68_Adnan Shaikh

TF-IDF document retrival with cosine similarity using Vector Space Model

In [1]:
```
import pandas as pd
import numpy as np
import string
import nltk
from nltk.corpus import gutenberg, stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
```

In [2]:
```
gutenberg.fileids()
```

Out[2]:
```
['austen-emma.txt',
 'austen-persuasion.txt',
 'austen-sense.txt',
 'bible-kjv.txt',
 'blake-poems.txt',
 'bryant-stories.txt',
 'burgess-busterbrown.txt',
 'carroll-alice.txt',
 'chesterton-ball.txt',
 'chesterton-brown.txt',
 'chesterton-thursday.txt',
 'edgeworth-parents.txt',
 'melville-moby_dick.txt',
 'milton-paradise.txt',
 'shakespeare-caesar.txt',
 'shakespeare-hamlet.txt',
 'shakespeare-macbeth.txt',
 'whitman-leaves.txt']
```

Statistics of Gutenberg Corpus (Frequency and Expectation)

In [3]:
```
gutenberg_analysis  = [
    [
        len(gutenberg.raw(text)),len(gutenberg.words(text)),
        len(gutenberg.sents(text)),len(set([w.lower() for w in gutenberg.wor
ds(text)])),
        text
    ]
    for text in gutenberg.fileids()
]
```

In [4]:

```
for chars,words,sents,vocubs,name in gutenberg_analysis:
    print(f"Text Name: {name} \n\
            Number of characters: {chars}\n\
            Number of words: {words}\n\
            Number of sentences: {sents}\n\
            Number of unique words: {vocubs}\n\
            Average word length: {int(chars/words)}\n\
            Average word in a sentence: {int(words/sents)}\n\
            Average frequency of a word: {int(words/vocubs)}\n\
            ")
```

Text Name: austen-emma.txt
Number of characters: 887071
Number of words: 192427
Number of sentences: 7752
Number of unique words: 7344
Average word length: 4
Average word in a sentence: 24
Average frequency of a word: 26

Text Name: austen-persuasion.txt
Number of characters: 466292
Number of words: 98171
Number of sentences: 3747
Number of unique words: 5835
Average word length: 4
Average word in a sentence: 26
Average frequency of a word: 16

Text Name: austen-sense.txt
Number of characters: 673022
Number of words: 141576
Number of sentences: 4999
Number of unique words: 6403
Average word length: 4
Average word in a sentence: 28
Average frequency of a word: 22

Text Name: bible-kjv.txt
Number of characters: 4332554
Number of words: 1010654
Number of sentences: 30103
Number of unique words: 12767
Average word length: 4
Average word in a sentence: 33
Average frequency of a word: 79

Text Name: blake-poems.txt
Number of characters: 38153
Number of words: 8354
Number of sentences: 438
Number of unique words: 1535
Average word length: 4
Average word in a sentence: 19
Average frequency of a word: 5

Text Name: bryant-stories.txt
Number of characters: 249439
Number of words: 55563
Number of sentences: 2863
Number of unique words: 3940
Average word length: 4
Average word in a sentence: 19
Average frequency of a word: 14

Text Name: burgess-busterbrown.txt
Number of characters: 84663
Number of words: 18963
Number of sentences: 1054
Number of unique words: 1559
Average word length: 4
Average word in a sentence: 17

```
            Average frequency of a word: 12

Text Name: carroll-alice.txt
            Number of characters: 144395
            Number of words: 34110
            Number of sentences: 1703
            Number of unique words: 2636
            Average word length: 4
            Average word in a sentence: 20
            Average frequency of a word: 12

Text Name: chesterton-ball.txt
            Number of characters: 457450
            Number of words: 96996
            Number of sentences: 4779
            Number of unique words: 8335
            Average word length: 4
            Average word in a sentence: 20
            Average frequency of a word: 11

Text Name: chesterton-brown.txt
            Number of characters: 406629
            Number of words: 86063
            Number of sentences: 3806
            Number of unique words: 7794
            Average word length: 4
            Average word in a sentence: 22
            Average frequency of a word: 11

Text Name: chesterton-thursday.txt
            Number of characters: 320525
            Number of words: 69213
            Number of sentences: 3742
            Number of unique words: 6349
            Average word length: 4
            Average word in a sentence: 18
            Average frequency of a word: 10

Text Name: edgeworth-parents.txt
            Number of characters: 935158
            Number of words: 210663
            Number of sentences: 10230
            Number of unique words: 8447
            Average word length: 4
            Average word in a sentence: 20
            Average frequency of a word: 24

Text Name: melville-moby_dick.txt
            Number of characters: 1242990
            Number of words: 260819
            Number of sentences: 10059
            Number of unique words: 17231
            Average word length: 4
            Average word in a sentence: 25
            Average frequency of a word: 15

Text Name: milton-paradise.txt
            Number of characters: 468220
            Number of words: 96825
            Number of sentences: 1851
            Number of unique words: 9021
```

```
            Average word length: 4
            Average word in a sentence: 52
            Average frequency of a word: 10

Text Name: shakespeare-caesar.txt
            Number of characters: 112310
            Number of words: 25833
            Number of sentences: 2163
            Number of unique words: 3032
            Average word length: 4
            Average word in a sentence: 11
            Average frequency of a word: 8

Text Name: shakespeare-hamlet.txt
            Number of characters: 162881
            Number of words: 37360
            Number of sentences: 3106
            Number of unique words: 4716
            Average word length: 4
            Average word in a sentence: 12
            Average frequency of a word: 7

Text Name: shakespeare-macbeth.txt
            Number of characters: 100351
            Number of words: 23140
            Number of sentences: 1907
            Number of unique words: 3464
            Average word length: 4
            Average word in a sentence: 12
            Average frequency of a word: 6

Text Name: whitman-leaves.txt
            Number of characters: 711215
            Number of words: 154883
            Number of sentences: 4250
            Number of unique words: 12452
            Average word length: 4
            Average word in a sentence: 36
            Average frequency of a word: 12
```

