In [6]:

```
import os
import math
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /home/adel/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[6]:

```
True
```

In [7]:

```
#Read the corpus
corpus = {}

for filename in os.listdir('files/holmes/'):
  with open(f'files/holmes/{filename}') as f:
    corpus[filename] = f.read()
```

In [8]:

```
#Tokenize the content
for filename in corpus:
  corpus[filename] = [word.lower() for word in nltk.word_tokenize(corpus[filename]) if word.isalpha()]
```

In [9]:

```
#Get all words
words = set()
for filename in corpus:
  words.update(corpus[filename])
```

In [10]:

```
#Calculate term frequency
tf = {}

for filename in corpus:
  tf[filename] = {word: corpus[filename].count(word) for word in words}
```

In [11]:

```
#Calculate the inverse document frequency
idf = {}

for word in words:
  freq = sum(word in corpus[filename] for filename in corpus)
  idf[word] = math.log(len(corpus) / freq)
```

In [12]:

```
#Calculate the Term Frequence-Inverse Document Frequency
tfidf = {}

for filename in corpus:
  tfidf[filename] = [(word, tf[filename][word] * idf[word]) for word in words]
```

In [13]:

```
#Sort the values
for filename in corpus:
  tfidf[filename] = sorted(tfidf[filename], key=lambda x: x[1], reverse=True)
```

In [14]:

```
#Print the top five words
for filename in corpus:
  print(filename)
  for term, score in tfidf[filename][:5]:
    print(f' {term}: {score}')
```

```
gloria_scott.txt
 trevor: 70.02401606763873
 beddoes: 33.489746814957655
 hudson: 24.396436929918487
 prendergast: 21.31165706406396
 boat: 18.81100205730782
crooked.txt
 barclay: 103.51376288259638
 colonel: 25.05525936990736
 aldershot: 18.81100205730782
 nancy: 18.26713462634054
 regiment: 14.108251542980867
bohemia.txt
 majesty: 54.80140387902161
 briony: 33.489746814957655
 irene: 32.919253600288684
 adler: 30.56787834312521
 photograph: 26.30802233840273
```

```
squires.txt
 cunningham: 94.3801955694261
 alec: 57.845926316745036
 acton: 45.667836565851346
 william: 31.506333455467118
 colonel: 31.3190742123842
patient.txt
 blessington: 79.157583380809
 trevelyan: 48.71235900357477
 brook: 24.356179501787384
 consultation: 15.222612188617115
 resident: 14.108251542980867
speckled.txt
 roylott: 60.89044875446846
 stoner: 57.845926316745036
 ventilator: 42.62331412812792
 stepfather: 36.53426925268108
 stoke: 33.489746814957655
twisted.txt
 clair: 82.20210581853242
 neville: 57.845926316745036
 lascar: 36.53426925268108
 opium: 25.865127828798254
 whitney: 24.356179501787384
interpreter.txt
 melas: 57.845926316745036
 mycroft: 49.37888040043303
 greek: 37.62200411461564
 interpreter: 33.489746814957655
 latimer: 21.31165706406396
coronet.txt
 coronet: 82.20210581853242
 arthur: 44.676129886106075
 gems: 39.5787916904045
 holder: 29.848105378863583
 snow: 23.513752571634775
treaty.txt
 phelps: 118.7363750712135
 joseph: 70.02401606763873
 harrison: 60.89044875446846
 woking: 42.62331412812792
 holdhurst: 42.62331412812792
bachelor.txt
 simon: 121.78089750893692
 doran: 36.53426925268108
 lestrade: 32.919253600288684
 wedding: 30.56787834312521
 lord: 29.655425113552127
blaze.txt
 straker: 115.69185263349007
 colonel: 55.121570613796194
 horse: 54.93061443340549
 trainer: 51.75688144129819
 moor: 48.71235900357477
face.txt
 cottage: 56.433006171923466
 munro: 18.81100205730782
 jack: 18.240508842638857
 grant: 16.459626800144342
 effie: 15.222612188617115
league.txt
 wilson: 42.81002327921689
 league: 37.62200411461564
 merryweather: 36.53426925268108
 jones: 33.489746814957655
 assistant: 32.919253600288684
boscombe.txt
 mccarthy: 112.64733019576666
 lestrade: 56.433006171923466
 turner: 49.37888040043303
 boscombe: 45.667836565851346
 pool: 42.3247546289426
problem.txt
 moriarty: 60.89044875446846
 professor: 28.216503085961733
 spray: 18.26713462634054
 rock: 18.26713462634054
 meiringen: 15.222612188617115
engineer.txt
 hydraulic: 30.44522437723423
 stark: 27.400701939510807
 eyford: 27.400701939510807
 colonel: 25.05525936990736
 engineer: 24.356179501787384
ritual.txt
 brunton: 70.02401606763873
 musgrave: 61.13575668625042
 ritual: 35.27062885745217
 hurlstone: 27.400701939510807
 butler: 21.52626787933984
carbuncle.txt
 goose: 61.13575668625042
 geese: 51.75688144129819
 horner: 39.5787916904045
 ryder: 36.53426925268108
 peterson: 33.489746814957655
copper.txt
 rucastle: 115.69185263349007
 hunter: 51.73025565759651
 toller: 45.667836565851346
 beeches: 25.865127828798254
 copper: 25.865127828798254
clerk.txt
 pycroft: 70.02401606763873
 mawson: 51.75688144129819
 pinner: 25.865127828798254
 hardware: 21.31165706406396
 birmingham: 21.1623773144713
```