COMP3009J Information Retrieval

Worksheet 1

Before you begin, you should have installed Python 3 on your computer.

Download the file named "words.txt" from Brightspace (use the "Download" button on Brightspace so that it does not open in your browser).

In this worksheet, you will write a program to find all the words in this file and print them in order of their frequency (i.e. the most common word at the top). You will do this in 5 stages.

Q1. Write a file named "q1.py" that reads the contents of "words.txt" line-by-line (do not read the entire file at once), and prints each line to standard output.

Note: Each line in the file ends with a newline "\n" character. Also, the print() function adds a newline. Remove the newline from each line before printing.

Hint: The methods available for a Python string are in the manual here:
https://docs.python.org/3/library/stdtypes.html#text-sequence-type-str

Q2. Copy "q1.py" to a file named "q2.py" and modify it so that it divides each line into words, and prints the words with parentheses (round brackets) around them.

For example, the first line should look like this:
```
(iS)(ansWEr)(thiS)(IS)(iS)
```

Hint: Because the print() statement adds a newline by default, you will need to provide an extra parameter. The manual page is here:
https://docs.python.org/3/library/functions.html#print

Q3. Copy "q2.py" to a file named "q3.py" and modify it so that it prints the lowercase version of each word instead.

For example, the first line should look like this:
```
(is)(answer)(this)(is)(is)
```

Hint: One of the standard string methods from the hint in Q1 might help.

Q4. Copy “q3.py” to a file named “q4.py” and modify it so that it counts the number of times each word appears in the text. For each word, print the word and the number of times it appears.

Sample output (the order of your output might be different):

```
--------------------
2048: is
128: answer
4096: this
...
--------------------
```

Hint: You will need to choose an appropriate data structure for this task.

Q5. Copy “q4.py” to a file named “q5.py” and modify it so that it prints the words (and their frequencies) in order of frequency. The most common word should be first, followed by the others in descending order.

Advanced

If you finish the above exercises early, here is something else you can try.

Q6. The file preprocessed_documents.txt contains some documents that have undergone some preprocessing (removal of punctuation and stopwords, stemming using Porter’s stemmer). Each document begins with its unique document ID and the terms in the document are then separated by spaces on the same line.

Create an inverted index of this corpus, using suitable built-in Python data structures.