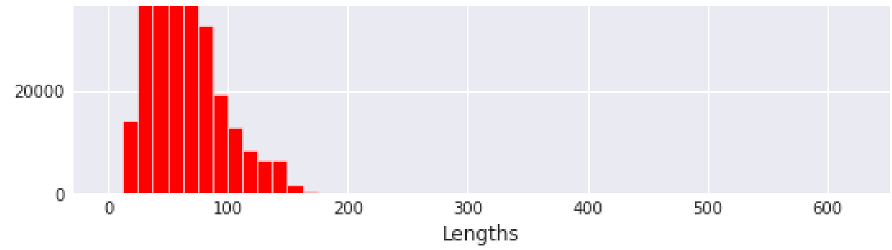
In [ ]:

```
q1_lengths = [len(q1)for q1 in train_df.question1]
print("Mean sentence length for Question1:", np.mean(q1_lengths))

plt.figure(figsize=(8,6))
plt.hist(q1_lengths,bins=50,color='r')
plt.xlabel('Lengths')
plt.ylabel('Count')
plt.title('Distribution of Question1 sentence lengths')
plt.show()
```

```
Mean sentence length for Question1: 59.53670879813995
```



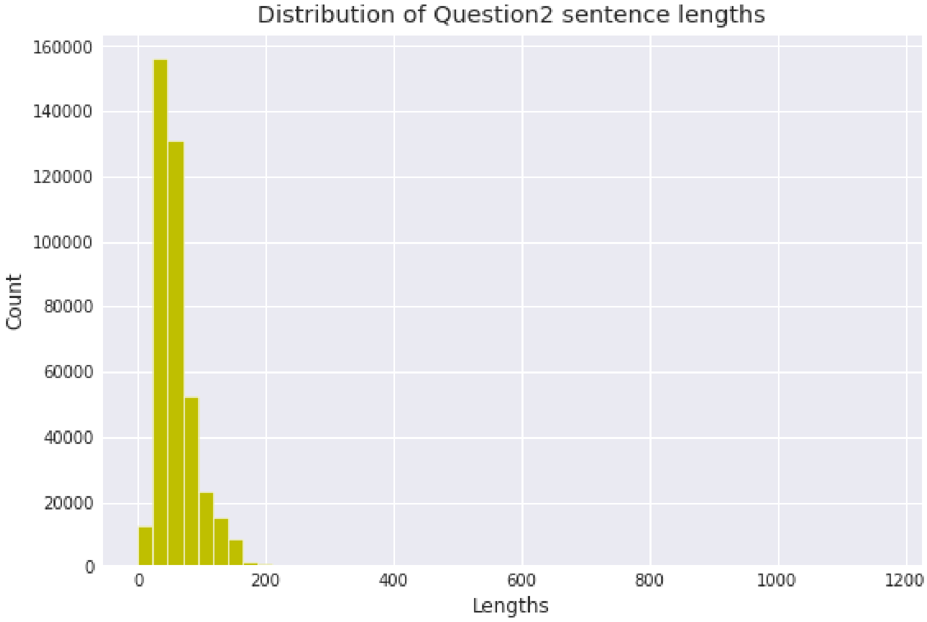


In [ ]:

```
q2_lengths = [len(q2)for q2 in train_df.question2]
print("Mean sentence length for Question2:", np.mean(q2_lengths))

plt.figure(figsize=(8,6))
plt.hist(q2_lengths,bins=50,color='y')
plt.xlabel('Lengths')
plt.ylabel('Count')
plt.title('Distribution of Question2 sentence lengths')
plt.show()
```

```
Mean sentence length for Question2: 60.10836528234683
```



5. S-BERT Embeddings

> **a) 100000 rows**

In [ ]:

```
st_model = SentenceTransformer('bert-base-nli-mean-tokens')
```

In [ ]:

```
sbert_df = train_df[:100000]
```

In [ ]:

```
sentences_question1 = list(sent for sent in sbert_df['question1'].values)
sentences_question2 = list(sent for sent in sbert_df['question2'].values)
```

In [ ]:

```
def generate_sent_embeddings(data):
    return st_model.encode(data)
```

In [ ]:

```
question1_sent_embeddings = generate_sent_embeddings(sentences_question1)
print("shape of question1 sentence embeddings:", question1_sent_embeddings.shape)
```

```
shape of question1 sentence embeddings: (100000, 768)
```

In [ ]:

```
question2_sent_embeddings = generate_sent_embeddings(sentences_question2)
print("shape of question2 sentence embeddings:", question2_sent_embeddings.shape)
```

```
shape of question2 sentence embeddings: (100000, 768)
```

In [ ]:

```
sbert_df['question1_sent_embeddings'] = pd.DataFrame({'question1_sent_embeddings' : list(question1_sent_em
beddings)})
sbert_df['question2_sent_embeddings'] = pd.DataFrame({'question2_sent_embeddings' : list(question2_sent_em
beddings)})
```

