In [1]:

```
# web crawler
import multiprocessing
from bs4 import BeautifulSoup
from queue import Queue, Empty
from concurrent.futures import ThreadPoolExecutor
from urllib.parse import urljoin, urlparse
import requests

class MultiThreadedCrawler:
  def __init__(self, seed_url):
    self.seed_url = seed_url
    self.root_url = '{}://{}'.format(urlparse(self.seed_url).scheme,
                    urlparse(self.seed_url).netloc)
    self.pool = ThreadPoolExecutor(max_workers=5)
    self.scraped_pages = set([])
    self.crawl_queue = Queue()
    self.crawl_queue.put(self.seed_url)



  def parse_links(self, html):
    soup = BeautifulSoup(html, 'html.parser')
    Anchor_Tags = soup.find_all('a', href=True)
    for link in Anchor_Tags:
      url = link['href']
      if url.startswith('/') or url.startswith(self.root_url):
        url = urljoin(self.root_url, url)
        if url not in self.scraped_pages:
          self.crawl_queue.put(url)

  def scrape_info(self, html):
    soup = BeautifulSoup(html, "html5lib")
    web_page_paragraph_contents = soup('p')
    text = ''
    for para in web_page_paragraph_contents:
      if not ('https:' in str(para.text)):
        text = text + str(para.text).strip()
    #print(f'\n <---Text Present in The WebPage is --->\n', text, '\n')
    return

  def post_scrape_callback(self, res):
    result = res.result()
    if result and result.status_code == 200:
      self.parse_links(result.text)
      self.scrape_info(result.text)

  def scrape_page(self, url):
    try:
      res = requests.get(url, timeout=(3, 30))
      return res
    except requests.RequestException:
      return

  def run_web_crawler(self):
    for i in range(1,2000):
      try:
        print("\n Name of the current executing process: ",
          multiprocessing.current_process().name, '\n')
        target_url = self.crawl_queue.get(timeout=60)
        if target_url not in self.scraped_pages:
          print("Scraping URL: {}".format(target_url))
          self.current_scraping_url = "{}".format(target_url)
          self.scraped_pages.add(target_url)
          job = self.pool.submit(self.scrape_page, target_url)
          job.add_done_callback(self.post_scrape_callback)

      except Empty:
        return
      except Exception as e:
        print(e)
        continue

  def info(self):
    #print('\n Seed URL is: ', self.seed_url, '\n')
    print('Scraped pages are: ', self.scraped_pages, '\n')

if __name__ == '__main__':
  cc = MultiThreadedCrawler("https://www.crictracker.com/cricket-news/")
  cc.run_web_crawler()
  cc.info()
```

```
Scraping URL: https://www.crictracker.com/cricket-news/faf-du-plessis-recalls-steve-smiths-arrogant-behavior-during-2018-test-series/ (https://www.crictracker.com/cricket-news/faf
-du-plessis-recalls-steve-smiths-arrogant-behavior-during-2018-test-series/)

 Name of the current executing process:  MainProcess


 Name of the current executing process:  MainProcess

Scraping URL: https://www.crictracker.com/cricket-news/they-showed-such-bad-attitude-people-would-leave-themselves-salman-butt-slams-pcb-for-mistreating-experienced-players/ (http
s://www.crictracker.com/cricket-news/they-showed-such-bad-attitude-people-would-leave-themselves-salman-butt-slams-pcb-for-mistreating-experienced-players/)

 Name of the current executing process:  MainProcess


 Name of the current executing process:  MainProcess

Scraping URL: https://www.crictracker.com/cricket-news/i-have-dismissed-virat-kohli-earlier-so-it-doesnt-need-to-be-the-highlight-of-my-career-taijul-islam/ (https://www.crictrack
er.com/cricket-news/i-have-dismissed-virat-kohli-earlier-so-it-doesnt-need-to-be-the-highlight-of-my-career-taijul-islam/)

 Name of the current executing process:  MainProcess
```

In [1]:

```
# Scraped links
links=['https://www.crictracker.com/live-scores/ghana-vs-ew-2nd-match-t20i-icc-mens-t20-world-cup-sub-regional-africa-qualifier-group-b-2022/fixtures-and-results/', 'https://www.crict
```

In [2]:

```

len(links)
```

Out[2]:

```
1361
```

In [3]:

```
# First thousand links
s = links[:1000]
```

In [87]:

```
# Dataset creatin by HTML parsing
from bs4 import BeautifulSoup as Soup
import os
import re
import requests
def downloader(link):
    req = requests.get(link)
    req.encoding = "utf8"
    return req.text
for i in range(len(s)):
  name = f'IR_Project/Docs/'+"doc"+str(i)+".txt"
  contents = downloader(s[i])
  soup = Soup(contents, "html5lib")
  res = ''
  for i in range(len(soup.find_all('p'))):
    text = soup.find_all('p')[i].get_text()
    res = res + text
  with open(name,'a', encoding="UTF-8") as f1:
    f1.write(res)
  f1.close()
```

...

