Exercise 1: Increasing the size of the collection

Several collection files of increasing size are available on the website of the course:

```
55k 01-Text_Only-Ascii-Coll-1-10-NoSem.gz
52k 02-Text_Only-Ascii-Coll-11-20-NoSem.gz
103k 03-Text_Only-Ascii-Coll-21-50-NoSem.gz
96k 04-Text_Only-Ascii-Coll-51-100-NoSem.gz
357k 05-Text_Only-Ascii-Coll-101-200-NoSem.gz
559k 06-Text_Only-Ascii-Coll-201-500-NoSem.gz
747k 07-Text_Only-Ascii-Coll-501-1000-NoSem.gz
1.2M 08-Text_Only-Ascii-Coll-1001-2000-NoSem.gz
4.1M 09-Text_Only-Ascii-Coll-2001-5000-NoSem.gz
```

Index each of these files using your indexing program (cf. Practical session n°1: dictionary, postings lists, *df*, *tf*). Build a time efficiency graph of your program.

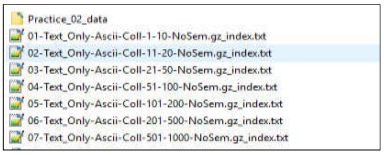
	Collection size	Time efficiency (seconds)
Col 1-10	10	0,08
Col 11-20	20	0,13
Col 21-50	50	0,28
Col 51-100	100	0,43
Col 101-200	200	1,87
Col 201-500	500	5,72
Col 501-1000	1000	14,03
Col 1001-2000	2000	39,00
Col 2001-5000	5000	314,70

As the collection size increases, the time it takes to process also increases

Variants: read several files instead of only one, uncompress a file if it is compressed, insert an option to print or not the index, print the indexing time. For large collections, you must not print the index!

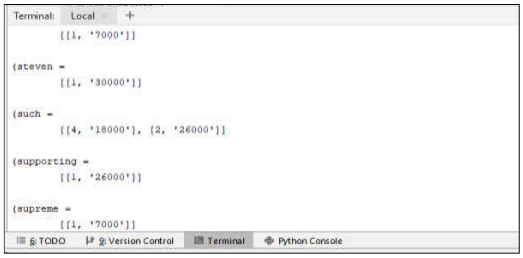
Read several files and uncompress if it's compressed:

```
for indice, archivo in enumerate(listaArchivos):
    start = time.time()
    with gzip.open("Practice_02_data/"+archivo, 'rt', encoding='utf8') as myfile:
        data = myfile.read().replace('\n', '')
```



Print the index:

```
with open(archivo +'_index.txt', 'w') as file:
    for key, value in sortedIndex.items():
        print(key + ' = \n\t' + str(value) + '\n')
        file.write(key + ' = \n\t' + str(value) + '\n')
```



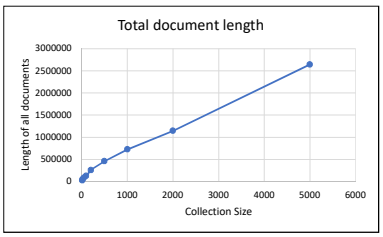
Exercise 2: Collection Statistics

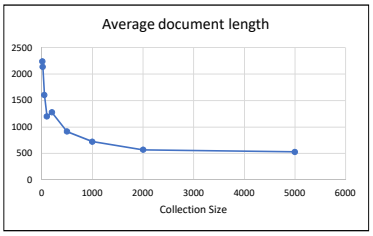
Modify your indexing program so that it computes different statistics on the indexed collection, for instance:

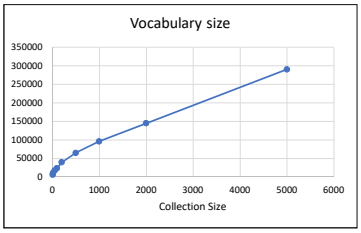
1. **document length,**
2. **term length,**
3. **vocabulary size,**
4. **collection frequency of terms.**

Plot the evolution of these statistics as the collection size grows

	Collection size	Total document length	Average document length	Vocabulary size
Col 1-10	10	22373	2237	6078
Col 11-20	20	42737	2137	9874
Col 21-50	50	80113	1602	16967
Col 51-100	100	119744	1197	22891
Col 101-200	200	255814	1279	39701
Col 201-500	500	456532	913	65027
Col 501-1000	1000	722913	723	96036
Col 1001-2000	2000	1139876	570	145008
Col 2001-5000	5000	2642185	528	290490







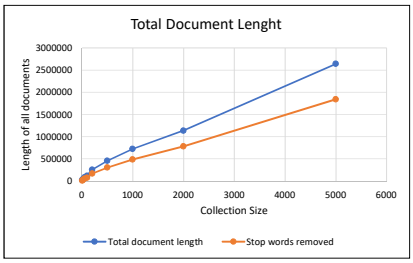
As the collections size increases the average documente decreases. The vocabulary size increases, on the other hand.

