UE19CS332 : Algorithms for Web and Information Retrieval
ASSIGNMENT – 2

Problem definition:

- **Build a search engine for any 3 corpora of your choice,**
- **Your Code should be able to: Search for the terms in the query – Create Postings list**
- **Fill the Inverted Index**
- **Retrieve the data from the dictionary – Query response time.**

TEAM MEMBERS :

NAME	SRN	SECTION
Priya Mohata	PES2UG19CS301	E
R Sharmila	PES2UG19CS309	E
Ritik	PES2UG19CS332	E

CORPUS – 2 : *TWITTER SENTIMENT ANALYSIS CORPUS*

DATASET LOCATION :
https://drive.google.com/file/d/1BfVUFqMUrA5enM0Fb7SJi3SMQk2YWRoK/view?usp=sharing

NOTEBOOK NAME : A3_P2_TEAM-20.ipynb

Link to the notebook :
https://colab.research.google.com/drive/1X2T4E6vhMbka1JOlcU4yR-_r7l-u1rhB?usp=sharing

STEPS :

Importing all libraries

```
# Assignment - 2

# PRIYA MOHATA - PES2UG19CS301
# R SHARMILA  - PES2UG19CS309
# RITIK - PES2UG19CS332

# Twitter Sentiment Analysis

import pandas as pd
import numpy as np
import nltk
from nltk.tokenize import word_tokenize
from nltk.tokenize import sent_tokenize
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

```
[2] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Case folding

```
[3]  # CASE FOLDING

     train_data=pd.read_csv('/content/drive/MyDrive/DATASETS-AIWIR/tweet_sentiment.csv')
     train_data["text"] = train_data["text"].str.lower()
     train_data.head()
```

	textID	text	sentiment_main	sentiment
0	p1000000000	i know i was listenin to bad habit earlier a...	empty	neutral
1	p1000000001	i should be sleep, but im not! thinking about ...	sadness	negative
2	2dfbe0b7fb	hmmm. http://www.djhero.com/ is down	worry	negative
3	6d846d7d50	i`m sorry at least it`s friday?	sadness	neutral
4	p1000000002	the storm is here and the electricity is gone	sadness	negative

```
[4]  train_data.shape
```

```
(12454, 4)
```

```
train_data.columns
```

```
Index(['textID', 'text', 'sentiment_main', 'sentiment'], dtype='object')
```

Sentence Tokenization

```
[5]  # SENTENCE TOKENIZATION
     df=train_data['text']
     l=list()
     for line in df:
       token=sent_tokenize(line)
       l.append(token)
```

```
[6]  df=train_data['text']
     train_data['sent_token']=l
```

```
[7]  train_data.head()
```

	textID	text	sentiment_main	sentiment	sent_token
0	p1000000000	i know i was listenin to bad habit earlier a...	empty	neutral	[i know i was listenin to bad habit earlier ...
1	p1000000001	i should be sleep, but im not! thinking about ...	sadness	negative	[i should be sleep, but im not!, thinking abou...
2	2dfbe0b7fb	hmmm. http://www.djhero.com/ is down	worry	negative	[hmmm., http://www.djhero.com/ is down]
3	6d846d7d50	i`m sorry at least it`s friday?	sadness	neutral	[i`m sorry at least it`s friday?]
4	p1000000002	the storm is here and the electricity is gone	sadness	negative	[the storm is here and the electricity is gone]

Word Tokenization

```
[8]  # WORD TOKENIZATION
     df=train_data['text']
     l1=list()
     for line in df:
       tokens=word_tokenize(line)
       l1.append(tokens)
```

```
[9]  df=train_data['text']
     train_data['word_token']=l1
```

```
[10] train_data.head()
```