In [4]:

```
import re
from os import listdir
from os.path import join, abspath
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

In [5]:

```

data_dir = '../IR_Project/Docs/'
DATA_SET_DIR = abspath(data_dir)
print('\nGetting List of text files from' + DATA_SET_DIR)
files = listdir(DATA_SET_DIR)
files = filter(lambda x: re.match(r'.*\.txt$', x, re.I), files)
print('\nFile list retrieved from ' + DATA_SET_DIR)
```

```

Getting List of text files fromC:\Users\sirik\IR_Project\Docs

File list retrieved from C:\Users\sirik\IR_Project\Docs
```

In [6]:

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\sirik\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\sirik\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[6]:

```
True
```

In [7]:

```

ps = PorterStemmer()
nltk_stop_words = set(stopwords.words('english'))
```

In [8]:

```
# build corpus: list of documents
# stop words are ignored, words are stemmed using PorterStemmer
corpus = []
for f in files:
    strm = open(DATA_SET_DIR + '/' + f, 'r', encoding="utf8")
    # using nltk word tokenizer to split file into word list
    words = word_tokenize(strm.read())
    # using filter to remove stop words from word list
    words = filter(lambda w: w not in nltk_stop_words, words)
    # using map to stem words in word list
    words = map(lambda w: ps.stem(str(w)), words)
    # joining words into string and adding to corpus list
    corpus.append(' '.join(words))
```

In [9]:

```
# Using sklearn's tfidfvectorizer to construct tfidf matrix
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(corpus)
terms = vectorizer.get_feature_names()
```

In [10]:

```
cos_sim_matrix = cosine_similarity(tfidf_matrix)
```

In [11]:

```
# Doc to doc cosine similarity matrix
cos_sim_matrix
```

Out[11]:

```
array([[1.        , 0.0395061 , 0.02679482, ..., 0.01074411, 0.05136209,
        0.01478737],
       [0.0395061 , 1.        , 0.07693528, ..., 0.01801358, 0.03732684,
        0.02671725],
       [0.02679482, 0.07693528, 1.        , ..., 0.01246013, 0.04090779,
        0.01269622],
       ...,
       [0.01074411, 0.01801358, 0.01246013, ..., 1.        , 0.01647358,
        0.35916592],
       [0.05136209, 0.03732684, 0.04090779, ..., 0.01647358, 1.        ,
        0.09638425],
       [0.01478737, 0.02671725, 0.01269622, ..., 0.35916592, 0.09638425,
        1.        ]])
```

In [12]:

```
pip install bm25
```

```
Requirement already satisfied: bm25 in c:\users\sirik\anaconda3\lib\site-packages (1.0.0)
Note: you may need to restart the kernel to use updated packages.
```

In [13]:

```
pip install rank_bm25
```

```
Requirement already satisfied: rank_bm25 in c:\users\sirik\anaconda3\lib\site-packages (0.2.2)
Requirement already satisfied: numpy in c:\users\sirik\anaconda3\lib\site-packages (from rank_bm25) (1.20.3)
Note: you may need to restart the kernel to use updated packages.
```

In [14]:

```
# calculating bm25 precision
from rank_bm25 import BM25Okapi

def bm25_precision10(query, corpus):
    tokenized_corpus = [doc.split(" ") for doc in corpus]
    bm25 = BM25Okapi(tokenized_corpus)
    tokenized_query = query.split(" ")
    print("Top 10 relevant docs in the query :")
    doc_scores = bm25.get_scores(tokenized_query)
    docscores = list(doc_scores)
    topten = sorted(range(len(docscores)), key=lambda i: docscores[i], reverse=True)[:10]
    rl=[]
    for i in(topten):
        rl.append(s[i])


