In [2]:

```python
import os
import math
```

In [3]:

```python
#This function displays the menu as follows
#     1. Indexing
#     3. Exit

def printMenu():
    print('Menu:')
    print('Please enter 1 for indexing, 2 to search, and 3 to exit')
    result_menu = int(input('1. Indexzing\n2. Searching\n3. Exit\n\n'))
    return result_menu
printMenu()
```

```
Menu:
Please enter 1 for indexing, 2 to search, and 3 to exit
1. Indexzing
2. Searching
3. Exit

1
```

Out[3]: 1

In [5]:

```python
# This function takes a text file as input and replaces all punctuations int
# Input: text
# Output: text with no punctuations

def punctuationsRemoval(text: str) -> str:
    """
    (str) -> str
    Return a text after replacing any punctuations to blank, " ".

    >>>punctuationsRemoval('abcdefg##@@higklmn,*()%opqr""stu')
    'abcdefg     higklmn      opqr  stu'
    """
    PUNCTUATIONS = ["!", "(", ")", "—", "–", "-", "[", "]", "{", "}", ";", "

    clear_text = ""
    for word in text:
        clear_text += word if word not in PUNCTUATIONS else " "
            # if a variable "word" is not punctuation, then append "word" it

    return clear_text
```

In [7]:
```python
# This function takes a text as input and removes all stopwords.
# Input: text
# Output: text with nostop words
def stopWordRemoval(text: str) -> list:
    """
    (str) -> list
    Return word list based on the input data(txt format), which are excluded
    Before running this function, should be removed punctuations from the te

    >>>stopWordRemoval("The monkeys jump on the bed.")
    ['monkeys', 'jump', 'bed.']
    """
    word_list = list(filter(lambda word: word != "", text.split(" ")))
        # excluding "" from the text after splitting by the blnk and change

    with open('Stop_Words.txt', 'r') as f: # Load stopwords file
        Stop_Words = f.read()

    remove_list = [i.strip("'").strip('"') for i in Stop_Words.split(", ")
    clean_list = list(filter(lambda word: word not in remove_list, word_list

    return clean_list

stopWordRemoval("The monkeys jump on the bed.")
```

Out[7]:
```
['The', 'monkeys', 'jump', 'bed.']
```

In [9]:
```python
def appendTermDocFreq(docid: int, cleanText: list, termDocFreqFile: file_nam
    """
    Appends TermDocFreqFile with the term(lower case), the document number,
    The format is like below.

    ontario 1 2\n
    government 1 3\n
    """
    term_freq = {} # format -> {term: freqency}
    for word in cleanText:
        word = word.lower() # every word changes lower case
        if word not in term_freq:
            term_freq[word] = 1
        else:
            term_freq[word] += 1

    append_text = ''
    for k, v in term_freq.items():
        append_text += '{} {} {}\n'.format(k, docid, v)

    termDocFreqFile.write(append_text)
```

In [10]:
```python
def genIndex(termDocFreqFile):
    index_file = {}
    # Format -> {term_01: {doc#: freq, doc#: freq,...}, term_02:{doc#: freq,

    termDocFreqFile = open("TermDocFreq.txt", 'r', encoding='utf-8')
    for line in termDocFreqFile: # read text document line by line
        read = line[:-1].split(" ") # apply split method after removing the
        if read[0] not in index_file: # read[0], read[1], read[2] = term, do
            index_file[read[0]] = {read[1]: read[2]}
        else:
            index_file[read[0]][read[1]] = read[2]

    return index_file
```

In [11]:
```python
def readFolderContent():
    files = []
    file_list = os.listdir('dataset')
    for filename in sorted(file_list):
        with open('dataset' + '/' + filename, 'r', encoding='utf-8') as infi
            files.append(infile.read())
    return files
```

In [12]:
```python
def indexing():
    termDocFreqFile = open("TermDocFreq.txt", 'w', encoding='utf-8')

    # readFolderContent is called to create a list of files.
    files = readFolderContent()
    id=1
    for file in files:
        puncRemoved = punctuationsRemoval(file) # remove all punctuations
        stopWordsRemoved = stopWordRemoval(puncRemoved) # remove all stop wo
        appendTermDocFreq(id, stopWordsRemoved, termDocFreqFile)  # Call app
        id += 1

    global global_index_file
    global_index_file = genIndex(termDocFreqFile) # Call genIndex function t
    termDocFreqFile.close()
```

In [32]:

```
 1  def search(query: str) -> str:
 2      """
 3      Returns the document with the highest score as the most relevant one.
 4      Before running this function, must run "index()" function at first.
 5      """
 6      query_words = stopWordRemoval(punctuationsRemoval(query.lower())) # deal
 7  
 8      file_list = os.listdir('dataset')
 9      num_docs_total = len(file_list) # the total number of documents in the c
10      global scores
11      scores = {} # {document_number: scores}, which will be updated
12      for query in query_words:
13          num_docs_query= len([doc for doc in global_index_file[query].keys()]
14          for doc, freq in global_index_file[query].items(): # document number
15              scores[int(doc)] = 0 if int(doc) not in scores else scores[int(d
16              scores[int(doc)] += int(freq) * math.log(num_docs_total/num_docs
17  
18      most_relevant_doc_num = max(scores, key=scores.get)
19      file_name = 'dataset/'+ file_list[most_relevant_doc_num-1]
20  
21      with open(file_name, 'r', encoding='utf-8') as f:
22          most_relevant_document = f.read()
23  
24      return most_relevant_document
```