	textID	text	sentiment_main	sentiment	sent_token	word_token
0	p1000000000	i know i was listenin to bad habit earlier a...	empty	neutral	[i know i was listenin to bad habit earlier ...	[i, know, i, was, listenin, to, bad, habit, ea...
1	p1000000001	i should be sleep, but im not! thinking about ...	sadness	negative	[i should be sleep, but im not!, thinking abou...	[i, should, be, sleep, ,, but, im, not, !, thi...
2	2dfbe0b7fb	hmmm. http://www.djhero.com/ is down	worry	negative	[hmmm., http://www.djhero.com/ is down]	[hmmm, ., http, :, //www.djhero.com/, is, down]
3	6d846d7d50	i`m sorry at least it`s friday?	sadness	neutral	[i`m sorry at least it`s friday?]	[i`m, sorry, at, least, it`s, friday, ?]
4	p1000000002	the storm is here and the electricity is gone	sadness	negative	[the storm is here and the electricity is gone]	[the, storm, is, here, and, the, electricity, ...

Stop Words Removal

```
[11] # STOP WORDS REMOVAL
     import nltk
     nltk.download('stopwords')
     from nltk.corpus import stopwords
     stoplist= stopwords.words('english')

     [nltk_data] Downloading package stopwords to /root/nltk_data...
     [nltk_data]   Unzipping corpora/stopwords.zip.

[12] stoplist=set(stoplist)
     l2=list()
     for i in l1:
       output = [w for w in i if not w in stoplist]
       l2.append(output)
     train_data['stop_words_removed']=l2

[13] train_data.head()
```

	textID	text	sentiment_main	sentiment	sent_token	word_token	stop_words_removed
0	p1000000000	i know i was listenin to bad habit earlier a...	empty	neutral	[i know i was listenin to bad habit earlier ...	[i, know, i, was, listenin, to, bad, habit, ea...	[know, listenin, bad, habit, earlier, started,...
1	p1000000001	i should be sleep, but im not! thinking about ...	sadness	negative	[i should be sleep, but im not!, thinking abou...	[i, should, be, sleep, ,, but, im, not, !, thi...	[sleep, ,, im, !, thinking, old, friend, want,...
2	2dfbe0b7fb	hmmm. http://www.djhero.com/ is down	worry	negative	[hmmm., http://www.djhero.com/ is down]	[hmmm, ., http, :, //www.djhero.com/, is, down]	[hmmm, ., http, :, //www.djhero.com/]
3	6d846d7d50	i`m sorry at least it`s friday?	sadness	neutral	[i`m sorry at least it`s friday?]	[i`m, sorry, at, least, it`s, friday, ?]	[i`m, sorry, least, it`s, friday, ?]
4	p1000000002	the storm is here and the electricity is gone	sadness	negative	[the storm is here and the electricity is gone]	[the, storm, is, here, and, the, electricity, ...	[storm, electricity, gone]

Stemming

```
[18] # STEMMING
     from nltk.stem import WordNetLemmatizer
     from nltk.stem import PorterStemmer

     # Stemming :
     final_train_stem_list=[]
     ps = PorterStemmer()
     for line in train_data['stop_words_removed']:
       Stem_words=[]
       for i in line:
         rootWord = ps.stem(i)
         Stem_words.append(rootWord)
       Stem_words= [word for word in Stem_words if word.isalnum()]
       final_train_stem_list.append(Stem_words)

     print(final_train_stem_list[0:5])

     [['know', 'listenin', 'bad', 'habit', 'earlier', 'start', 'freakin', 'part'], ['sleep', 'im', 'think', 'old', 'friend', 'want', 'marri', 'want', '2', 'scand

[ ] train_data['stemmed_words']=final_train_stem_list
     train_data.head()
```