In [ ]:

```
cos_sim = []
spear_corr = []
for index, row in sbert_df.iterrows():
  cos_sim.append(cosine_similarity([row['question1_sent_embeddings']],[row['question2_sent_embeddings']]))
  spear_corr.append(scipy.stats.spearmanr(row['question1_sent_embeddings'],row['question2_sent_embeddings
'])[0])
sbert_df['cos_sim'] = cos_sim
sbert_df['spear_corr'] = spear_corr
```

In [16]:

```
def similarity_to_predictions(cos_sim, threshold):
    if (cos_sim >= threshold):
        return 1
    else:
        return 0
```

In [ ]:

```
sbert_df['pred_res(cos_sim)'] = sbert_df['cos_sim'].apply(similarity_to_predictions, threshold=0.87)
sbert_df['pred_res(spear_corr)'] = sbert_df['spear_corr'].apply(similarity_to_predictions, threshold=0.86
)
```

In [ ]:

```
sbert_df.head(3)
```

Out[ ]:

	id	qid1	qid2	question1	question2	is_duplicate	question1_sent_embeddings	question2_sent_embeddings	cos_sim	spear_corr	pr
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	[-0.009722352, -0.32162306, 0.9211391, 0.12629...	[0.15146354, -0.20154329, 0.9581177, 0.0159406...	[[0.84010166]]	0.810172	
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	[0.27386734, 0.47279105, -0.6623544, 0.1045286...	[0.19313551, 0.09134984, -1.0451194, 0.5032031...	[[0.7469238]]	0.731909	
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0	[-0.20832907, -0.15172529, 1.1032256, 0.248804...	[0.27955115, 0.0012331137, -0.03924411, 0.3699...	[[0.89106655]]	0.881655	

In [ ]:

```
print("Accuracy for SBERT embeddings using cosine similarity - ", metrics.accuracy_score(sbert_df['is_dupl
icate'], sbert_df['pred_res(cos_sim)']))
print("Accuracy for SBERT embeddings using spearman's correlation- ", metrics.accuracy_score(sbert_df['is_
duplicate'], sbert_df['pred_res(spear_corr)']))
```

```
Accuracy for SBERT embeddings using cosine similarity -  0.72892
Accuracy for SBERT embeddings using spearman's correlation-  0.7286
```

> **b) 500 rows**

In [ ]:

```
sbert_df1 = train_df[:500]
```

In [ ]:

```
sentences_question1 = list(sent for sent in sbert_df1['question1'].values)
sentences_question2 = list(sent for sent in sbert_df1['question2'].values)
```

In [ ]:

```
question1_sent_embeddings = generate_sent_embeddings(sentences_question1)
print("shape of question1 sentence embeddings:", question1_sent_embeddings.shape)
```

```
shape of question1 sentence embeddings: (500, 768)
```

In [ ]:

```
question2_sent_embeddings = generate_sent_embeddings(sentences_question2)
print("shape of question2 sentence embeddings:", question2_sent_embeddings.shape)
```

```
shape of question2 sentence embeddings: (500, 768)
```

In [ ]:

```
sbert_df1['question1_sent_embeddings'] = pd.DataFrame({'question1_sent_embeddings' : list(question1_sent_e
mbeddings)})
sbert_df1['question2_sent_embeddings'] = pd.DataFrame({'question2_sent_embeddings' : list(question2_sent_e
mbeddings)})
```

In [ ]:

```
cos_sim = []
spear_corr = []
for index, row in sbert_df1.iterrows():
  cos_sim.append(cosine_similarity([row['question1_sent_embeddings']],[row['question2_sent_embeddings']]))
  spear_corr.append(scipy.stats.spearmanr(row['question1_sent_embeddings'],row['question2_sent_embeddings
'])[0])
sbert_df1['cos_sim'] = cos_sim
sbert_df1['spear_corr'] = spear_corr
```

In [ ]:

```
sbert_df1['pred_res(cos_sim)'] = sbert_df1['cos_sim'].apply(similarity_to_predictions, threshold=0.86)
sbert_df1['pred_res(spear_corr)'] = sbert_df1['spear_corr'].apply(similarity_to_predictions, threshold=0.
86)
```

In [ ]:

```
sbert_df1.head(3)
```

Out[ ]:

	id	qid1	qid2	question1	question2	is_duplicate	question1_sent_embeddings	question2_sent_embeddings	cos_sim	spear_corr	pr
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	[-0.009722352, -0.32162306, 0.9211391, 0.12629...	[0.15146354, -0.20154329, 0.9581177, 0.0159406...	[[0.84010166]]	0.810172	
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	[0.27386734, 0.47279105, -0.6623544, 0.1045286...	[0.19313551, 0.09134984, -1.0451194, 0.5032031...	[[0.7469238]]	0.731909	
				How can I increase	How can Internet						

	id	qid1	qid2	question1	question2	is_duplicate	question1_sent_embeddings	question2_sent_embeddings	cos_sim	spear_corr	pr
2	2	5	6	the speed of my internet co...	speed be increased by hacking...	0	[-0.20832907, -0.15172529, 1.1032256, 0.248804...	[0.27955115, 0.0012331137, -0.03924411, 0.3699...	[[0.8910655]]	0.881655	