If I obeyed your instruction or its formula on MS2, a programmer might execute the "for" loop against all files, just like this.

```
for D in range(len(os.listdir('dataset'))):
...
```

However, the above code repeats all files, even if query words are not contained. I think it takes more time, so I tried to calculate the score of query each word at first. That is, the computer does not need to search around all the files to find user-specified query words. If a file(document) contains multiple query words, the score of each words are adding up (referring to **code line 13&14**).

In [33]:

```
def main():
    option=printMenu()
    if option == 1:
        indexing()
    elif option == 2:
        query= input("What's your query word or sentence?\n")
        return search(query)

#if __name__ == "__main__":
#       main()
```

```
In [34]: main()
```

```
Menu:
Please enter 1 for indexing, 2 to search, and 3 to exit
1. Indexzing
2. Searching
3. Exit

2
What's your query word or sentence?
Ontario document
```

```
Out[34]: 'London, Ont., Mayor Joe Fontana says in retrospect it was "stupid" of him to a
lter a document he submitted for expenses while he was a Liberal member of Parl
iament, but insists it was no forgery.Fontana took the stand Wednesday in his o
wn defence after pleading not guilty to fraud, breach of trust and uttering for
ged documents from his time as a cabinet minister.He admitted making seven chan
ges — including whiting out his wife\'s signature and replacing it with his own
— to an existing contract for a hall rental for his son\'s 2005 wedding to refl
ect an event he planned for then-finance minister Ralph Goodale at the same ven
ue.Other alterations on the contract were changing the date of the event from J
une 25, 2005 to Feb. 25, 2004, the word "wedding" to "reception" and the additi
on of a yellow sticky note saying "misc constituents reception."The event didn
\'t end up going ahead at the Marconi Club, but Fontana believed the club was o
wed a $1,700 deposit from his MP budget. Since he had only spoken with the club
\'s president over the phone and didn\'t have any paperwork, Fontana changed se
veral details on the wedding contract from a few months prior and submitted it,
he testified.One of the changes was to write the word original in quotation mar
ks at the top of the document."I took a document that I thought was null and vo
id...put \'original\' there so it wouldn\'t be confused with anything else," Fo
ntana said.During a testy cross-examination, Fontana used various terms to desc
ribe what he did to the contract — modified, changed, altered, and his most com
mon refrain, that he just created a new document — but he bristled at Crown att
orney Timothy Zuber\'s suggestion that it was a forgery."Yeah, excuse me?" Font
ana said after his lawyer objected. "Dumb, stupid, yes. I was busy, it was avai
lable...Things were harried at the time in Ottawa — a minority government."Zube
r asked Fontana why he wouldn\'t have just gone to the Marconi Club and asked f
or an invoice that he could submit."I submitted that document as proof," Fontan
a replied.Zuber wondered if there were any other occasions in which third-party
service providers for MP functions didn\'t provide him with a bill."Well no, be
cause they wouldn\'t be paid," Fontana said."My point exactly," Zuber replied.S
ince the Marconi Club never sent Fontana or the government a bill, Zuber sugges
ted that meant they didn\'t expect to be paid. Fontana suggested he took it upo
n himself to see that they be paid a deposit — he decided to use the same amoun
t as he deposited for his son\'s wedding — for reserving the hall."They held th
eir hall for me on a Friday night and therefore they couldn\'t use it for anyth
ing else," he testified. "Therefore I felt obligated to pay them for the use of
that hall."As the current mayor of London — Fontana has refused to step down —
would Fontana stop doing business with someone if he got wind they were submitt
ed altered documents, Zuber wondered."Probably," Fontana replied.Closing argume
nts are scheduled for Thursday, and Ontario Superior Court Judge Bruce Thomas a
sked the lawyers to address in their submissions, if he is to accept the defenc
e theory, "Did Mr. Fontana nonetheless commit an offence as set out in count 3
(forgery)?"The Crown said that a $1,700 Government of Canada cheque was ultimat
ely sent to the Marconi Club, where it was listed on their books as payment for
Fontana\'s son\'s wedding.The court has heard conflicting evidence about whethe
```

```
r government officials had been instructed to reimburse Fontana or the Marconi
Club the $1,700. Fontana insisted it was intended for the club, but the Crown s
uggested otherwise.Zuber suggested that Fontana didn\'t intend for the Marconi
Club to put the money toward his son\'s wedding, but that the cheque was suppos
ed to be sent to him to "line your pockets with $1,700 cash." Fontana disagree
d.The president of the Marconi Club at the time testified Wednesday that he exp
lained to the general manager that the government cheque was for a cancelled re
ception, but the general manager testified he did not recall that conversatio
n.'
```

```
In [ ]:
```