**Assignment 1**

**Group 37**

**Hritika Shah MT22101**

**Gnanesh Patel MT22030**

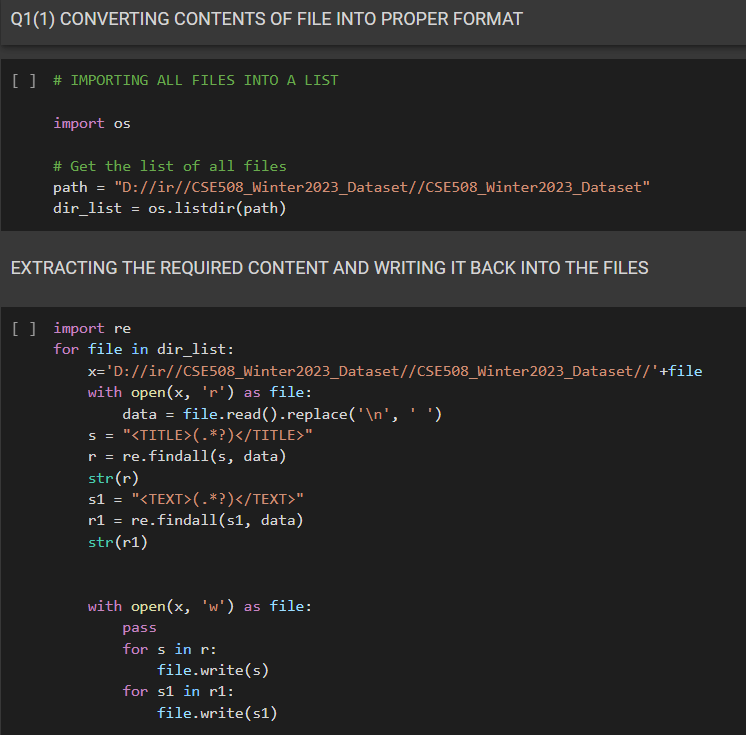
**Nikita Kapoor 2020531**

All the questions done in this assignment were an equal collaborative effort from the whole group.

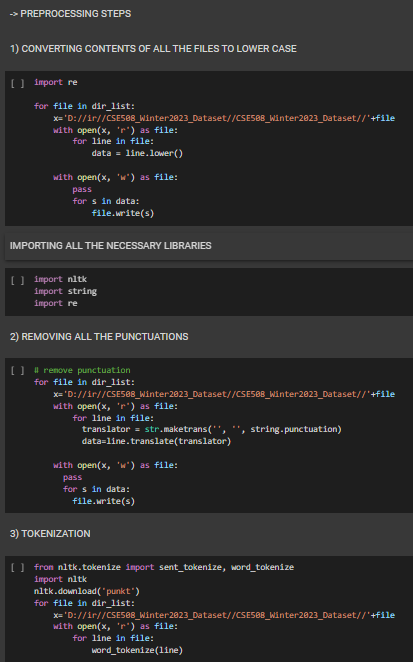
Report of the questions is provided below.

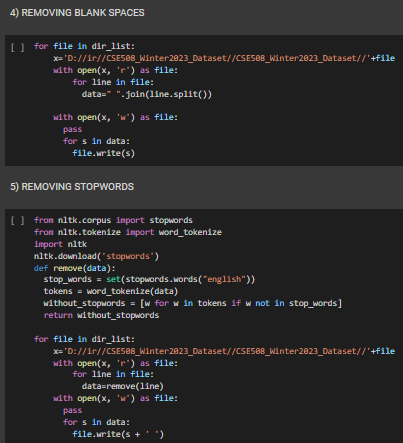
Question 1:

(i) The contents of the files was converted into the right format for pre-processing and subsequently the relevant text extraction was done.



(ii) The pre-processing steps were performed and compiled in a well documented format. The necessary libraries were imported for the same following which the code was executed and all the files from the given dataset were updated as per the requirements of the question.