	textID	text	sentiment_main	sentiment	sent_token	word_token	stop_words_removed	stemmed_words
0	p1000000000	i know i was listenin to bad habit earlier a...	empty	neutral	[i know i was listenin to bad habit earlier ...	[i, know, i, was, listenin, to, bad, habit, ea...	[know, listenin, bad, habit, earlier, started,...	[know, listenin, bad, habit, earlier, start, f...
1	p1000000001	i should be sleep, but im not! thinking about ...	sadness	negative	[i should be sleep, but im not!, thinking abou...	[i, should, be, sleep, ,, but, im, not, !, thi...	[sleep, ,, im, !, thinking, old, friend, want,...	[sleep, im, think, old, friend, want, marri, w...
2	2dfbe0b7fb	hmmm. http://www.djhero.com/ is down	worry	negative	[hmmm., http://www.djhero.com/ is down]	[hmmm, ., http, :, //www.djhero.com/, is, down]	[hmmm, ., http, :, //www.djhero.com/]	[hmmm, http]
3	6d846d7d50	i`m sorry at least it`s friday?	sadness	neutral	[i`m sorry at least it`s friday?]	[i`m, sorry, at, least, it`s, friday, ?]	[i`m, sorry, least, it`s, friday, ?]	[sorri, least, friday]
4	p1000000002	the storm is here and the electricity is gone	sadness	negative	[the storm is here and the electricity is gone]	[the, storm, is, here, and, the, electricity, ...	[storm, electricity, gone]	[storm, electr, gone]

Lemmatization

```
[16] # LEMMATIZATION

     import nltk
     nltk.download('wordnet')
     final_train_lemma_word = []
     wordnet_lemmatizer = WordNetLemmatizer()
     for line in train_data['stop_words_removed']:
       lemma_word=[];
       for w in line:
           word1 = wordnet_lemmatizer.lemmatize(w, pos = "n")
           word2 = wordnet_lemmatizer.lemmatize(word1, pos = "v")
           word3 = wordnet_lemmatizer.lemmatize(word2, pos = ("a"))
           lemma_word.append(word1)
       lemma_word= [word for word in lemma_word if word.isalnum()]
       final_train_lemma_word.append(lemma_word)

     print(final_train_lemma_word[0:5])
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
[['know', 'listenin', 'bad', 'habit', 'earlier', 'started', 'freakin', 'part'], ['sleep', 'im', 'thinking', 'old', 'friend', 'want', 'married', 'want', '2',
```

```
[17] train_data['lemmatized_words']=final_train_lemma_word
     train_data.head()
```

[17]

	textID	text	sentiment_main	sentiment	sent_token	word_token	stop_words_removed	stemmed_words	lemmatized_words
0	p1000000000	i know i was listenin to bad habit earlier a...	empty	neutral	[i know i was listenin to bad habit earlier ...	[i, know, i, was, listenin, to, bad, habit, ea...	[know, listenin, bad, habit, earlier, started,...	[know, listenin, bad, habit, earlier, start, f...	[know, listenin, bad, habit, earlier, started,...
1	p1000000001	i should be sleep, but im not! thinking about ...	sadness	negative	[i should be sleep, but im not!, thinking abou...	[i, should, be, sleep, ,, but, im, not, !, thi...	[sleep, ,, im, !, thinking, old, friend, want,...	[sleep, im, think, old, friend, want, marri, w...	[sleep, im, thinking, old, friend, want, marri...
2	2dfbe0b7fb	hmmm. http://www.djhero.com/ is down	worry	negative	[hmmm., http://www.djhero.com/ is down]	[hmmm, ., http, :, //www.djhero.com/, is, down]	[hmmm, ., http, :, //www.djhero.com/]	[hmmm, http]	[hmmm, http]
3	6d846d7d50	i`m sorry at least it`s friday?	sadness	neutral	[i`m sorry at least it`s friday?]	[i`m, sorry, at, least, it`s, friday, ?]	[i`m, sorry, least, it`s, friday, ?]	[sorri, least, friday]	[sorry, least, friday]
4	p1000000002	the storm is here and the electricity is gone	sadness	negative	[the storm is here and the electricity is gone]	[the, storm, is, here, and, the, electricity, ...	[storm, electricity, gone]	[storm, electr, gone]	[storm, electricity, gone]