#     print(topten)
#     print("Top 3 relevant docs are :")
    return rl,bm25.get_top_n(tokenized_query, corpus, n=3)
```

In [15]:
```
# results without GUI
qs = "Kohli is evergreen classic player"
corp = corpus
bm25_precision10(qs, corp)
```

```
Top 10 relevant docs in the query :
```

Out[15]:
```
(['https://www.crictracker.com/live-scores/msc-w-vs-rac-w-match-4-woment20-bengal-womens-t20-challenge-2022/fantasy-tips/',
  'https://www.crictracker.com/live-scores/ban-vs-ind-1st-test-india-tour-of-bangladesh-2022/',
  'https://www.crictracker.com/cricket-players/sunil-gavaskar/',
  'https://www.crictracker.com/cricket-news/pak-vs-eng-haris-rauf-ruled-out-of-2nd-test-match-owing-to-an-injury-4445/',
  'https://www.crictracker.com/t20-wc-sub-regional-africa-qualifier-b/archives/',
  'https://www.crictracker.com/live-scores/tba-vs-tba-eliminator-t20-lanka-premier-league-2022/',
  'https://www.crictracker.com/live-scores/srh-vs-pbks-match-70-t20-indian-premier-league-2022/full-scorecard/',
  'https://www.crictracker.com/live-scores/pak-vs-eng-6th-t20i-england-tour-of-pakistan-2022/full-scorecard/',
  'https://www.crictracker.com/live-scores/ghana-vs-ew-2nd-match-t20i-icc-mens-t20-world-cup-sub-regional-africa-qualifier-group-b-2022/news/',
  'https://www.crictracker.com/cricket-videos/zambia-t10-league-live-streaming,-match-12,-lusaka-heats-vs-kitwe-kings/'],
 ["author view : 2.8k2 min readupd - nov 30 , 2022 , 12:20 istget everi cricket updat ! follow us onth mini-auct ipl 2023 schedul take place decemb 23 kochi , hugh edmead set return auction . the ipl auction host edmead sinc succeed previou host richard madley 2018 . he collaps due postur hypotens first day ipl 2022 mega auction held februari 2022.the auction halt unfortun incid board control cricket india ( bcci ) office-bear ask charu sharma step edmead medic team attend . member franchis bcci offici applaud loudli gave edmead stand ovat return lead auction 's final phase.ther doubt , nevertheless , whether board would appoint conduct auction . howev , edmead state tuesday lead auction upcom season cash-rich event travel kochi via dubai decemb 21. “ i thrill ask bcci conduct 2023 ipl auction excit visit kochi first time , ” edmead told sportstar.befor go independ 2016 , edmead , independ fine art , classic car , chariti auction , spent 38 year work renown christi 's . even experi conduct 2500 auction global , edmead experienc someth entir new visit jaipur decemb 2018 ipl auction , first term indian cricket.aft complet ipl 2022 mega auction februari , edmead said `` stupid '' miss meal , may one reason low blood pressur . put past behind , season auction , howev , anticip '' excit '' auction kochi 10 franchise go make wise purchases.© 2013 - 2022 crictrack pvt ltd all right reserv .",
  "author view : 1.4k2 min readupd - dec 05 , 2022 , 16:21 istget everi cricket updat ! follow us onin open one-day intern dhaka , team india shockingli defeat host bangladesh , one wicket take 1-0 seri lead . against bangladeshi attack , led shakib al hasan ebadot hossain , indian bat lineup crumbl like cooki . shakib ebadot share nine wicket , former claim five-wicket haul.th indian bowler kept thing control second inning defend low total . it turn cut-throat competit bangladesh reduc 136/9 39.3 over india need one wicket win first game . but blunder indian fielder gave bangladesh bonu life , indian bowler fail pick final one wicket . as result , mehidi hasan miraz game hosts.mani cricket pundit shock india ’ defeat , even share harsh critic rohit sharma men . one among former team india cricket mohammad kaif , also rais import question indian team ought take consideration. “ it india 's game , taken nine wicket . the bowl excel , got india back game batter bad day . the bowl cover 40th , last 10 over , death bowler ? is deepak chahar kuldeep sen ? , ” question kaif speak soni sport network post-match discussion.kl rahul drop sitter mehidi hasan miraz 15 . india need one wicket win drop catch chang cours game . soon mistak , washington sundar miss chanc take catch , eventu cost india game . the bowler conced run pressur result , bangladesh game one wicket. “ we drop catch . kl rahul n't keep often . he good fielder , ran litton da direct hit deep t20 world cup . sundar n't dive tri take catch . the fielder seen pressur . we made mistak pressur . we bowl wide ball no-bal . you overcom pressur win world cup . that team emerg , whether talk new zealand england , top white-bal cricket , ” added.© 2013 - 2022 crictrack pvt ltd all right reserv .",
  'sub-editor view : 3.8k3 min readupd - aug 09 , 2022 , 22:25 istget everi cricket updat ! follow us onsport hero take back good old day , induc nostalgia , provid someth treasur anticip . the legend leagu cricket offer . the leagu ’ first season held januari year muscat , oman , india maharaja , world giant , asia lions.llc set return alter format 2022 , four team compet event , held six indian citi : kolkata , lucknow , delhi , jodhpur , cuttack , rajkot , septemb 17 octob 8 . a draft mechan use pick player respect squad . the competit held india countri celebr 75th independ anniversari , indian fan present see favourit superstar action.meanwhil , crictrack caught raman raheja , co-found ceo legend leagu cricket , open leagu ’ vision goal , potenti icc accredit , possibl particip pakistani player , thing . raheja also associ sport entertain sector , includ ipl world kabaddi league.1 . with mani cricket leagu take place around globe , main vision legend ’ leagu cricket ? we talk cricket , legend seen grow . what happen , although good , compet new boy , unabl offer kind competit cricket . they forc take retir , fade oblivion . that want . there may five-six year cricket left . so , come bring cricket left them.2 . how hard sustain qualiti cricket competit go tournament ? when first season , sign multipl cricket , agre play . we look physic competit cricket – got andrew leipu director sport scienc , work cricket physic fit . the cricket land oman attend practic session took thing serious . the fan also love cricket back got chanc watch favourit star past back play competit cricket.3 . are plan get icc certif go forward ? what long-term goal project ? these retir cricket , ’ need take permiss respect board play cricket . so , suppos get permiss bcci icc , said , even season one , want adopt best practic cricket author , whether bcci icc . we start get touch icc ’ anticorrupt unit also panel umpir . and virtu , abl get support bcci icc . in fact , anti-corrupt head icc also visit oman . that ’ look carri forward.4 . chri gayl yet offici retir intern cricket , expect play season . could give clariti ? the cricket retir intern setup play intern leagu , purview board , consid retir us . but said , except mayb taken consider cricket play domest leagu featur intern cricket.5 . would pakistan player particip leagu ? that challeng . we seek govern advic ; input ministri extern affair , necessari permiss sought . we would want bring govern permits.6 . is possibl pakistan cricket come india tournament ? there definit possibl ; ’ rule . we announc pakistan cricket pool . but point , tournament schedul take place oman . but shift india , hold back pakistani cricket till get necessari permiss . we would like get appropri author give us permiss grant visa approvals.© 2013 - 2022 crictrack pvt ltd all right reserv .'])
```

In [16]:

```
# calculating cosine similarity
from collections import Counter
from math import sqrt

def cosine_similarity(query, corpus):
    """
    Calculate the cosine similarity between a query and a document in a corpus.
    """
    csres=[]
    # Convert the query and document to lowercase and remove any punctuation or stop words
    for i in corpus:
        #query = preprocess(query)
        document = i

    # Convert the query and document to Counter objects to easily count the frequencies of words
        query_counter = Counter(query)
        document_counter = Counter(document)

    # Get the set of words that are in either the query or the document
        words = set(query_counter.keys()).union(set(document_counter.keys()))

    # Calculate the dot product of the two vectors
        dot_product = sum(query_counter.get(w, 0) * document_counter.get(w, 0) for w in words)

    # Calculate the magnitude of the query vector
        query_mag = sqrt(sum(query_counter.get(w, 0) ** 2 for w in words))

    # Calculate the magnitude of the document vector
        document_mag = sqrt(sum(document_counter.get(w, 0) ** 2 for w in words))

    # Return the cosine similarity
        csres.append(dot_product / (query_mag * document_mag))
    cl=[]
    cr = sorted(range(len(csres)), key=lambda i: csres[i], reverse=True)[:10]
    for i in cr:
        cl.append(s[i])


    print("top 10 relevant through cosine similarity :")
#     print(cr)

    return cr, cl#, sorted(range(len(csres)), key=lambda i: csres[i], reverse=True)[:10]
```

In [17]:

```
# results without GUI
qs = "Kohli is the king"
corp = corpus
cosine_similarity(qs, corp)
```

```
top 10 relevant through cosine similarity :
```

Out[17]:

```
([616, 700, 538, 453, 150, 305, 530, 48, 680, 971],
 ['https://www.crictracker.com/cricket-news/when-dhoni-bhai-was-speaking-it-literally-felt-like-it-was-gaikwad-azim-kazi/',
  'https://www.crictracker.com/fantasy-cricket-tips/dream11-bn-a-vs-in-a-dream11-prediction-test-series-2022-fantasy-cricket-tips-playing-xi-pitch-report-injury-updates-for-1st-unoff
icial-test/',
  'https://www.crictracker.com/if-we-take-psl-to-auction-model-then-well-see-who-goes-to-play-ipl-over-psl-ramiz-raja/',
  'https://www.crictracker.com/cricket-news/sir-vivian-richards-appointed-as-brand-ambassador-of-lanka-premier-league/',
  'https://www.crictracker.com/ban-vs-ind/fantasy-tips/',
  'https://www.crictracker.com/6ixty-men-2022-semi-final-2-br-vs-tkr-dream11-prediction-fantasy-cricket-tips-playing-11-pitch-report-and-injury-update/',
  'https://www.crictracker.com/cricket-news/you-two-wont-get-out-alive-in-barbados-brad-haddin-recalls-ugly-on-field-confrontation-with-sulieman-benn/',
  'https://www.crictracker.com/cricket-players/scott-boland/',
  'https://www.crictracker.com/cricket-news/pak-vs-eng-haris-rauf-ruled-out-of-2nd-test-match-owing-to-an-injury-4445/',
  'https://www.crictracker.com/cricket-teams/india-women/'])
```

In [18]:

```
pip install PySimpleGUI
```

```
Requirement already satisfied: PySimpleGUI in c:\users\sirik\anaconda3\lib\site-packages (4.60.4)
Note: you may need to restart the kernel to use updated packages.
```

In [19]:

```
!apt-get install -y xvfb
```

```
'apt-get' is not recognized as an internal or external command,
operable program or batch file.
```

In [20]:

```
from rank_bm25 import BM25Okapi

import PySimpleGUI as sg

layout = [
    [sg.VPush()],
    [sg.Text("Search: "), sg.Input(key='INPUT')],
    [sg.Ok()],
    [sg.Text("", size=(0, 1), key='OUTPUT2'),],
    [sg.Text("", size=(170, 20), key='OUTPUT'),  ],
    [sg.Text("", size=(100, 1), key='OUTPUT1'),  ],
    [sg.VPush()],
]



window = sg.Window("BM25", layout, size=( 1400, 600), element_justification='c')

while True:
    event, values = window.read()
    if event == sg.WINDOW_CLOSED:
        break
    elif event == 'Ok':
        name = values['INPUT']
        tokenized_query = name.split(" ")
#         doc_scores = bm25.get_scores(tokenized_query)
        window['OUTPUT2'].update(name)
        window['OUTPUT'].update(value=bm25_precision10(name, corpus))
        #window['OUTPUT1'].update(cos_sim_matrix)

window.close()
```

```
Top 10 relevant docs in the query :
```

In [21]:
```
import PySimpleGUI as sg

layout = [
    [sg.VPush()],
    [sg.Text("Search: "), sg.Input(key='INPUT')],
    [sg.Ok()],
    [sg.Text("", size=(0, 1), key='OUTPUT2'),],
    [sg.Text("", size=(170, 20), key='OUTPUT'),  ],
    [sg.Text("", size=(100, 1), key='OUTPUT1'),  ],
    [sg.VPush()],
]



window = sg.Window("Cosine Similarity", layout, size=( 1400, 600), element_justification='c')

while True:
    event, values = window.read()
    if event == sg.WINDOW_CLOSED:
        break
    elif event == 'Ok':
        name = values['INPUT']
        tokenized_query = name.split(" ")
#         doc_scores = bm25.get_scores(tokenized_query)
        window['OUTPUT2'].update(name)
        window['OUTPUT'].update(value=cosine_similarity(name, corpus))
        #window['OUTPUT1'].update(cos_sim_matrix)

window.close()
```

```
top 10 relevant through cosine similarity :
```

In [ ]:

In [ ]:

In [ ]:
```
vi
```