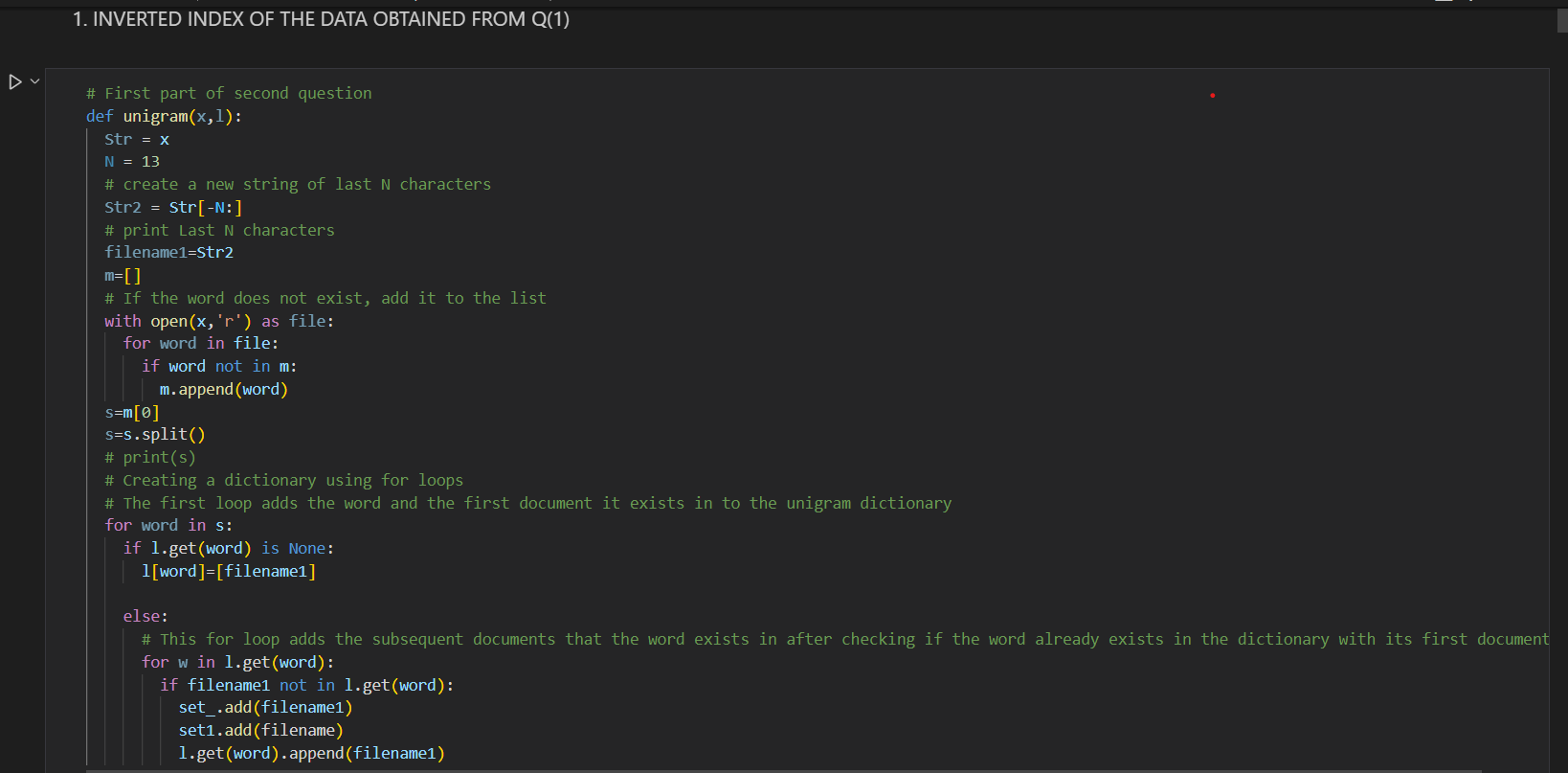


Following this, we attached screenshots of 5 sample files before and after each of the pre-processing steps. They were also printed on the terminal. The compilation of this information can be found here containing all the relevant information and code snippets:

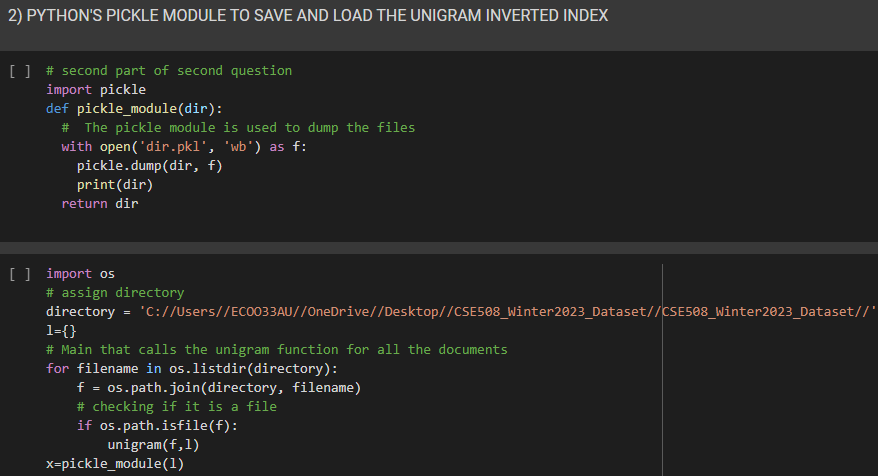
<https://docs.google.com/document/d/1J-0gTiI02hvulLIt0rC_Ds8oUMAhVGyD-rxQDb94occ/edit?usp=sharing>

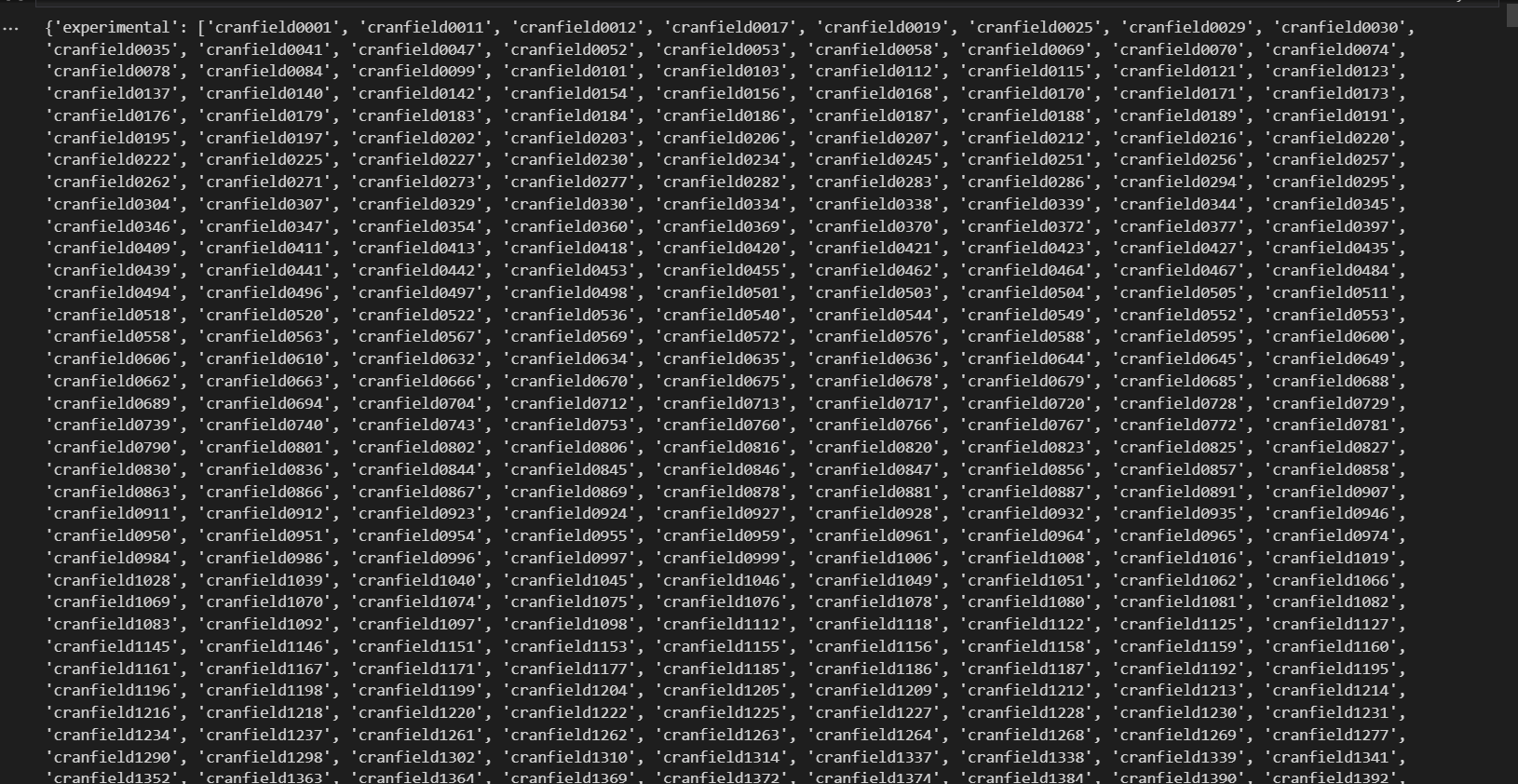
Question2:

1. A unigram inverted index was created from scratch using the help of dictionaries and lists in python for the dataset obtained in question1.



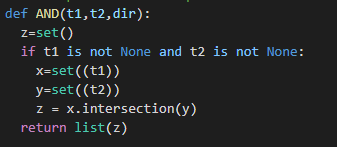
2. Using the pickle module in python, the unigram inverted index was saved and loaded.



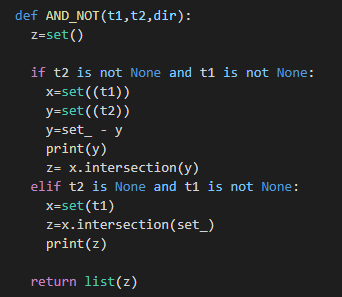


3. Support was provided for the operations as per the question. All edge cases depending on whether of not T1 and T2 exist in the unigram inverted index have been covered.

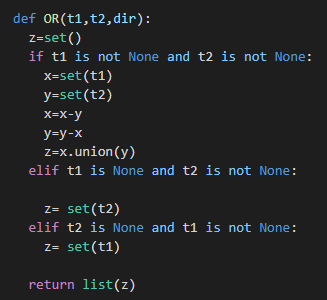
1. T1 AND T2



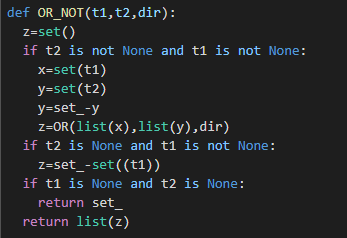
1. T1 AND NOT T2

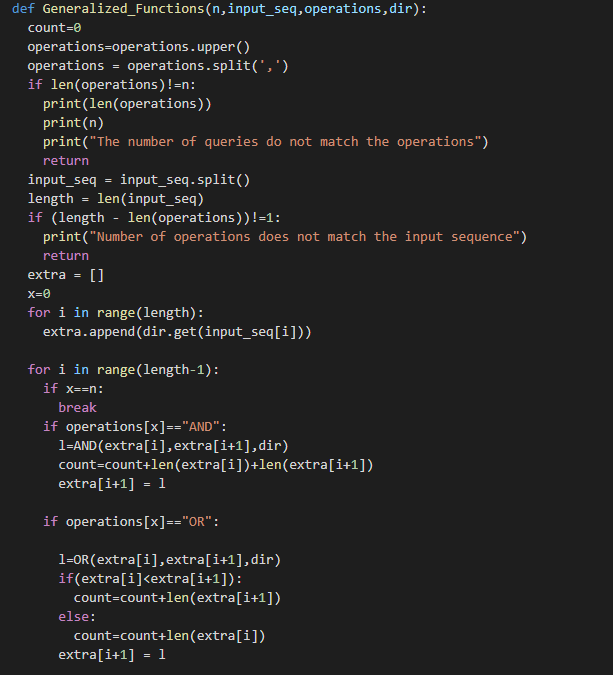


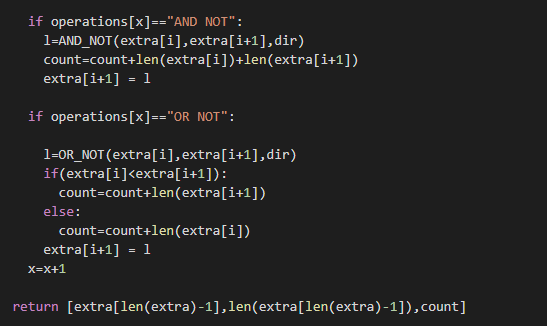
1. T1 OR T2



1. T1 OR NOT T2

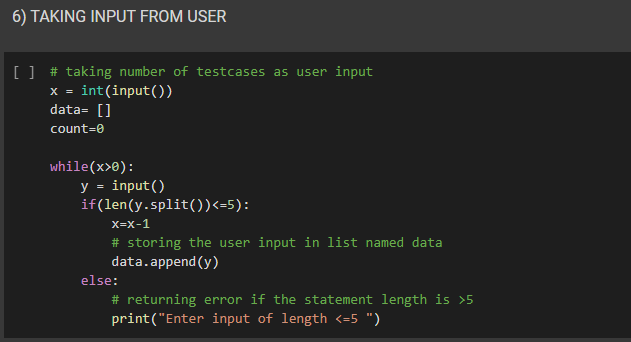


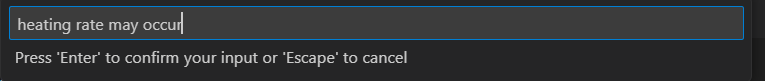
4,5. A generalized function was created for the queries. Number of comparisons done to execute the queries have been incorporated into the same function

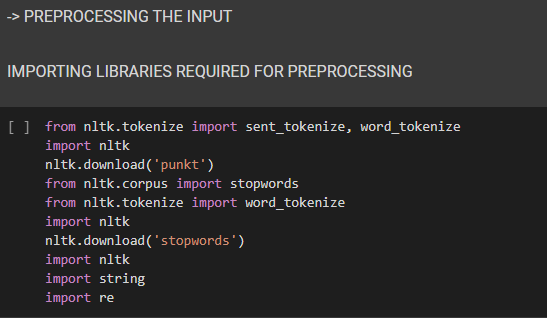


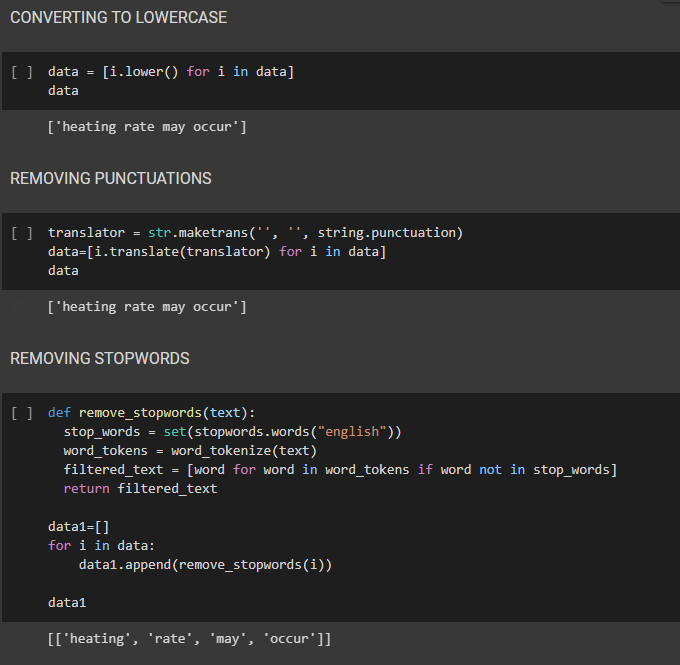
Number of comparisons is returned.