Building Inverted Index

```
# GENERATING INVERTED INDEX
def generate_inverted_index(data: list):
    inv_idx_dict = {}
    for index, doc_text in enumerate(data):
        for word in doc_text:
            if word not in inv_idx_dict.keys():
                inv_idx_dict[word] = [index]
            elif index not in inv_idx_dict[word]:
                inv_idx_dict[word].append(index)
    return inv_idx_dict

final_train=generate_inverted_index(final_train_stem_list)

j=0
for i in final_train:
  print(i,":",final_train[i])
  if j==20:
    break;
  j=j+1
```

```
know : [0, 6, 64, 104, 114, 134, 153, 166, 171, 210, 217, 243, 278, 292, 295, 325, 344, 449, 465, 467, 487, 492, 511, 550, 566, 593, 595, 667, 680, 726, 754,
listenin : [0, 5864, 9881, 11009]
bad : [0, 16, 52, 121, 130, 158, 232, 257, 278, 324, 415, 424, 479, 480, 625, 696, 722, 732, 733, 752, 822, 829, 855, 903, 936, 943, 971, 1041, 1097, 1112,
habit : [0, 6345, 8659, 10521]
earlier : [0, 633, 1372, 1804, 2053, 2892, 3031, 3765, 3943, 4064, 4456, 6123, 7910, 8572, 8700, 11104, 12030]
start : [0, 61, 122, 144, 163, 263, 326, 343, 387, 582, 711, 861, 885, 1007, 1040, 1081, 1145, 1297, 1313, 1415, 1531, 1581, 1586, 1590, 1647, 1732, 1958,
freakin : [0, 794, 1325, 2103, 2222, 2374, 3293, 3714, 3824, 6355, 6823, 9057, 11436]
part : [0, 24, 436, 2544, 3304, 3410, 3961, 4068, 4187, 6209, 6277, 6626, 6924, 7195, 7908, 8768, 8872, 9347, 9897, 9900, 10060, 10304, 11015, 11259, 11275,
sleep : [1, 24, 116, 192, 196, 214, 250, 293, 301, 305, 312, 340, 352, 373, 392, 412, 481, 491, 507, 519, 525, 565, 637, 638, 711, 720, 723, 729, 733, 737,
im : [1, 35, 39, 54, 76, 77, 144, 217, 265, 269, 290, 339, 385, 408, 441, 478, 481, 491, 493, 497, 523, 551, 564, 584, 591, 597, 601, 608, 636, 637, 736,
think : [1, 7, 11, 21, 34, 38, 44, 48, 70, 74, 109, 115, 171, 172, 187, 189, 197, 217, 267, 317, 341, 343, 363, 370, 382, 404, 457, 462, 464, 474, 483, 495,
old : [1, 40, 155, 366, 499, 575, 677, 681, 755, 913, 1549, 1772, 1804, 1817, 1953, 2390, 2818, 2900, 2971, 3067, 3079, 3354, 3375, 3669, 4187, 4254, 4511,
friend : [1, 52, 72, 134, 193, 217, 266, 357, 403, 405, 445, 512, 516, 576, 620, 732, 849, 922, 1066, 1099, 1152, 1162, 1199, 1456, 1465, 1486, 1563, 1581,
want : [1, 29, 48, 61, 64, 91, 124, 150, 153, 226, 281, 292, 293, 298, 306, 312, 326, 335, 345, 346, 355, 410, 425, 440, 444, 450, 454, 471, 474, 479, 493,
marri : [1, 2061, 3125, 3447, 4551, 4935, 7970, 9224, 11085, 11132]
2 : [1, 69, 77, 144, 313, 341, 343, 349, 352, 389, 478, 503, 534, 622, 632, 658, 712, 719, 840, 879, 899, 943, 944, 961, 987, 1022, 1035, 1137, 1202, 1298,
scandal : [1, 10223]
```

```
[19] hmmm : [2, 1921, 2267, 4416, 5510, 6122, 6436, 6679, 7131, 7203, 8351, 8583, 8731, 8891, 10870, 10950, 11560, 11574, 11713, 12083, 12426]
     http : [2, 5, 33, 85, 191, 204, 207, 212, 230, 242, 272, 329, 330, 368, 376, 380, 389, 416, 437, 458, 520, 541, 556, 580, 591, 593, 600, 622, 663, 668, 677,
     sorri : [3, 103, 121, 128, 175, 181, 218, 243, 267, 302, 324, 351, 428, 441, 489, 630, 736, 815, 845, 979, 990, 1033, 1158, 1174, 1207, 1327, 1374, 1421, 15
     least : [3, 312, 571, 870, 961, 1023, 1318, 1345, 1506, 1705, 1714, 1818, 1859, 1861, 2064, 2213, 2384, 2406, 2707, 2793, 2877, 3156, 3321, 3601, 3737, 3744,
```

```
# Sorting index based on terms
final_train1=sorted(final_train.items())
final_train1
```

```
  11487]),
 ('awwwww', [4286, 4905, 5200, 7554, 9687, 9945]),
 ('awwwwww', [1547, 4118]),
 ('awwwwwww', [1343, 12353]),
 ('axayi', [6303]),
 ('axe', [11112]),
 ('aye', [4005, 6745, 9323, 12310]),
 ('ayi', [3241]),
 ('ayshea', [6949]),
 ('az', [1035, 2777]),
 ('azeroth', [3786]),
 ('aztec', [9218]),
 ('aztex', [3617]),
 ('azz', [6181, 10098]),
 ('azza', [10967]),
 ('b',
  [217,
   432,
   485,
   588,
   1071,
   1698,
   1766,
   2166,
   2716,
   2924,
   3692,
   3755,
   4584,
   4774,
```