In [ ]:

```
print("Accuracy for SBERT embeddings using cosine similarity - ", metrics.accuracy_score(sbert_df1['is_duplicate'], sbert_df1['pred_res(cos_sim)']))
print("Accuracy for SBERT embeddings using spearman's correlation- ", metrics.accuracy_score(sbert_df1['is_duplicate'], sbert_df1['pred_res(spear_corr)']))
```

```
Accuracy for SBERT embeddings using cosine similarity -  0.728
Accuracy for SBERT embeddings using spearman's correlation-  0.726
```

6. BERT Embeddings

In [2]:

```
!pip install BERTSimilarity
```

```
Collecting BERTSimilarity
  Downloading BERTSimilarity-0.1.tar.gz (2.7 kB)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from BERTSimilarity) (1.19
.5)
Requirement already satisfied: torch in /usr/local/lib/python3.7/dist-packages (from BERTSimilarity) (1.10
.0+cu111)
Requirement already satisfied: transformers in /usr/local/lib/python3.7/dist-packages (from BERTSimilarity
) (4.12.5)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from BERTSimilarity) (1.4.
1)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from torch->BE
RTSimilarity) (3.10.0.2)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transfor
mers->BERTSimilarity) (4.8.2)
Requirement already satisfied: huggingface-hub<1.0,>=0.1.0 in /usr/local/lib/python3.7/dist-packages (from
transformers->BERTSimilarity) (0.1.2)
Requirement already satisfied: tokenizers<0.11,>=0.10.1 in /usr/local/lib/python3.7/dist-packages (from tr
ansformers->BERTSimilarity) (0.10.3)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-packages (from transformers->B
ERTSimilarity) (6.0)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.7/dist-packages (from transformers->BE
RTSimilarity) (4.62.3)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from transformers->BERT
Similarity) (2.23.0)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-packages (from transformer
s->BERTSimilarity) (21.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transform
ers->BERTSimilarity) (2019.12.20)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from transformers->BERT
Similarity) (3.3.2)
Requirement already satisfied: sacremoses in /usr/local/lib/python3.7/dist-packages (from transformers->BE
RTSimilarity) (0.0.46)
Requirement already satisfied: pyparsing<3,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packagi
ng>=20.0->transformers->BERTSimilarity) (2.4.7)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadat
a->transformers->BERTSimilarity) (3.6.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-pa
ckages (from requests->transformers->BERTSimilarity) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->tran
sformers->BERTSimilarity) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests
->transformers->BERTSimilarity) (2021.10.8)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests-
>transformers->BERTSimilarity) (3.0.4)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from sacremoses->transform
ers->BERTSimilarity) (7.1.2)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from sacremoses->transformer
s->BERTSimilarity) (1.15.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from sacremoses->transfor
mers->BERTSimilarity) (1.1.0)
Building wheels for collected packages: BERTSimilarity
  Building wheel for BERTSimilarity (setup.py) ... done
  Created wheel for BERTSimilarity: filename=BERTSimilarity-0.1-py3-none-any.whl size=3612 sha256=0dfd843f
7ee268ab52d5b733fc6cd868ca7593062ca9a551aae0b44ba5aadc4a
  Stored in directory: /root/.cache/pip/wheels/fa/f7/22/510c1c7131e536fb02b71c619dddcce9636913654ba2f22f22
Successfully built BERTSimilarity
Installing collected packages: BERTSimilarity
Successfully installed BERTSimilarity-0.1
```

In [3]:

```
import torch
from transformers import BertTokenizer,BertModel
from scipy.spatial.distance import cosine
class BERTSimilarity():
    def bert_tokenize(self,data):
        self.data=data
        self.output_tokens=''
        self.output_tokens+='[CLS] ' +self.data+' [SEP]'
        return self.output_tokens
    def sentential_embeddings(self,tokenizer,tokenized_text):
        self.tokenizer=tokenizer
        self.tokenized_text=tokenized_text
        self.idx_tokens=self.tokenizer.convert_tokens_to_ids(self.tokenized_text)
        self.segmenter_idx=[1]*len(self.tokenized_text)
        self.tokens_tensor=torch.tensor([self.idx_tokens])
        self.segmenter_tensor=torch.tensor([self.segmenter_idx])
        self.model=BertModel.from_pretrained('bert-base-uncased',output_hidden_states=True)
        self.model.eval()
        with torch.no_grad():
            self.outputs=self.model(self.tokens_tensor,self.segmenter_tensor)
            self.hidden_state=self.outputs[2]
        self.embedding_token=torch.stack(self.hidden_state,dim=0)
        self.embedding_token=torch.squeeze(self.embedding_token,dim=1)
        self.embedding_token=self.embedding_token.permute(1,0,2)
        self.vs_sum_cat=[]
        for i in self.embedding_token:
            vs_li=torch.sum(i[-4:],dim=0)
            self.vs_sum_cat.append(vs_li)
        self.token_vecs=self.hidden_state[-2][0]
        self.sentence_embeddings=torch.mean(self.token_vecs,dim=0)
        return self.sentence_embeddings,self.vs_sum_cat
    def calculate_distance(self,sentence_1,sentence_2):
        self.sentence_1=sentence_1
        self.sentence_2=sentence_2
        self.tokenizer=BertTokenizer.from_pretrained('bert-base-uncased')
        self.preprocess_1=self.bert_tokenize(self.sentence_1)
        self.preprocess_2=self.bert_tokenize(self.sentence_2)
        self.tokenized_text_1=self.tokenizer.tokenize(self.preprocess_1)
        self.tokenized_text_2=self.tokenizer.tokenize(self.preprocess_2)
        self.sentence_1,self.vs_sum_cat1=self.sentential_embeddings(self.tokenizer,self.tokenized_text_1)
        self.sentence_2,self.vs_sum_cat2=self.sentential_embeddings(self.tokenizer,self.tokenized_text_2)
        self.distance=1-cosine(self.sentence_1,self.sentence_2)
        return self.distance
    def corr(self,sentence_1,sentence_2):
        self.sentence_1=sentence_1
        self.sentence_2=sentence_2
        self.tokenizer=BertTokenizer.from_pretrained('bert-base-uncased')
        self.preprocess_1=self.bert_tokenize(self.sentence_1)
        self.preprocess_2=self.bert_tokenize(self.sentence_2)
        self.tokenized_text_1=self.tokenizer.tokenize(self.preprocess_1)
        self.tokenized_text_2=self.tokenizer.tokenize(self.preprocess_2)
        self.sentence_1,self.vs_sum_cat1=self.sentential_embeddings(self.tokenizer,self.tokenized_text_1)
        self.sentence_2,self.vs_sum_cat2=self.sentential_embeddings(self.tokenizer,self.tokenized_text_2)
        self.spcorr=scipy.stats.spearmanr(self.sentence_1,self.sentence_2)[0]
        return self.spcorr
```