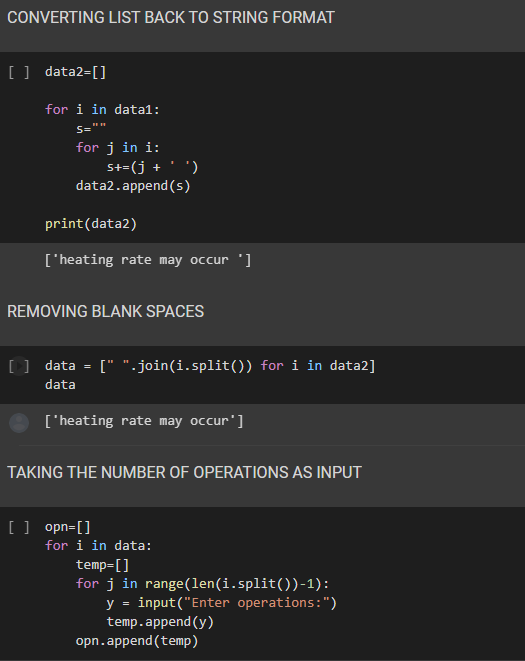
6-10. These parts have been completed taking the input and providing the output as required in the question. Appropriate format has been followed. Pre-processing steps have been performed on the input taken from the user as well.

