```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import sqlite3
import csv
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import re
import os
from sqlalchemy import create_engine # database connection
import datetime as dt
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.multiclass import OneVsRestClassifier
from sklearn.linear_model import SGDClassifier
from sklearn import metrics
from sklearn.metrics import f1_score,precision_score,recall_score
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from datetime import datetime
from tqdm import tqdm
from nltk.corpus import stopwords
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics import pairwise_distances
```

```
data_main_clean_v4=pd.read_pickle('data_main_clean_v4.pickle')
```

```
! pip install transformers
```

```
Collecting transformers
  Downloading https://files.pythonhosted.org/packages/27/3c/91ed8f5c4e7ef3227b4119200fc0ed4b4fd965b1f0172021c25701087825/transformers-3.0.2-py3-none-any.whl (769kB)
     |████████████████████████████████| 778kB 2.6MB/s
Requirement already satisfied: numpy in /usr/local/lib/python3.6/dist-packages (from transformers) (1.18.5)
Requirement already satisfied: dataclasses; python_version < "3.7" in /usr/local/lib/python3.6/dist-packages (from transformers) (0.7)
Collecting tokenizers==0.8.1.rc1
  Downloading https://files.pythonhosted.org/packages/40/d0/30d5f8d221a0ed981a186c8eb986ce1c94e3a6e87f994eae9f4aa5250217/tokenizers-0.8.1rc1-cp36-cp36m-manylinux1_x86
     |████████████████████████████████| 3.0MB 12.7MB/s
Requirement already satisfied: requests in /usr/local/lib/python3.6/dist-packages (from transformers) (2.23.0)
Collecting sentencepiece!=0.1.92
  Downloading https://files.pythonhosted.org/packages/d4/a4/d0a884c4300004a78cca907a6ff9a5e9fe4f090f5d95ab341c53d28cbc58/sentencepiece-0.1.91-cp36-cp36m-manylinux1_x8
     |████████████████████████████████| 1.1MB 36.0MB/s
Collecting sacremoses
  Downloading https://files.pythonhosted.org/packages/7d/34/09d19aff26edcc8eb2a01bed8e98f13a1537005d31e95233fd48216eed10/sacremoses-0.0.43.tar.gz (883kB)
     |████████████████████████████████| 890kB 36.4MB/s
Requirement already satisfied: filelock in /usr/local/lib/python3.6/dist-packages (from transformers) (3.0.12)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.6/dist-packages (from transformers) (4.41.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.6/dist-packages (from transformers) (20.4)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.6/dist-packages (from transformers) (2019.12.20)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.6/dist-packages (from requests->transformers) (1.24.3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.6/dist-packages (from requests->transformers) (2020.6.20)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.6/dist-packages (from requests->transformers) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.6/dist-packages (from requests->transformers) (3.0.4)
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from sacremoses->transformers) (1.15.0)
Requirement already satisfied: click in /usr/local/lib/python3.6/dist-packages (from sacremoses->transformers) (7.1.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.6/dist-packages (from sacremoses->transformers) (0.16.0)
Requirement already satisfied: pyparsing>=2.0.2 in /usr/local/lib/python3.6/dist-packages (from packaging->transformers) (2.4.7)
Building wheels for collected packages: sacremoses
  Building wheel for sacremoses (setup.py) ... done
  Created wheel for sacremoses: filename=sacremoses-0.0.43-cp36-none-any.whl size=893260 sha256=4a97e5e4901f1f42b12d038fa0676f1020afcd9a9d0527904489f1abe68f1063
  Stored in directory: /root/.cache/pip/wheels/29/3c/fd/7ce5c3f0666dab31a50123635e6fb5e19ceb42ce38d4e58f45
```

BERT Hugging Face

```
from transformers import AutoTokenizer, pipeline, TFDistilBertModel
from tqdm import tqdm
```

```
title=data_main_clean_v4['Cleaned_Title'].values
```

```
title_list=list(title)
```

```
#Using the pretained 'distilbert-base-uncased' model and saving in model
model = TFDistilBertModel.from_pretrained('distilbert-base-uncased')

#Tokenizing the pretrained model and storing in tokenizer
tokenizer = AutoTokenizer.from_pretrained('distilbert-base-uncased')

#Creating a pipeline from the defined model and tokenizer for our data
pipe = pipeline('feature-extraction', model=model, tokenizer=tokenizer)
```

```
Downloading: 100%  442/442 [00:21<00:00, 20.3B/s]

Downloading: 100%  363M/363M [00:08<00:00, 45.0MB/s]

Some weights of the model checkpoint at distilbert-base-uncased were not used when initializing TFDistilBertModel: ['vocab_projector', 'vocab_layer_norm', 'vocab_tran
- This IS expected if you are initializing TFDistilBertModel from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a
- This IS NOT expected if you are initializing TFDistilBertModel from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenc
All the weights of TFDistilBertModel were initialized from the model checkpoint at distilbert-base-uncased.
