In [1]:
```
!wget --header="Host: s3-ap-southeast-1.amazonaws.com" --header="User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWe
```

```
--2019-11-05 07:47:10--  https://s3-ap-southeast-1.amazonaws.com/internshala-uploads/chat-uploads/5dc121c2da7871572938178-694
9877.zip (https://s3-ap-southeast-1.amazonaws.com/internshala-uploads/chat-uploads/5dc121c2da7871572938178-6949877.zip)
Resolving s3-ap-southeast-1.amazonaws.com (s3-ap-southeast-1.amazonaws.com)... 52.219.128.102
Connecting to s3-ap-southeast-1.amazonaws.com (s3-ap-southeast-1.amazonaws.com)|52.219.128.102|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6949877 (6.6M) [application/zip]
Saving to: ‘Text_Similarity_Dataset.zip’

Text_Similarity_Dat 100%[===================>]   6.63M  5.31MB/s    in 1.2s

2019-11-05 07:47:12 (5.31 MB/s) - ‘Text_Similarity_Dataset.zip’ saved [6949877/6949877]
```

In [0]:
```
import zipfile
zip_ref = zipfile.ZipFile("Text_Similarity_Dataset.zip", 'r')
zip_ref.extractall("/content")
zip_ref.close()
```

In [0]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
data = pd.read_csv("Text_Similarity_Dataset.csv")
data.head()
```

Out[2]:

	Unique_ID	text1	text2
0	0	savvy searchers fail to spot ads internet sear...	newcastle 2-1 bolton kieron dyer smashed home ...
1	1	millions to miss out on the net by 2025 40% o...	nasdaq planning $100m share sale the owner of ...
2	2	young debut cut short by ginepri fifteen-year-...	ruddock backs yapp s credentials wales coach m...
3	3	diageo to buy us wine firm diageo the world s...	mci shares climb on takeover bid shares in us ...
4	4	be careful how you code a new european directi...	media gadgets get moving pocket-sized devices ...

In [0]:
```
!pip3 install --quiet "tensorflow>=1.7"
!pip3 install --quiet tensorflow-hub
```

In [0]:
```
import tensorflow as tf
import tensorflow_hub as hub
import matplotlib.pyplot as plt
module_url = "https://tfhub.dev/google/universal-sentence-encoder/2" #@param ["https://tfhub.dev/google/universal-sentence-enc
```

In [0]:
```
# Import the Universal Sentence Encoder's TF Hub module
embed = hub.Module(module_url)
```

In [0]:
```
sts_input1 = tf.placeholder(tf.string, shape=(None))
sts_input2 = tf.placeholder(tf.string, shape=(None))

# For evaluation we use exactly normalized rather than
# approximately normalized.
sts_encode1 = tf.nn.l2_normalize(embed(sts_input1), axis=1)
sts_encode2 = tf.nn.l2_normalize(embed(sts_input2), axis=1)
cosine_similarities = tf.reduce_sum(tf.multiply(sts_encode1, sts_encode2), axis=1)
clip_cosine_similarities = tf.clip_by_value(cosine_similarities, -1.0, 1.0)
sim_scores = 1.0 - tf.acos(clip_cosine_similarities)
```

In [0]:
```
text_a = data['text1'].tolist()
text_b = data['text2'].tolist()

def run_sts_benchmark(session):
  """Returns the similarity scores"""
  emba, embb, scores = session.run(
      [sts_encode1, sts_encode2, sim_scores],
      feed_dict={
          sts_input1: text_a,
          sts_input2: text_b
      })
  return scores


with tf.Session() as session:
  session.run(tf.global_variables_initializer())
  session.run(tf.tables_initializer())
  scores = run_sts_benchmark(session)
```

In [13]:
```
len(scores)
```

Out[13]: 4023

In [0]:
```
data["Similarity_Score"] = scores
```

In [15]:
```
data.head()
```

Out[15]:

	Unique_ID	text1	text2	Similarity_Score
0	0	savvy searchers fail to spot ads internet sear...	newcastle 2-1 bolton kieron dyer smashed home ...	-0.463501
1	1	millions to miss out on the net by 2025 40% o...	nasdaq planning $100m share sale the owner of ...	-0.194790
2	2	young debut cut short by ginepri fifteen-year-...	ruddock backs yapp s credentials wales coach m...	-0.002321
3	3	diageo to buy us wine firm diageo the world s...	mci shares climb on takeover bid shares in us ...	-0.130618
4	4	be careful how you code a new european directi...	media gadgets get moving pocket-sized devices ...	-0.147261

In [21]:
```
data[data["Similarity_Score"]>=0.45]
```

Out[21]:

	Unique_ID	text1	text2	Similarity_Score
1613	1613	economy strong in election year uk businesse...	australia rates at four year high australia is...	0.504565
2284	2284	dvd copy protection strengthened dvds will be ...	dvd copy protection strengthened dvds will be ...	0.854296
2731	2731	brown shrugs off economy fears gordon brown is...	brown names 16 march for budget chancellor gor...	0.452971
3013	3013	call for kenteris to be cleared kostas kenteri...	iaaf to rule on greek sprint pair greek sprint...	0.453533
3056	3056	howard dismisses tory tax fears michael howard...	defection timed to hit tax pledge with impecca...	0.607239
3260	3260	benitez delight after crucial win liverpool ma...	wenger dejected as arsenal slump arsenal manag...	0.485157
3289	3289	ruddock backs yapp s credentials wales coach m...	white prepared for battle tough-scrummaging pr...	0.471721
3403	3403	holmes starts 2005 with gb events kelly holmes...	holmes starts 2005 with gb events kelly holmes...	1.000000
3544	3544	bank holds interest rate at 4.75% the bank of ...	economy strong in election year uk businesse...	0.540780
3859	3859	troubled marsh under sec scrutiny the us stock...	marsh executive in guilty plea an executive at...	0.456028

In [0]:
```
def partition(x):
    if x < 0.45:
        return 0
    return 1
```

In [0]:
```
data["Label"] = data["Similarity_Score"].apply(partition)
```

In [24]:
```
data.head()
```

Out[24]:

	Unique_ID	text1	text2	Similarity_Score	Label
0	0	savvy searchers fail to spot ads internet sear...	newcastle 2-1 bolton kieron dyer smashed home ...	-0.463501	0
1	1	millions to miss out on the net by 2025 40% o...	nasdaq planning $100m share sale the owner of ...	-0.194790	0
2	2	young debut cut short by ginepri fifteen-year-...	ruddock backs yapp s credentials wales coach m...	-0.002321	0
3	3	diageo to buy us wine firm diageo the world s...	mci shares climb on takeover bid shares in us ...	-0.130618	0
4	4	be careful how you code a new european directi...	media gadgets get moving pocket-sized devices ...	-0.147261	0

In [25]:
```
data["Label"].value_counts()
```

Out[25]:
```
0    4013
1      10
Name: Label, dtype: int64
```

In [0]:
```

```