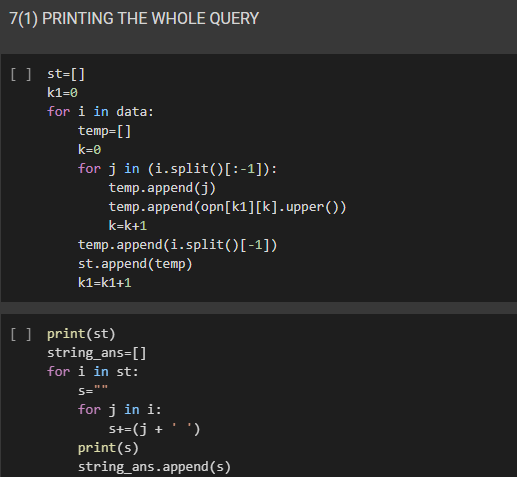


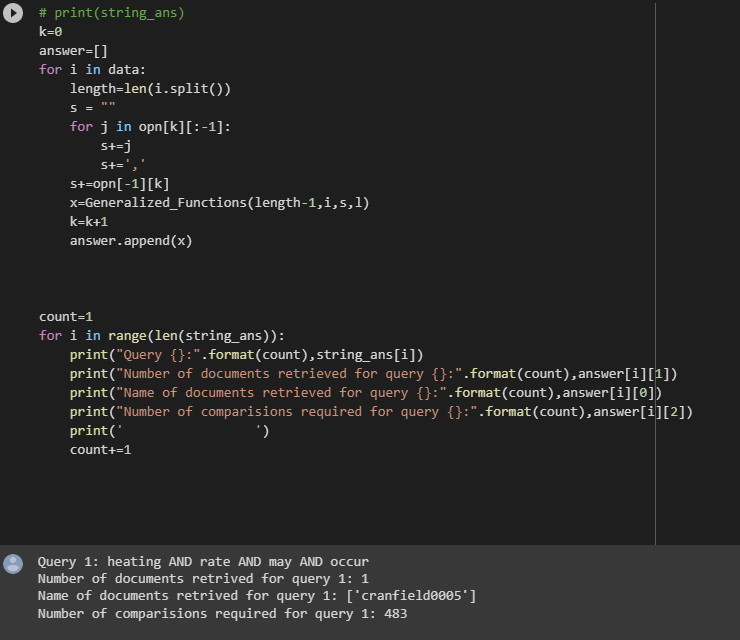






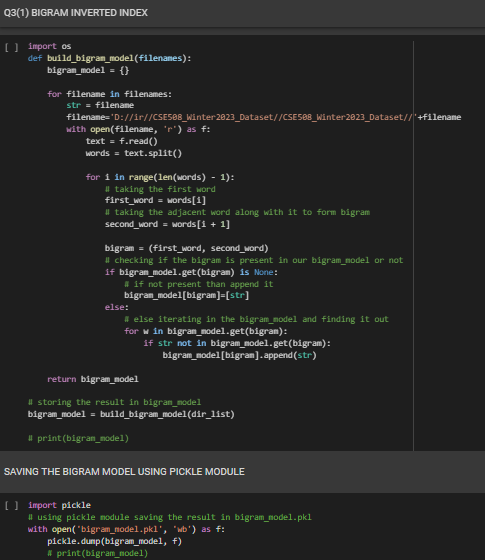




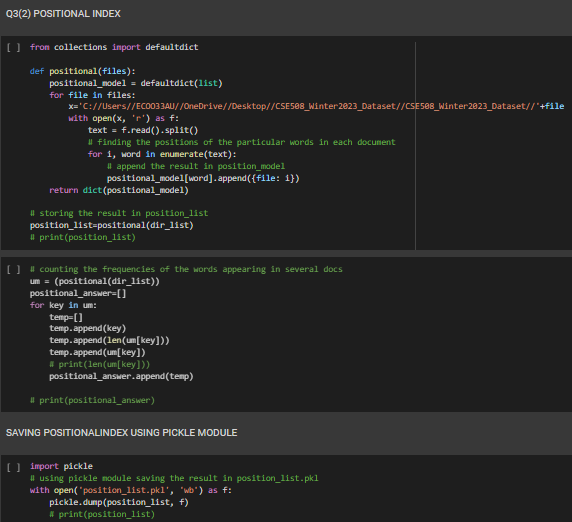


Question3:

(i) In this question, a bigram inverted index was created from scratch of the dataset that was obtained from the first question following which Python’s pickle module was used to save and load the bigram inverted index.





(ii) A positional index was created of the obtained dataset following which it was saved and loaded using the Python pickle module.

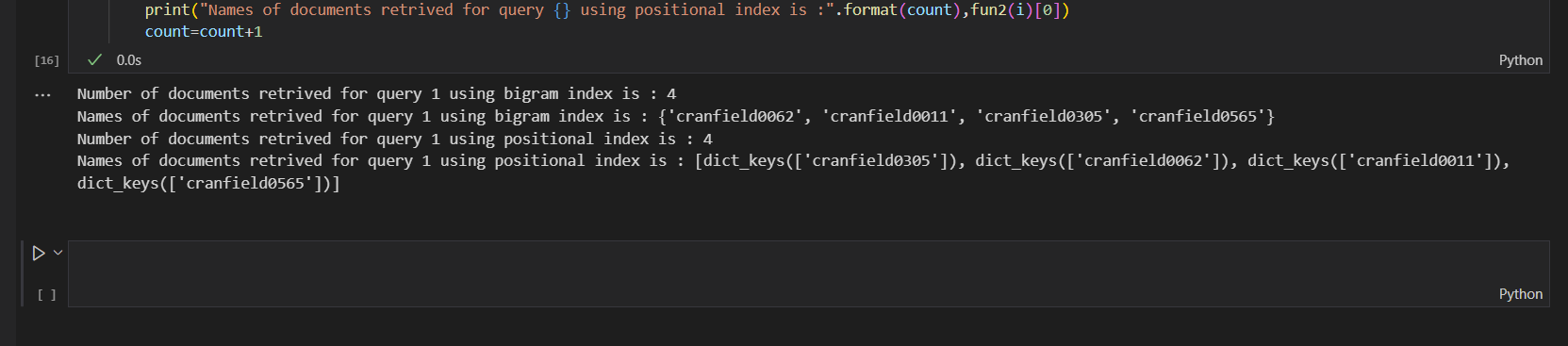


(iii) Upon comparing the results from (i) and (ii) :

-Result 1 is a bigram index query so it will give the documents as a result even if it does not contain the whole phrase together in the document.

-Result 2 is positional listing phrase query. It will show the results that contain the whole phrase together in a document.

-In the second method we are even checking the position of words to check if it is a continuous phrase in a document or not.



Following the above, the required input/output format was followed and the pre-processing steps were followed as well:

