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

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

```
path='/content/drive/MyDrive/cfc'
```

```
import os
import pandas as pd
data= pd.DataFrame( columns=['Doc_Id', 'Abstract'])
files = os.listdir(path)
doc_names=[]
for fl in files:
  if fl[:3]=='cf7':
    filepath=path+'/'+fl
    with open(filepath) as f:
      lines = f.readlines()
      for i in range(len(lines)):
        if lines[i][:3]=="RN ":
          rn=lines[i][3:-2]
          #print('Paper:',rn)
          i+=6
        if lines[i][:3]=="AB " or lines[i][:3]=="EX ":
            ab=lines[i][3:-1]
            i+=1
            while(lines[i][:3]!='RF '):
              ab+=lines[i][2:-1]
              i+=1
            print('Paper:',rn)
            print(ab)
            dic={'Doc_Id':rn,'Abstract':ab}
            data=data.append(dic,ignore_index=True)
            rn+='.txt'
            doc_names.append(rn)
            f = open(rn, 'w')  # Open file
            f.write(ab)  # Write string
            f.close()
```

```
Paper: 0016
Cystic Fibrosis is a generalized hereditary disorder of children, adolescents, and young adults in which there is widespread dysfunct
Paper: 0016
In five patients with cystic fibrosis of the pancreas the mucous glandular system of the conjunctiva was studied, as changes, if any,
Paper: 0017
A study has been made of plasma tocopherol concentrations in normal children and in children with intestinal abnormalities. A positiv
Paper: 0017
Five glycosidases, alpha-fucosidase, alpha-galactosidase, alpha- glucosidase, beta-mannosidase and N-acetyl-alpha-glucosaminidase wer
Paper: 0017
The electrical potential difference (PD) across the rectal wall was measured in 26 patients with cystic fibrosis of pancreas (CFP) ar
Paper: 0017
By measuring potential difference between rectal mucosa and perianal skin using a reference electrode placed on the forearm, we demor
Paper: 0017
Fifty cystic fibrosis (CF) patients, of whom 9 had multilobular cirrhosis, were observed regularly for a period of 3 years and variou
Paper: 0017
A clinical study of the albumin content in meconium was performed on two categories of newborn infants: a screening series of 8,830 i
Paper: 0017
The simultaneously occurring mucoid (M) and non-mucoid (NM) variants of Pseudomonas aeruginosa frequently observed in cultures from t
Paper: 0017
The relative prevalence of mucoid strains compared with non-mucoid strains of Pseudomonas aeruginosa has been investigated in all rou
Paper: 0017
The occurrence of antibodies against antigens prepared from strains representing 13 O groups of Pseudomonas aeruginosa and against a
Paper: 0017
During the recent decade, 1651 isolates of Staphylococcus aureus from 111 patients with cystic fibrosis have been tested for antibiot
Paper: 0018
During recent years, more and more data have been accumulated implicating early malnutrition in subsequent small stature and behavior
Paper: 0018
Assessment of nutritional status of patients with cystic fibrosis of the pancreas (CFP) showed that poor growth was associated with l
Paper: 0018
A patient with cystic fibrosis was found to have pneumatosis coli associted with rectal prolapse.  In cystic fibrosis there are sever
Paper: 0018
Three patients with cystic fibrosis were noted to have swelling of knee and ankle joints during exacerbation of their lung disease. S
Paper: 0018
Sixty-one patients with cystic fibrosis were studied to determine the relationship between degree of compliance with taking antibioti
Paper: 0018
Duke Medical Center and the National Institutes of Health have analyzed the maximum achieved heights and weights of 60 persons with C
Paper: 0018
Intraluminal bowel obstruction secondary to inspissated feces is a known complication of cystic fibrosis. When seen in the older chil
Paper: 0018
```

```
The respiratory flora of patients with cystic fibrosis (CF) frequently includes Aspergillus, and 30% of their serum samples have beer
Paper: 0018
A heptavalent lipopolysaccharide Pseudomonas vaccine was evaluated in 22 patients with acute leukemia and 12 patients with cystic fib
Paper: 0018
A nurse and mother describes the clinical effects of cystic fibrosis on patients and the long-term demands the disease places on her
Paper: 0019
The characteristic increased salinity of sweat and other abnormalities of exocrine secretions in patients with cystic fibrosis (CF) s
Paper: 0019
Assays of carboxypeptidase B-like activity and C3 in serum from patients with cystic fibrosis and appropriate control subjects failed
Paper: 0019
Pulmonary function in children with cystic fibrosis was assessed by the arterial-alveolar PN2 difference adjusted to sublingual tempe
Paper: 0019
The results of open thoractomy and pleurectomy or pleural abrasion for 17 episodes of pneumothorax in patients with cystic fibrosis w
Paper: 0019
Normal children as well as those with asthma and cystic fibrosis were studied to assess the contribution of lung zones emptying at di
Paper: 0019
Tracheal mucous velocity was measured by observing the motion of teflon discs across the tracheal mucosa through a fiberoptic broncho
Paper: 0019
```

```
from nltk.stem import PorterStemmer
import numpy as np

def tokenize(txt):
  symbols = "!\"#$%&()*+-./:;,<=>?@[\]^_`{|}~\n"
  for i in symbols:
    txt = txt.replace(i,' ')
  return txt.split()

def remove_stopwords(words):
  stopwords=['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",
           "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself',
           'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
           'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
           'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had',
           'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as',
           'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during',
           'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
           'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
           'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so',
           'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd',
           'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn',
           "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
           'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "wo
  return [word for word in words if word not in stopwords]

def porter_stemmer(words):
  stemmer = PorterStemmer()
  return [stemmer.stem(word) for word in words]
```

```
a=tokenize(data['Abstract'][0])
a=remove_stopwords(a)
print(porter_stemmer(a))
```

```
['cystic', 'fibrosi', 'gener', 'hereditari', 'disord', 'children,', 'adolescents,', 'young', 'adult', 'widespread', 'dysfunct', 'mucu',
```

```
data['Tokenize'] = data['Abstract'].apply(tokenize)
```

```
data['Removed_stopwords'] = data['Tokenize'].apply(remove_stopwords)
```

```
data['Stemmed'] = data['Removed_stopwords'].apply(porter_stemmer)
```

```
data.head()
```