Adding the module for timing the query response

```
[21] !pip install ipython-autotime
     %load_ext autotime

     Collecting ipython-autotime
       Downloading ipython_autotime-0.3.1-py2.py3-none-any.whl (6.8 kB)
     Requirement already satisfied: ipython in /usr/local/lib/python3.7/dist-packages (from ipython-autotime) (5.5.0)
     Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (57.4.0)
     Requirement already satisfied: pickleshare in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (0.7.5)
     Requirement already satisfied: pexpect in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (4.8.0)
     Requirement already satisfied: decorator in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (4.4.2)
     Requirement already satisfied: prompt-toolkit<2.0.0,>=1.0.4 in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (1.0.18)
     Requirement already satisfied: pygments in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (2.6.1)
     Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (5.1.1)
     Requirement already satisfied: simplegeneric>0.8 in /usr/local/lib/python3.7/dist-packages (from ipython->ipython-autotime) (0.8.1)
     Requirement already satisfied: wcwidth in /usr/local/lib/python3.7/dist-packages (from prompt-toolkit<2.0.0,>=1.0.4->ipython->ipython-autotime) (0.2.5)
     Requirement already satisfied: six>=1.9.0 in /usr/local/lib/python3.7/dist-packages (from prompt-toolkit<2.0.0,>=1.0.4->ipython->ipython-autotime) (1.15.0)
     Requirement already satisfied: ptyprocess>=0.5 in /usr/local/lib/python3.7/dist-packages (from pexpect->ipython->ipython-autotime) (0.7.0)
     Installing collected packages: ipython-autotime
     Successfully installed ipython-autotime-0.3.1
     time: 136 µs (started: 2022-03-27 11:24:18 +00:00)
```