In [4]:

```
bertsimilarity=BERTSimilarity()
```

In [12]:

```
from transformers import logging
logging.set_verbosity_error()
```

In [13]:

```
bert_df = train_df[:500]
```

In [14]:

```
distances=[]
spear_corr = []
for i in range(len(bert_df)):
  q1=bert_df['question1'][i]
  q2=bert_df['question2'][i]
  distances.append(bertsimilarity.calculate_distance(q1,q2))
  spear_corr.append(bertsimilarity.corr(q1,q2))

bert_df['cos_sim']=distances
```

```
bert_df['spear_corr']=spear_corr
```

In [29]:

```
bert_df['pred_res(cos_sim)'] = bert_df['cos_sim'].apply(similarity_to_predictions, threshold=0.89)
bert_df['pred_res(spear_corr)'] = bert_df['spear_corr'].apply(similarity_to_predictions, threshold=0.87)
```

In [30]:

```
bert_df.head()
```

Out[30]:

	id	qid1	qid2	question1	question2	is_duplicate	cos_sim	spear_corr	pred_res(cos_sim)	pred_res(spear_corr)
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	0.970151	0.937000	1	1
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	0.905713	0.770143	1	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0	0.923254	0.811675	1	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0	0.610703	0.356798	0	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0	0.761775	0.490774	0	0

In [31]:

```
print("Accuary for BERT embeddings using cosine similarity- ", metrics.accuracy_score(bert_df['is_duplicat
e'], bert_df['pred_res(cos_sim)']))
print("Accuary for BERT embeddings using spearman's correlation- ", metrics.accuracy_score(bert_df['is_dup
licate'], bert_df['pred_res(spear_corr)']))
```

```
Accuary for BERT embeddings using cosine similarity-  0.692
Accuary for BERT embeddings using spearman's correlation-  0.662
```