	Doc_Id	Abstract	Tokenize	Removed_stopwords
0	0016	Cystic Fibrosis is a generalized hereditary di...	[Cystic, Fibrosis, is, a, generalized, heredit...	[Cystic, Fibrosis, generalized, hereditary, di...
1	0016	In five patients with cystic fibrosis of the p...	[In, five, patients, with, cystic, fibrosis, o...	[In, five, patients, cystic, fibrosis, pancrea...
2	0017	A study has been made of plasma tocopherol con...	[A, study, has, been, made, of, plasma, tocoph...	[A, study, made, plasma, tocopherol, concentra...

```
print(data['Doc_Id'], end=' ')
```

```
0        0016
1        0016
2        0017
3        0017
4        0017
        ...
1243    01235
1244    01236
1245    01237
1246    01238
1247    01239
Name: Doc_Id, Length: 1248, dtype: object
```

```
dictionary=[]
for i in range(len(data)):
  for j in data['Stemmed'][i]:
    dictionary.append(j)
```

```
dictionary =set(dictionary)
```

```
def D(W):
  count=0
  for i in range(len(data)):
      if W in data['Stemmed'][i]:
        count+=1
  return count

def F(docID, w):
      count=0
      for i in range(len(data)):
        if data['Doc_Id'][i]==docID:
            for j in data['Stemmed'][i]:
              if j==w:
                count+=1
            break
      return count

def C(docID):
      count=0
      for i in range(len(data)):
          if data['Doc_Id'][i]==docID:
              count=len(data['Stemmed'][i])
              break
      return count

def TF(docID,w):
  return F(docID, w) / C(docID)

def IDF(size,w):
   return size / D(w)


print(D('sweat'))
print(F('00981','120'))
print(C('00981'))
```

```
145
0
85
```

```
Term_set=[]
for i in dictionary:
    if D(i)>=3:
      Term_set.append(i)
```

```
dic={}
for i in range(len(data)):
    lst=[]
    for j in Term_set:
        if j in data['Stemmed'][i]:
          lst.append(1)
        else:
          lst.append(0)
    dic[data['Doc_Id'][i]]=lst


from sklearn.metrics import jaccard_score
from heapq import nsmallest,nlargest
def top_similar_doc_jaccard(bol_vec,doc,k=3):
        lst={}
        for i in bol_vec.keys():
            lst[i]=jaccard_score(bol_vec[doc],bol_vec[i])
        top_similar_doc = nlargest(k+1, lst, key = lst.get)
        return top_similar_doc

print(top_similar_doc_jaccard(dic,'00001',3))
print(top_similar_doc_jaccard(dic,'00002',3))
print(top_similar_doc_jaccard(dic,'00003',3))


     ['00001', '00415', '00983', '00987']
     ['00002', '0020', '00777', '00486']
     ['00003', '00004', '00379', '00639']
```