Calculating TF, DF. WF, IDF, WF-IDF Using Vector Space Model

Normalization

TF to WF:

wf = 1 + log(tf) if tf > 0 else 0

DF to IDF:

idf = log(N/df)

WF-IDF

wf-idf = wf x idf

In [5]:
```
stemmer = PorterStemmer()
sw = set(stopwords.words("english") + list(string.punctuation) + list(string
.ascii_letters))
freq = {
    "doc_freq": {},
}
for text in gutenberg.fileids():
    doc_name = text.replace("-","_").replace(".txt","")
    freq[doc_name] = {}

    for word in word_tokenize(gutenberg.raw(text)):
        if word in sw:
            continue
        stemmed_word = stemmer.stem(word)
        if stemmed_word in freq[doc_name]:
            freq[doc_name][stemmed_word] += 1
        else:
            freq[doc_name][stemmed_word] = 1
            if stemmed_word in freq["doc_freq"]:
                freq["doc_freq"][stemmed_word] += 1
            else:
                freq["doc_freq"][stemmed_word] = 1
```

In [6]:
```
vector_model = pd.DataFrame(freq).fillna(0)
vector_model.head()
```

Out[6]:

	doc_freq	austen_emma	austen_persuasion	austen_sense	bible_kjv	blake_poems	bryant_storie
emma	2	855.0	1.0	0.0	0.0	0.0	0
jane	3	301.0	1.0	1.0	0.0	0.0	0
austen	3	1.0	1.0	1.0	0.0	0.0	0
1816	1	1.0	0.0	0.0	0.0	0.0	0
volum	13	3.0	6.0	3.0	2.0	0.0	0

In [7]:
```
nvm = pd.DataFrame()  #normalized vector space
N = len(vector_model.columns[1:])
nvm["idf"] = np.log10(N/vector_model["doc_freq"])

for column in vector_model.columns[1:]:
    nvm["wf_"+column] = vector_model[column].apply(lambda x: 1+np.log10(x) i
f x>0 else 0)
    nvm["wf_idf_"+column] = nvm["idf"]*nvm["wf_"+column]
```

In [8]:
```
nvm
```

Out[8]:

	idf	wf_austen_emma	wf_idf_austen_emma	wf_austen_persuasion	wf_idf_austen_persuasio
emma	0.954243	3.931966	3.752049	1.000000	0.95424
jane	0.778151	3.478566	2.706851	1.000000	0.77815
austen	0.778151	1.000000	0.778151	1.000000	0.77815
1816	1.255273	1.000000	1.255273	0.000000	0.00000
volum	0.141329	1.477121	0.208760	1.778151	0.25130
...	...	...	...	...	
times'	1.255273	0.000000	0.000000	0.000000	0.00000
they.	1.255273	0.000000	0.000000	0.000000	0.00000
germin	1.255273	0.000000	0.000000	0.000000	0.00000
heart-thud	1.255273	0.000000	0.000000	0.000000	0.00000
blither	1.255273	0.000000	0.000000	0.000000	0.00000

36944 rows × 37 columns

In [9]:
```
def cosine_score(q,d):
    L2_q, L2_d = np.linalg.norm(q,ord=2), np.linalg.norm(d,ord=2)
    L2 = np.around(L2_q*L2_d,4)
    dot_product = np.around(np.dot(q,d),4)
    return np.around(dot_product/L2,4) if L2 >0 else dot_product

def query(q):
    qtf = {}
    for word in word_tokenize(q):

        if word in sw:
            continue

        if word in qtf:
            qtf[word] += 1
        else:
            qtf[word] = 1

    qtf = pd.Series(qtf)
    q_idf =  pd.Series({key:(nvm["idf"].loc[key] if key in nvm.index else 0)
for key in qtf.index})
    qtf_wf_idf = qtf.apply(lambda x: 1+np.log10(x)) * q_idf
    result = {}

    for doc in vector_model.columns[1:]:
        doc_wf_idf = pd.Series({key:(nvm["wf_idf_"+doc].loc[key] if key in n
vm.index else 0) for key in qtf.index})
        result[doc] = cosine_score(qtf_wf_idf,doc_wf_idf)

    return sorted(result.items(),key= lambda x: x[1],reverse=True)
```

In [26]:
```
query("caesar is smart")
```

Out[26]:
```
[('melville_moby_dick', 1.0),
 ('chesterton_thursday', 0.999),
 ('chesterton_brown', 0.9984),
 ('shakespeare_hamlet', 0.9962),
 ('bible_kjv', 0.9762),
 ('edgeworth_parents', 0.9631),
 ('shakespeare_macbeth', 0.9281),
 ('shakespeare_caesar', 0.928),
 ('milton_paradise', 0.3731),
 ('whitman_leaves', 0.3731),
 ('austen_persuasion', 0.3725),
 ('bryant_stories', 0.3725),
 ('chesterton_ball', 0.3725),
 ('austen_sense', 0.3723),
 ('burgess_busterbrown', 0.3721),
 ('austen_emma', 0.0),
 ('blake_poems', 0.0),
 ('carroll_alice', 0.0)]
```