7. Universal Sentene Encoder embeddings

In [ ]:

```
!pip3 install --upgrade tensorflow-gpu
!pip3 install tensorflow-hub
```

```
Collecting tensorflow-gpu
  Downloading tensorflow_gpu-2.7.0-cp37-cp37m-manylinux2010_x86_64.whl (489.6 MB)
     |████████████████████████████████| 489.6 MB 24 kB/s
Requirement already satisfied: opt-einsum>=2.3.2 in /usr/local/lib/python3.7/dist-packages (from tensorflo
w-gpu) (3.3.0)
Requirement already satisfied: termcolor>=1.1.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow
-gpu) (1.1.0)
Requirement already satisfied: tensorflow-estimator<2.8,~=2.7.0rc0 in /usr/local/lib/python3.7/dist-packag
es (from tensorflow-gpu) (2.7.0)
Requirement already satisfied: wheel<1.0,>=0.32.0 in /usr/local/lib/python3.7/dist-packages (from tensorfl
ow-gpu) (0.37.0)
Requirement already satisfied: google-pasta>=0.1.1 in /usr/local/lib/python3.7/dist-packages (from tensorf
low-gpu) (0.2.0)
Requirement already satisfied: keras-preprocessing>=1.1.1 in /usr/local/lib/python3.7/dist-packages (from
tensorflow-gpu) (1.1.2)
Requirement already satisfied: typing-extensions>=3.6.6 in /usr/local/lib/python3.7/dist-packages (from te
nsorflow-gpu) (3.10.0.2)
Requirement already satisfied: six>=1.12.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow-gpu)
(1.15.0)
Requirement already satisfied: flatbuffers<3.0,>=1.12 in /usr/local/lib/python3.7/dist-packages (from tens
orflow-gpu) (2.0)
Requirement already satisfied: absl-py>=0.4.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow-g
pu) (0.12.0)
Requirement already satisfied: gast<0.5.0,>=0.2.1 in /usr/local/lib/python3.7/dist-packages (from tensorfl
ow-gpu) (0.4.0)
Requirement already satisfied: wrapt>=1.11.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow-gp
u) (1.13.3)
Requirement already satisfied: astunparse>=1.6.0 in /usr/local/lib/python3.7/dist-packages (from tensorflo
w-gpu) (1.6.3)
Requirement already satisfied: h5py>=2.9.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow-gpu)
(3.1.0)
```

```
Requirement already satisfied: keras<2.8,>=2.7.0rc0 in /usr/local/lib/python3.7/dist-packages (from tensor
flow-gpu) (2.7.0)
Requirement already satisfied: grpcio<2.0,>=1.24.3 in /usr/local/lib/python3.7/dist-packages (from tensorf
low-gpu) (1.41.1)
Requirement already satisfied: libclang>=9.0.1 in /usr/local/lib/python3.7/dist-packages (from tensorflow-
gpu) (12.0.0)
Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.21.0 in /usr/local/lib/python3.7/dist-packa
ges (from tensorflow-gpu) (0.22.0)
Requirement already satisfied: protobuf>=3.9.2 in /usr/local/lib/python3.7/dist-packages (from tensorflow-
gpu) (3.17.3)
Requirement already satisfied: numpy>=1.14.5 in /usr/local/lib/python3.7/dist-packages (from tensorflow-gp
u) (1.19.5)
Requirement already satisfied: tensorboard~=2.6 in /usr/local/lib/python3.7/dist-packages (from tensorflow
-gpu) (2.7.0)
Requirement already satisfied: cached-property in /usr/local/lib/python3.7/dist-packages (from h5py>=2.9.0
->tensorflow-gpu) (1.5.2)
Requirement already satisfied: tensorboard-data-server<0.7.0,>=0.6.0 in /usr/local/lib/python3.7/dist-pack
ages (from tensorboard~=2.6->tensorflow-gpu) (0.6.1)
Requirement already satisfied: google-auth<3,>=1.6.3 in /usr/local/lib/python3.7/dist-packages (from tenso
rboard~=2.6->tensorflow-gpu) (1.35.0)
Requirement already satisfied: setuptools>=41.0.0 in /usr/local/lib/python3.7/dist-packages (from tensorbo
ard~=2.6->tensorflow-gpu) (57.4.0)
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in /usr/local/lib/python3.7/dist-packages (fr
om tensorboard~=2.6->tensorflow-gpu) (1.8.0)
Requirement already satisfied: requests<3,>=2.21.0 in /usr/local/lib/python3.7/dist-packages (from tensorb
oard~=2.6->tensorflow-gpu) (2.23.0)
Requirement already satisfied: google-auth-oauthlib<0.5,>=0.4.1 in /usr/local/lib/python3.7/dist-packages
(from tensorboard~=2.6->tensorflow-gpu) (0.4.6)
Requirement already satisfied: werkzeug>=0.11.15 in /usr/local/lib/python3.7/dist-packages (from tensorboa
rd~=2.6->tensorflow-gpu) (1.0.1)
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.7/dist-packages (from tensorboard
~=2.6->tensorflow-gpu) (3.3.4)
Requirement already satisfied: rsa<5,>=3.1.4 in /usr/local/lib/python3.7/dist-packages (from google-auth<3
,>=1.6.3->tensorboard~=2.6->tensorflow-gpu) (4.7.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /usr/local/lib/python3.7/dist-packages (from googl
e-auth<3,>=1.6.3->tensorboard~=2.6->tensorflow-gpu) (0.2.8)
Requirement already satisfied: cachetools<5.0,>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from goog
le-auth<3,>=1.6.3->tensorboard~=2.6->tensorflow-gpu) (4.2.4)
Requirement already satisfied: requests-oauthlib>=0.7.0 in /usr/local/lib/python3.7/dist-packages (from go
ogle-auth-oauthlib<0.5,>=0.4.1->tensorboard~=2.6->tensorflow-gpu) (1.3.0)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from markdown
>=2.6.8->tensorboard~=2.6->tensorflow-gpu) (4.8.2)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in /usr/local/lib/python3.7/dist-packages (from pyasn1
-modules>=0.2.1->google-auth<3,>=1.6.3->tensorboard~=2.6->tensorflow-gpu) (0.4.8)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests
<3,>=2.21.0->tensorboard~=2.6->tensorflow-gpu) (2021.10.8)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests<3,>=2
.21.0->tensorboard~=2.6->tensorflow-gpu) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests<
3,>=2.21.0->tensorboard~=2.6->tensorflow-gpu) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-pa
ckages (from requests<3,>=2.21.0->tensorboard~=2.6->tensorflow-gpu) (1.24.3)
Requirement already satisfied: oauthlib>=3.0.0 in /usr/local/lib/python3.7/dist-packages (from requests-oa
uthlib>=0.7.0->google-auth-oauthlib<0.5,>=0.4.1->tensorboard~=2.6->tensorflow-gpu) (3.1.1)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadat
a->markdown>=2.6.8->tensorboard~=2.6->tensorflow-gpu) (3.6.0)
Installing collected packages: tensorflow-gpu
Successfully installed tensorflow-gpu-2.7.0

Requirement already satisfied: tensorflow-hub in /usr/local/lib/python3.7/dist-packages (0.12.0)
Requirement already satisfied: protobuf>=3.8.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow-
hub) (3.17.3)
Requirement already satisfied: numpy>=1.12.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow-hu
b) (1.19.5)
Requirement already satisfied: six>=1.9 in /usr/local/lib/python3.7/dist-packages (from protobuf>=3.8.0->t
ensorflow-hub) (1.15.0)
```