Count vector

```
count_vector={}
for i in range(len(data)):
    lst=[]
    for j in Term_set:
         lst.append(data['Stemmed'][i].count(j))
    count_vector[data['Doc_Id'][i]]=lst



from  scipy.spatial.distance import cosine
from heapq import nsmallest,nlargest

def top_similar_doc_cosine(count_vec,doc,k=3):
        lst={}
        for i in count_vec.keys():
            lst[i]=1-cosine(count_vec[doc],count_vec[i])
        top_similar_doc = nlargest(k+1, lst, key = lst.get)
        return top_similar_doc

print(top_similar_doc_cosine(count_vector,'00001',3))
print(top_similar_doc_cosine(count_vector,'00002',3))
print(top_similar_doc_cosine(count_vector,'00003',3))


     ['00001', '00415', '00778', '00160']
     ['00002', '00051', '00420', '00619']
     ['00003', '00004', '00786', '00901']
```

TI_IDF Vector

```
TF_IDF_vector={}
size=len(data)
for i in range(size):
    lst=[]
    docid=data['Doc_Id'][i]
    for j in Term_set:
         lst.append(TF(docid,j)*IDF(size,j))
    TF_IDF_vector[docid]=lst
print(top_similar_doc_cosine(TF_IDF_vector,'00001',3))
print(top_similar_doc_cosine(TF_IDF_vector,'00002',3))
print(top_similar_doc_cosine(TF_IDF_vector,'00003',3))
```

```
IDF(1239,'the')
TF('00001','the')
```

```
0.038461538461538464
```

```
data.head()
```

	Doc_Id	Abstract	Tokenize	Removed_stopwords	Stemmed
0	0016	Cystic Fibrosis is a generalized hereditary di...	[Cystic, Fibrosis, is, a, generalized, heredit...	[Cystic, Fibrosis, generalized, hereditary, di...	[cystic, fibrosi, gener, hereditari, disord, c...
1	0016	In five patients with cystic fibrosis of the p...	[In, five, patients, with, cystic, fibrosis, o...	[In, five, patients, cystic, fibrosis, pancrea...	[In, five, patient, cystic, fibrosi, pancrea, ...
2	0017	A study has been made of plasma tocopherol con...	[A, study, has, been, made, of, plasma, tocoph...	[A, study, made, plasma, tocopherol, concentra...	[A, studi, made, plasma, tocopherol, concentr,...
3	0017	Five glycosidases, alpha-fucosidase,	[Five, glycosidases,, alpha, fucosidase,,	[Five, glycosidases,, alpha, fucosidase,, alph...	[five, glycosidases,, alpha, fucosidase,,