Building Positional Index

```
# GENERATING POSITIONAL INDEX
pos_index = {}
file_map = {}
def generate_positional_index(data:list):
  fileno=0
  lineno=-1
  for line in data:
    lineno+=1;
    for pos, term in enumerate(line):
      if term in pos_index:
          pos_index[term][0] = pos_index[term][0] + 1
          if fileno in pos_index[term][1]:
            pos_index[term][1][lineno].append(pos)
          else:
            pos_index[term][1][lineno] = [pos]
      else:
          pos_index[term] = []
          pos_index[term].append(1)
          pos_index[term].append({})
          pos_index[term][1][lineno] = [pos]
      fileno += 1
  return pos_index

final=generate_positional_index(final_train_stem_list)
count=0
for i in final:
  count=count+1;
  if count<=20:
    print(i,final[i])
  else:
    break;
```

```
[30] know [430, {0: [0], 6: [2], 64: [8], 104: [0], 114: [1], 134: [8], 153: [8], 166: [0], 171: [4], 210: [3], 217: [9], 243: [5], 278: [13], 292: [1], 295: [1],
     listenin [4, {0: [1], 5864: [3], 9881: [0], 11009: [2]}]
     bad [189, {0: [2], 16: [10], 52: [0], 121: [7], 130: [0], 158: [7], 232: [1], 257: [6], 278: [12], 324: [5], 415: [0], 424: [5], 479: [5], 480: [3], 625: [6
     habit [4, {0: [3], 6345: [2], 8659: [17], 10521: [8]}]
     earlier [17, {0: [4], 633: [6], 1372: [4], 1804: [6], 2053: [3], 2892: [1], 3031: [9], 3765: [3], 3943: [8], 4064: [4], 4456: [3], 6123: [4], 7910: [1], 857
     start [139, {0: [5], 61: [9], 122: [2], 144: [3], 163: [12], 263: [5], 326: [1], 343: [7], 387: [2], 582: [3], 711: [8], 861: [0], 885: [3], 1007: [4], 1040
     freakin [14, {0: [6], 794: [0], 1325: [4], 2103: [5], 2222: [7], 2374: [0], 3293: [1], 3714: [1], 3824: [1], 6355: [7], 6823: [5], 9057: [4], 11436: [2]}]
     part [30, {0: [7], 24: [6], 436: [8], 2544: [2], 3304: [6], 3410: [1], 3961: [10], 4068: [2], 4187: [15], 6209: [8], 6277: [2], 6626: [10], 6924: [2], 7195:
     sleep [191, {1: [0], 24: [1], 116: [3], 192: [1], 196: [1], 214: [0], 250: [4], 293: [7], 301: [0], 305: [2], 312: [4], 340: [7], 352: [6], 373: [4], 392: [2
     im [409, {1: [1], 35: [3], 39: [6], 54: [2], 76: [4], 77: [18], 144: [5], 217: [1], 265: [3], 269: [1], 290: [3], 339: [2], 385: [0], 408: [3], 441: [10], 4
     think [407, {1: [2], 7: [1], 11: [7], 21: [0], 34: [5], 38: [2], 44: [3], 48: [0], 70: [4], 74: [8], 109: [3], 115: [3], 171: [9], 172: [0], 187: [0], 189:
     old [90, {1: [3], 40: [17], 155: [5], 366: [10], 499: [5], 575: [3], 677: [5], 681: [1], 755: [1], 913: [1], 1549: [5], 1772: [6], 1804: [8], 1817: [4], 195
     friend [203, {1: [4], 52: [1], 72: [0], 134: [3], 193: [1], 217: [4], 266: [3], 357: [9], 403: [3], 405: [4], 445: [3], 512: [4], 516: [3], 576: [2], 620: [4
     want [396, {1: [7], 29: [0], 48: [1], 61: [8], 64: [0], 91: [5], 124: [0], 150: [0], 153: [9], 226: [0], 281: [5], 292: [5], 293: [0], 298: [3], 306: [3], 3
     marri [11, {1: [6], 2061: [0], 3125: [2], 3447: [3], 4551: [2], 4935: [11], 7970: [7], 9224: [3], 11085: [7], 11132: [1]}]
     2 [251, {1: [8], 69: [2], 77: [4], 144: [7], 313: [10], 341: [7], 343: [1], 349: [11], 352: [10], 389: [5], 478: [14], 503: [1], 534: [2], 622: [5], 632: [12
     scandal [2, {1: [9], 10223: [1]}]
     hmmm [23, {2: [0], 1921: [0], 2267: [0], 4416: [10], 5510: [0], 6122: [2], 6436: [0], 6679: [0], 7131: [0], 7203: [9], 8351: [0], 8583: [0], 8731: [2], 8891
     http [591, {2: [1], 5: [9], 33: [4], 85: [7], 191: [0], 204: [0], 207: [4], 212: [10], 230: [0], 242: [5], 272: [4], 329: [0], 330: [6], 368: [6], 376: [6],
     sorri [201, {3: [0], 103: [0], 121: [2], 128: [0], 175: [0], 181: [2], 218: [7], 243: [0], 267: [0], 302: [0], 324: [1], 351: [12], 428: [0], 441: [11], 489
     time: 297 ms (started: 2022-03-27 11:28:51 +00:00)
```

```
[31] # SORTING THE POSITIONAL INDEX BASED ON TERMS
     final_train1=sorted(final.items())
     final_train1

     [6, {4286: [0], 4905: [1], 5200: [0], 7554: [0], 9687: [0], 9945: [2]}]),
     ('awwwwww', [2, {1547: [0], 4118: [0]}]),
     ('awwwwwww', [2, {1343: [0], 12353: [9]}]),
     ('axayi', [1, {6303: [5]}]),
     ('axe', [1, {11112: [4]}]),
     ('aye', [4, {4005: [5], 6745: [3], 9323: [0], 12310: [0]}]),
     ('ayi', [1, {3241: [0]}]),
     ('ayshea', [1, {6949: [2]}]),
     ('az', [2, {1035: [2], 2777: [1]}]),
     ('azeroth', [1, {3786: [6]}]),
     ('aztec', [1, {9218: [0]}]),
     ('aztex', [1, {3617: [2]}]),
     ('azz', [2, {6181: [1], 10098: [4]}]),
     ('azza', [1, {10967: [1]}]),
     ('b',
      [45,
       {217: [8],
        432: [9],
        485: [1],
        588: [15],
        1071: [0],
        1698: [4],
        1766: [5],
        2166: [9],
```