In [ ]:

```
import tensorflow as tf
import tensorflow_hub as hub
```

In [ ]:

```
module_url = "https://tfhub.dev/google/universal-sentence-encoder/4"
model = hub.load(module_url)
print ("module %s loaded" % module_url)
```

```
INFO:absl:Using /tmp/tfhub_modules to cache modules.
INFO:absl:Downloading TF-Hub Module 'https://tfhub.dev/google/universal-sentence-encoder/4'.
INFO:absl:Downloaded https://tfhub.dev/google/universal-sentence-encoder/4, Total size: 987.47MB
INFO:absl:Downloaded TF-Hub Module 'https://tfhub.dev/google/universal-sentence-encoder/4'.
```

```
module https://tfhub.dev/google/universal-sentence-encoder/4 loaded
```

In [ ]:

```
use_df = train_df[:500]
```

In [ ]:

```
sentences_question1 = list(sent for sent in use_df['question1'].values)
sentences_question2 = list(sent for sent in use_df['question2'].values)
```

In [ ]:

```
sentence1_embeddings = model(sentences_question1)
print("shape of question1 sentence embeddings:", sentence1_embeddings.shape)
sentence2_embeddings = model(sentences_question2)
print("shape of question2 sentence embeddings:", sentence2_embeddings.shape)
```

```
shape of question1 sentence embeddings: (500, 512)
shape of question2 sentence embeddings: (500, 512)
```

In [ ]:

```
use_df['question1_sent_embeddings'] = pd.DataFrame({'question1_sent_embeddings' : list(sentence1_embedding
s)})
use_df['question2_sent_embeddings'] = pd.DataFrame({'question2_sent_embeddings' : list(sentence2_embedding
s)})
```

In [ ]:

```
cos_sim = []
spear_corr = []
for index, row in use_df.iterrows():
  cos_sim.append(cosine_similarity([row['question1_sent_embeddings']],[row['question2_sent_embeddings']]))
  spear_corr.append(scipy.stats.spearmanr(row['question1_sent_embeddings'],row['question2_sent_embeddings
'])[0])
use_df['cos_sim'] = cos_sim
use_df['spear_corr'] = spear_corr
```

In [ ]:

```
use_df['pred_res(cos_sim)'] = use_df['cos_sim'].apply(similarity_to_predictions, threshold=0.86)
use_df['pred_res(spear_corr)'] = use_df['spear_corr'].apply(similarity_to_predictions, threshold=0.88)
```

In [ ]:

```
use_df.head(3)
```

Out[ ]:

	id	qid1	qid2	question1	question2	is_duplicate	question1_sent_embeddings	question2_sent_embeddings	cos_sim	spear_corr	pr
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	(tf.Tensor(0.0021821507, shape=(), dtype=float...	(tf.Tensor(0.018747559, shape=(), dtype=float3...	[[0.9364382]]	0.934645	
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	(tf.Tensor(-0.0081168795, shape=(), dtype=floa...	(tf.Tensor(-0.026330141, shape=(), dtype=float...	[[0.68438935]]	0.675223	
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0	(tf.Tensor(-0.025076203, shape=(), dtype=float...	(tf.Tensor(-0.019373633, shape=(), dtype=float...	[[0.60938096]]	0.589570	

In [ ]:

```
print("Accuracy for USE embeddings using cosine similarity - ", metrics.accuracy_score(use_df['is_duplicat
e'], use_df['pred_res(cos_sim)']))
print("Accuracy for USE embeddings using spearman's correlation- ", metrics.accuracy_score(use_df['is_dupl
icate'], use_df['pred_res(spear_corr)']))
```

```
Accuracy for USE embeddings using cosine similarity -  0.688
Accuracy for USE embeddings using spearman's correlation-  0.686
```

8. RoBERTa embeddings

In [ ]:

```
rb_model = SentenceTransformer('roberta-base-nli-stsb-mean-tokens')
```

In [ ]:

```
roberta_df = train_df[:500]
```

In [ ]:

```
sentences_question1 = list(sent for sent in roberta_df['question1'].values)
sentences_question2 = list(sent for sent in roberta_df['question2'].values)
```

In [ ]:

```
def generate_sent_embeddings(data):
    return rb_model.encode(data)
```