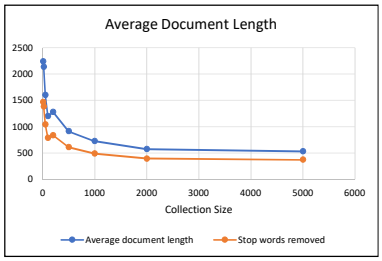
Exercise 3: Stop-words

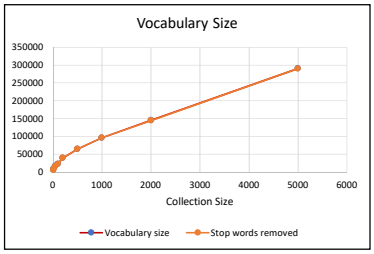
Download a stop-words list (cf. list, lecture n°2).
Refresh the index of the 9th file of the exercise n°1, removing stop words.
Compute again the statistics of the exercise n°2

Stop word list:
("also","although","always","am","among", "amongst", "amoungst", "amount", "an", "and", "another", "any","anyhow","anyone","anything","anyway", "anywhere", "are", "around", "as", "at", "back","be","became",
"because","become","becomes", "becoming", "been", "before", "beforehand", "behind", "being", "below", "beside", "besides", "between", "beyond", "bill", "both", "bottom","but", "by", "call", "can", "cannot",
"cant", "co", "con", "could", "couldnt", "cry", "de", "describe", "detail", "do", "done", "down", "due", "during", "each", "eg", "eight", "either", "eleven","else", "elsewhere", "empty", "enough", "etc",
"even", "ever", "every", "everyone", "everything", "everywhere", "except", "few", "fifteen", "fify", "fill", "find", "fire", "first", "five", "for", "former", "formerly", "forty", "found", "four", "from",
"front", "full", "further", "get", "give", "go", "had", "has", "hasnt", "have", "he", "hence", "her", "here", "hereafter", "hereby", "herein", "hereupon", "hers", "herself", "him", "himself", "his", "how",
"however", "hundred", "ie", "if", "in", "inc", "indeed", "interest", "into", "is", "it", "its", "itself", "keep", "last", "latter", "latterly", "least", "less", "ltd", "made", "many", "may", "me", "meanwhile",
"might", "mill", "mine", "more", "moreover", "most", "mostly", "move", "much", "must", "my", "myself", "name", "namely", "neither", "never", "nevertheless", "next", "nine", "no", "nobody", "none", "noone",
"nor", "not", "nothing", "now", "nowhere", "of", "off", "often", "on", "once", "one", "only", "onto", "or", "other", "others", "otherwise", "our", "ours", "ourselves", "out", "over", "own","part", "per",
"perhaps", "please", "put", "rather", "re", "same", "see", "seem", "seemed", "seeming", "seems", "serious", "several", "she", "should", "show", "side", "since", "sincere", "six", "sixty", "so", "some",
"somehow", "someone", "something", "sometime", "sometimes", "somewhere", "still", "such", "system", "take", "ten", "than", "that", "the", "their", "them", "themselves", "then", "thence", "there", "thereafter",
"thereby", "therefore", "therein", "thereupon", "these", "they", "thick", "thin", "third", "this", "those", "though", "three", "through", "throughout", "thru", "thus", "to", "together", "too", "top", "toward",
"towards", "twelve", "twenty", "two", "un", "under", "until", "up", "upon", "us", "very", "via", "was", "we", "well", "were", "what", "whatever", "when", "whence", "whenever", "where", "whereafter", "whereas",
"whereby", "wherein", "whereupon", "wherever", "whether", "which", "while", "whither", "who", "whoever", "whole", "whom", "whose", "why", "will", "with", "within", "without", "would", "yet", "you", "your",
"yours", "yourself", "yourselves", "the")

	Collection size	Total document length	Average document length	Vocabulary size
Col 1-10	10	14691	1469	5855
Col 11-20	20	27603	1380	9623
Col 21-50	50	52035	1041	16699
Col 51-100	100	78744	787	22620
Col 101-200	200	166708	834	39415
Col 201-500	500	303572	607	64737
Col 501-1000	1000	486670	487	95739
Col 1001-2000	2000	783766	392	144710
Col 2001-5000	5000	1843165	369	290192





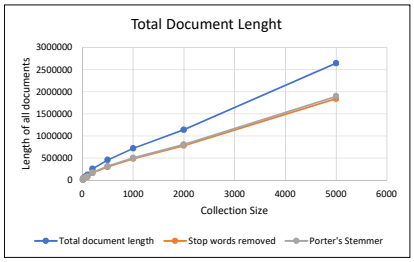


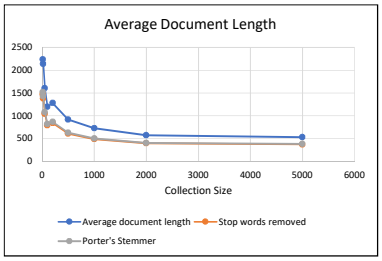
After removing the stop words the total document length decreases, but it almost doesn't affect the vocabulary size. It remains pretty much the same.

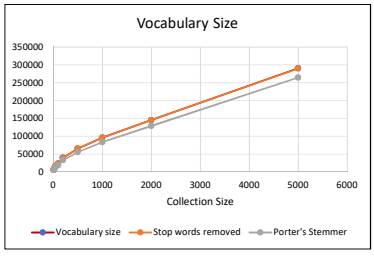
Exercise 4: Porter's Stemmer

Download a Porter's Stemmer (cf. list, lecture n°2)
Refresh the index of the 9th file of the exercise n°1, applying Porter's stemmer.
Compute again the statistics of the exercise n°2

	Collection size	Total document length	Average document length	Vocabulary size
Col 1-10	10	15211	1521	5043
Col 11-20	20	28491	1425	7863
Col 21-50	50	53627	1073	13692
Col 51-100	100	81394	814	18945
Col 101-200	200	172539	863	32905
Col 201-500	500	313693	627	55288
Col 501-1000	1000	502432	502	83290
Col 1001-2000	2000	807082	404	128745
Col 2001-5000	5000	1895289	379	264620







After applying the Porter's Stemmer the document size increases over removing the stop words, but the vocabulary size decreases significantly.