Performing Boolean Queries

```
[22] # Boolean Query
     # AND
     def and_query(l1, l2):
         p1 = 0
         p2 = 0
         result = list()
         while p1 < len(l1) and p2 < len(l2):
             if l1[p1] == l2[p2]:
                 result.append(l1[p1])
                 p1 += 1
                 p2 += 1
             elif l1[p1] > l2[p2]:
                 p2 += 1
             else:
                 p1 += 1
         return result

     time: 6.52 ms (started: 2022-03-27 11:25:03 +00:00)
```

```
def or_query(l1,l2):
  result=list();
  p1=0
  p2=0
  while p1 < len(l1) and p2 < len(l2):
    if l1[p1] == l2[p2]:
      result.append(l1[p1])
      p1 += 1
      p2 += 1
    elif l1[p1] > l2[p2]:
      result.append(l2[p2])
      p2 += 1
    else:
      result.append(l1[p1])
      p1 += 1
  while(p1 < len(l1)):
        result.append(l1[p1])
        p1 += 1
  while p2 < len(l2):
        result.append(l2[p2])
        p2 += 1
  return result

time: 15.3 ms (started: 2022-03-27 11:25:08 +00:00)
```

```
# PERFORMING THE BOOLEAN QUERY
print("Enter the first input word : ")
input1=input()
print("Enter the second input word : ")
input2=input()
```

```
Enter the first input word :
bad
Enter the second input word :
start
time: 7.11 s (started: 2022-03-27 11:26:03 +00:00)
```

```
[26] l1=final_train[input1]
     l2=final_train[input2]
     print("posting list for",input1 ,l1)
     print("posting list for",input2,l2)
     print("Resultant list: ",and_query(l1,l2))
```

```
posting list for bad [0, 16, 52, 121, 130, 158, 232, 257, 278, 324, 415, 424, 479, 480, 625, 696, 722, 732, 733, 752, 822, 829, 855, 903, 936, 943, 971, 104
posting list for start [0, 61, 122, 144, 163, 263, 326, 343, 387, 582, 711, 861, 885, 1007, 1040, 1081, 1145, 1297, 1313, 1415, 1531, 1581, 1586, 1590, 1647,
Resultant list:  [0, 1531, 1581, 7320]
time: 7.71 ms (started: 2022-03-27 11:26:37 +00:00)
```

```
[27] print("Resultant list: ",or_query(l1,l2))
     print("Length of posting list for",input1 ,len(l1))
     print("Length of posting list for",input2,len(l2))
     print("Length of and list: ",len(and_query(l1,l2)))
     print("Resultant list :",or_query(l1,l2))
     print("Length of the OR list:",len(or_query(l1,l2)))
```

```
[27] Resultant list:  [0, 16, 52, 61, 121, 122, 130, 144, 158, 163, 232, 257, 263, 278, 324, 326, 343, 387, 415, 424, 479, 480, 582, 625, 696, 711, 722, 732, 733,
     Length of posting list for bad 186
     Length of posting list for start 139
     Length of and list:  4
     Resultant list : [0, 16, 52, 61, 121, 122, 130, 144, 158, 163, 232, 257, 263, 278, 324, 326, 343, 387, 415, 424, 479, 480, 582, 625, 696, 711, 722, 732, 733,
     Length of the OR list: 321
     time: 16.3 ms (started: 2022-03-27 11:26:43 +00:00)
```

```
[28] print("Enter the third input word : ")
     input3=input()
     print("Enter the fourth input word : ")
     input4=input()
     l3=final_train[input3]
     l4=final_train[input4]
     resultant=or_query(or_query(and_query(l1,l4),l3),l2)
     print("Result:",resultant)
     print("Result length:",len(resultant))
```

```
Enter the third input word :
want
Enter the fourth input word :
sleep
Result: [0, 1, 29, 48, 61, 64, 91, 122, 124, 144, 150, 153, 163, 226, 263, 281, 292, 293, 298, 306, 312, 326, 335, 343, 345, 346, 355, 387, 410, 425, 440, 4
Result length: 523
time: 6.91 s (started: 2022-03-27 11:27:13 +00:00)
```

```
[29] resultant=and_query(or_query(l1,l3),or_query(l2,l4))
     print("Result:",resultant)
     print("Result length:",len(resultant))
```

```
Result: [0, 1, 61, 293, 312, 326, 519, 733, 1046, 1112, 1455, 1531, 1581, 2528, 3366, 4674, 4812, 4986, 7000, 7320, 10314, 10902]
Result length: 22
time: 7.39 ms (started: 2022-03-27 11:27:38 +00:00)
```