In [ ]:

```
sentence1_embeddings = generate_sent_embeddings(sentences_question1)
print("shape of question1 sentence embeddings:", sentence1_embeddings.shape)
sentence2_embeddings = generate_sent_embeddings(sentences_question2)
print("shape of question2 sentence embeddings:", sentence2_embeddings.shape)
```

```
shape of question1 sentence embeddings: (500, 768)
shape of question2 sentence embeddings: (500, 768)
```

In [ ]:

```
roberta_df['question1_sent_embeddings'] = pd.DataFrame({'question1_sent_embeddings' : list(question1_sent_
embeddings)})
roberta_df['question2_sent_embeddings'] = pd.DataFrame({'question2_sent_embeddings' : list(question2_sent_
embeddings)})
```

In [ ]:

```
cos_sim = []
spear_corr = []
for index, row in roberta_df.iterrows():
  cos_sim.append(cosine_similarity([row['question1_sent_embeddings']],[row['question2_sent_embeddings']]))
  spear_corr.append(scipy.stats.spearmanr(row['question1_sent_embeddings'],row['question2_sent_embeddings
'])[0])
roberta_df['cos_sim'] = cos_sim
roberta_df['spear_corr'] = spear_corr
```

In [ ]:

```
roberta_df['pred_res(cos_sim)'] = roberta_df['cos_sim'].apply(similarity_to_predictions, threshold=0.86)
roberta_df['pred_res(spear_corr)'] = roberta_df['spear_corr'].apply(similarity_to_predictions, threshold=
0.86)
```

In [ ]:

```
roberta_df.head(3)
```

Out[ ]:

	id	qid1	qid2	question1	question2	is_duplicate	question1_sent_embeddings	question2_sent_embeddings	cos_sim	spear_corr	pr
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	[-0.009722352, -0.32162306, 0.9211391, 0.12629...	[0.15146354, -0.20154329, 0.9581177, 0.0159406...	[[0.84010166]]	0.810172	

	id	qid1	qid2	question1	question2	is_duplicate	question1_sent_embeddings	question2_sent_embeddings	cos_sim	spear_corr	pr
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	[0.27386734, 0.47279105, -0.6623544, 0.1045286...	[0.19313551, 0.09134984, -1.0451194, 0.5032031...	[[0.7469238]]	0.731909	
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0	[-0.20832907, -0.15172529, 1.1032256, 0.248804...	[0.27955115, 0.0012331137, -0.03924411, 0.3699...	[[0.89106655]]	0.881655	

In [ ]:

```
print("Accuary for RoBERTa embeddings using cosine similarity- ", metrics.accuracy_score(roberta_df['is_du
plicate'], roberta_df['pred_res(cos_sim)']))
print("Accuary for RoBERTa embeddings using spearman's correlation- ", metrics.accuracy_score(roberta_df[
'is_duplicate'], roberta_df['pred_res(spear_corr)']))
```

```
Accuary for RoBERTa embeddings using cosine similarity-  0.728
Accuary for RoBERTa embeddings using spearman's correlation-  0.726
```

9. InferSent embeddings

In [ ]:

```
!wget -c http://nlp.stanford.edu/data/glove.840B.300d.zip
!wget -c https://dl.fbaipublicfiles.com/infersent/infersent1.pkl
!wget -c https://raw.githubusercontent.com/facebookresearch/InferSent/master/models.py
```

```
--2021-11-19 05:28:05--  http://nlp.stanford.edu/data/glove.840B.300d.zip
Resolving nlp.stanford.edu (nlp.stanford.edu)... 171.64.67.140
Connecting to nlp.stanford.edu (nlp.stanford.edu)|171.64.67.140|:80... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://nlp.stanford.edu/data/glove.840B.300d.zip [following]
--2021-11-19 05:28:06--  https://nlp.stanford.edu/data/glove.840B.300d.zip
Connecting to nlp.stanford.edu (nlp.stanford.edu)|171.64.67.140|:443... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: http://downloads.cs.stanford.edu/nlp/data/glove.840B.300d.zip [following]
--2021-11-19 05:28:06--  http://downloads.cs.stanford.edu/nlp/data/glove.840B.300d.zip
Resolving downloads.cs.stanford.edu (downloads.cs.stanford.edu)... 171.64.64.22
Connecting to downloads.cs.stanford.edu (downloads.cs.stanford.edu)|171.64.64.22|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2176768927 (2.0G) [application/zip]
Saving to: ‘glove.840B.300d.zip’

glove.840B.300d.zip 100%[===================>]   2.03G  5.03MB/s    in 6m 53s

2021-11-19 05:34:59 (5.03 MB/s) - ‘glove.840B.300d.zip’ saved [2176768927/2176768927]

--2021-11-19 05:34:59--  https://dl.fbaipublicfiles.com/infersent/infersent1.pkl
Resolving dl.fbaipublicfiles.com (dl.fbaipublicfiles.com)... 104.22.74.142, 172.67.9.4, 104.22.75.142, ...
Connecting to dl.fbaipublicfiles.com (dl.fbaipublicfiles.com)|104.22.74.142|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 154010676 (147M) [application/octet-stream]
Saving to: ‘infersent1.pkl’

infersent1.pkl      100%[===================>] 146.88M  28.7MB/s    in 5.6s

2021-11-19 05:35:05 (26.2 MB/s) - ‘infersent1.pkl’ saved [154010676/154010676]

--2021-11-19 05:35:06--  https://raw.githubusercontent.com/facebookresearch/InferSent/master/models.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.109.133, 185.19
9.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 9875 (9.6K) [text/plain]
Saving to: ‘models.py’

models.py           100%[===================>]   9.64K  --.-KB/s    in 0s

2021-11-19 05:35:06 (81.2 MB/s) - ‘models.py’ saved [9875/9875]
```