If your task is similar to the task the model of the ckeckpoint was trained on, you can already use TFDistilBertModel for predictions without further training.
Downloading: 100%  232k/232k [00:00<00:00, 296kB/s]
```

```
#Using the defined pipeline to convert our data into features and considering one 768 dimension vecotor for each datapoint
features = pipe(title_list)
features = np.squeeze(features)
features = features[:,0,:]
```

```
#Converting list to numpy array
features1=np.array(features)
```

```
features1.shape
```

```
(1396270,768)
```

```
a1=features1
```

```
def Recomend(string):
  stopwords_1 = stopwords.words("english")
  a=string
  sent_1=a.lower().strip()
  sent_1 = re.sub(r"won\'t", "will not", sent_1)
  sent_1 = re.sub(r"can\'t", "can not", sent_1)
  sent_1 = re.sub(r"n\'t", " not", sent_1)
  sent_1 = re.sub(r"\'re", " are", sent_1)
  sent_1 = re.sub(r"\'s", " is", sent_1)
  sent_1 = re.sub(r"\'d", " would", sent_1)
  sent_1 = re.sub(r"\'ll", " will", sent_1)
  sent_1 = re.sub(r"\'t", " not", sent_1)
  sent_1 = re.sub(r"\'ve", " have", sent_1)
  sent_1 = re.sub(r"\'m", " am", sent_1)
  sent_1 = re.sub('[^A-Za-z0-9-+]+', ' ', sent_1)
  sent_1 = ' '.join(e for e in sent_1.split() if e not in stopwords_1)

  sent_1=sent_1.lower().strip()
  print('QUERY ENTERED BY THE USER')
  print(sent_1)
  print('\n')

  query=a1[0]
  distance = pairwise_distances(a1, query.reshape(1,-1),metric='cosine')
  indices = np.argsort(distance.flatten())[0:10]
  pdists = np.sort(distance.flatten())[0:10]

  print('RECOMENDED SIMILAR QUESTIONS')
  g=0
  for i in indices:
    g=g+1
    print(g ,'th question','"',data_main_clean_v4['Cleaned_Title'][i],'"')
    print(g ,'th question distance is ',round((float(distance[i])),4))
    print('\n')
```

```
import time
start_time = time.time()
Recomend('implementing boundary value analysis software testing c++ program')
print('TIME TAKEN TO FETCH RESULTS')
print(time.time()-start_time,'seconds')
```

```
QUERY ENTERED BY THE USER
implementing boundary value analysis software testing c++ program

RECOMENDED SIMILAR QUESTIONS
1 th question ' implementing boundary value analysis software testing c++ program '
1 th question distance is 0.0

2 th question ' implementing data structures algorithms c++ '
2 th question distance is 0.4653

3 th question ' boundary value analysis c++ cppunit '
3 th question distance is 0.472

4 th question ' code metrics analysis unmanaged c++ code '
4 th question distance is 0.4922

5 th question ' c++ code profiling analysis mac mpi '
5 th question distance is 0.5023

6 th question ' calculating critical path dag c++ '
6 th question distance is 0.5064

7 th question ' methods implementing using graphs nodes c++ '
7 th question distance is 0.5136

8 th question ' complex data structures embedding extending python c++'
8 th question distance is 0.5187

9 th question ' obfuscate c++ variables functions '
9 th question distance is 0.5244

10 th question ' c++ static global non-pod theory practice '
10 th question distance is 0.5245

TIME TAKEN TO FETCH RESULTS
4.870607376098633 seconds
```

/