Performing Phrase Query on Inverted Index

```
[32] # PHRASE QUERY on Inverted Index :
     def phrase_query(phr):
         query=phr.split();
         for i in range(0,len(query)-1,2):
           result=and_query(final_train[query[i]],final_train[query[i+1]])
         print(result)
     print("Enter your query")
     q=input()
     phrase_query(q)
```

```
Enter your query
want sleep
[1, 293, 312, 519, 1046, 2528, 3366, 4812, 4986, 10314]
time: 17.1 s (started: 2022-03-27 11:31:56 +00:00)
```

```
print("Enter your query")
q=input()
phrase_query(q)
```

```
Enter your query
old friend want
[1, 2971, 3067, 3079, 11077]
time: 4.38 s (started: 2022-03-27 11:32:17 +00:00)
```

Performing Phrase Query on Positional Index

```
[34] # Phrase query on positional index :
     def fetch_list(d):
         l=list();
         d1=d[1];
         for i in d1:
           l.append(i)
         return l;
     def post_phrase_query(phr):
       query=phr.split()
       for i in range(0,len(query)-1,2):
         l1=fetch_list(final[query[i]])
         l2=fetch_list(final[query[i+1]])
         result=and_query(l1,l2)
       print(result)

     time: 7.98 ms (started: 2022-03-27 11:33:24 +00:00)

     print("Enter your query")
     q=input()
     post_phrase_query(q)

     Enter your query
     want sleep friend
     [1, 293, 312, 519, 1046, 2528, 3366, 4812, 4986, 10314]
     time: 5.12 s (started: 2022-03-27 11:33:32 +00:00)

[36] print("Enter your query")
     q=input()
     post_phrase_query(q)

     Enter your query
     think sleep old friend
     [1, 2971, 3067, 3079, 11077]
     time: 5.56 s (started: 2022-03-27 11:33:41 +00:00)
```