In [ ]:

```
with zipfile.ZipFile("/content/glove.840B.300d.zip", "r") as zipread:
```

```
with zipfile.ZipFile("/content/glove.840B.300d.zip", "r") as zipread:
  zipread.extractall("/content/")
  zipread.close
```

In [ ]:

```
from models import InferSent

MODEL_PATH = '/content/infersent1.pkl'
params_model = {'bsize': 64, 'word_emb_dim': 300, 'enc_lstm_dim': 2048,
                'pool_type': 'max', 'dpout_model': 0.0, 'version': 2}
infersent = InferSent(params_model)
infersent.load_state_dict(torch.load(MODEL_PATH))

infersent.set_w2v_path("/content/glove.840B.300d.txt")
```

In [ ]:

```
infersent_df = train_df[:500]
```

In [ ]:

```
import itertools
from itertools import chain

infersent.build_vocab(list(chain(infersent_df.question1, infersent_df.question2)), tokenize=True)
```

```
Found 2665(/2739) words with w2v vectors
Vocab size : 2665
```

In [ ]:

```
embeddings1 = infersent.encode(infersent_df.question1, tokenize=True)
embeddings2 = infersent.encode(infersent_df.question2, tokenize=True)
```

In [ ]:

```
cos_sim = []
spear_corr = []
for (e1, e2) in zip(embeddings1, embeddings2):
  cos_sim.append(cosine_similarity(e1.reshape(1,-1), e2.reshape(1,-1))[0][0])
  spear_corr.append(scipy.stats.spearmanr(e1, e2)[0])
infersent_df["cos_sim"] = cos_sim
infersent_df["spear_corr"] = spear_corr
```

In [ ]:

```
infersent_df['pred_res(cos_sim)'] = infersent_df['cos_sim'].apply(similarity_to_predictions, threshold=0.85)
infersent_df['pred_res(spear_corr)'] = infersent_df['spear_corr'].apply(similarity_to_predictions, threshold=0.85)
```

In [ ]:

```
infersent_df.head(3)
```

Out[ ]:

	id	qid1	qid2	question1	question2	is_duplicate	cos_sim	spear_corr	pred_res(cos_sim)	pred_res(spear_corr)
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	0.953249	0.921047	1	1
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	0.832436	0.702478	0	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0	0.868755	0.768777	1	0

In [ ]:

```
print("Accuary for InferSent embeddings using cosine similarity- ", metrics.accuracy_score(infersent_df['is_duplicate'], infersent_df['pred_res(cos_sim)']))
print("Accuary for InferSent embeddings using spearman's correlation- ", metrics.accuracy_score(infersent_df['is_duplicate'], infersent_df['pred_res(spear_corr)']))
```

```
Accuary for InferSent embeddings using cosine similarity-  0.682
Accuary for InferSent embeddings using spearman's correlation-  0.654
```

10. Semantic Search using SBERT

In [ ]:

```
semser_df = train_df[:100]
semser_df.shape
```

Out[ ]:

```
(100, 6)
```

In [ ]:

```
q1 = semser_df.question1.tolist()
q2 = semser_df.question2.tolist()
```

In [ ]:

```
sentences = q1 + q2
sentence_embeddings = st_model.encode(sentences)
```

In [ ]:

```
print('Sample BERT embedding vector - length', len(sentence_embeddings[0]))
```

```
Sample BERT embedding vector - length 768
```

In [ ]:

```
query = semser_df['question1'][28]

queries = [query]
query_embeddings = st_model.encode(queries)
n = 4

print("Semantic Search Results")

for query, query_embedding in zip(queries, query_embeddings):
    distances = scipy.spatial.distance.cdist([query_embedding], sentence_embeddings, "cosine")[0]
    results = zip(range(len(distances)), distances)
    results = sorted(results, key=lambda x: x[1])
    print("Query:", query)
    print("\nTop 3 most similar sentences - ")
    for idx, distance in results[0:n]:
        print(sentences[idx].strip(), "(Cosine Score: %.4f)" % (1-distance))
```

```
Semantic Search Results
Query: What is best way to make money online?

Top 3 most similar sentences -
What is best way to make money online? (Cosine Score: 1.0000)
What is best way to ask for money online? (Cosine Score: 0.9583)
How can I make money through the Internet? (Cosine Score: 0.7752)
What are some different ways to make money online, excluding selling things? (Cosine Score: 0.7561)
```

In [ ]:

```

```