**FOBDA**

**Assignment # 02**

**Muhammad Awais**

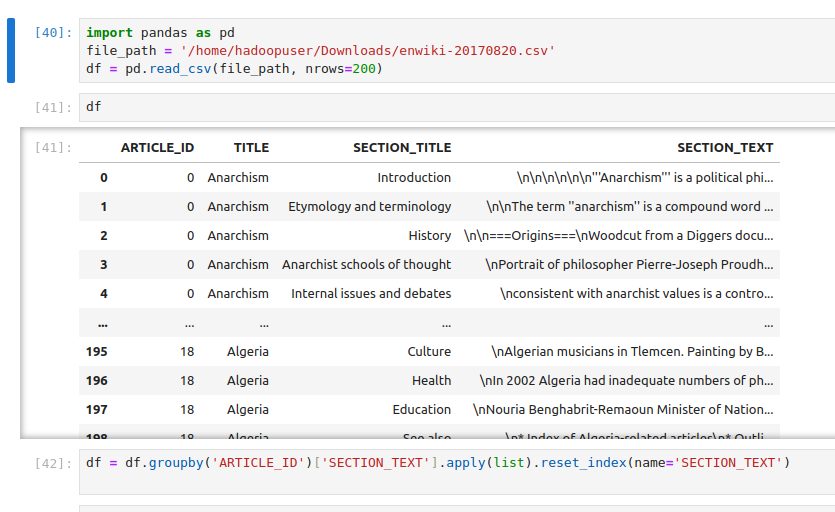
**Malaika Azhar**

**Muhammad hadi**

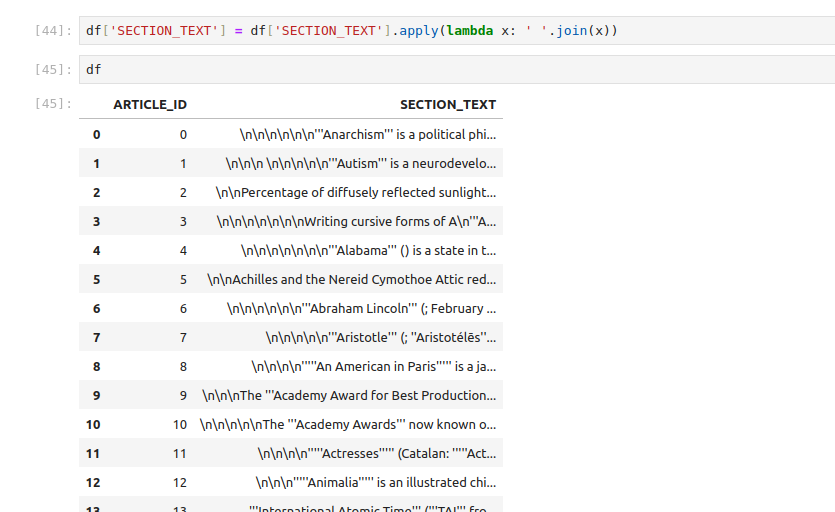
***Detailed report:***

**Dataset PrepRocessing:**

**Image # 01:**



Loading data and then concatenating string of all article IDs which are separated on the basis of topic 🙂 We removed all remaining columns from dataset because we are not interested in them.



Now

we are performing the lower() method to make all text lower and lemmatization to map to root words after that we have the final result like this.



Now saving into a text file to proceed further



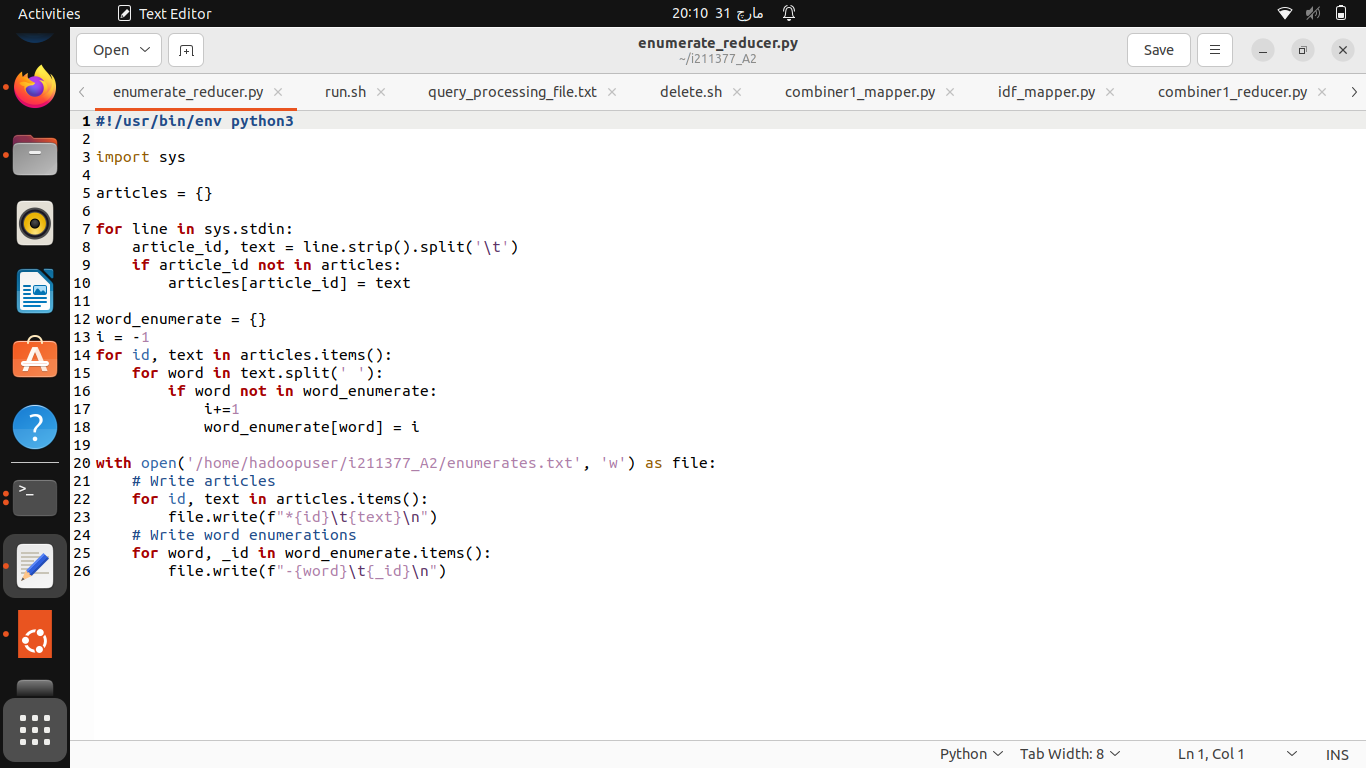
**Indexer part:**

1. ***Word ENumerates:***

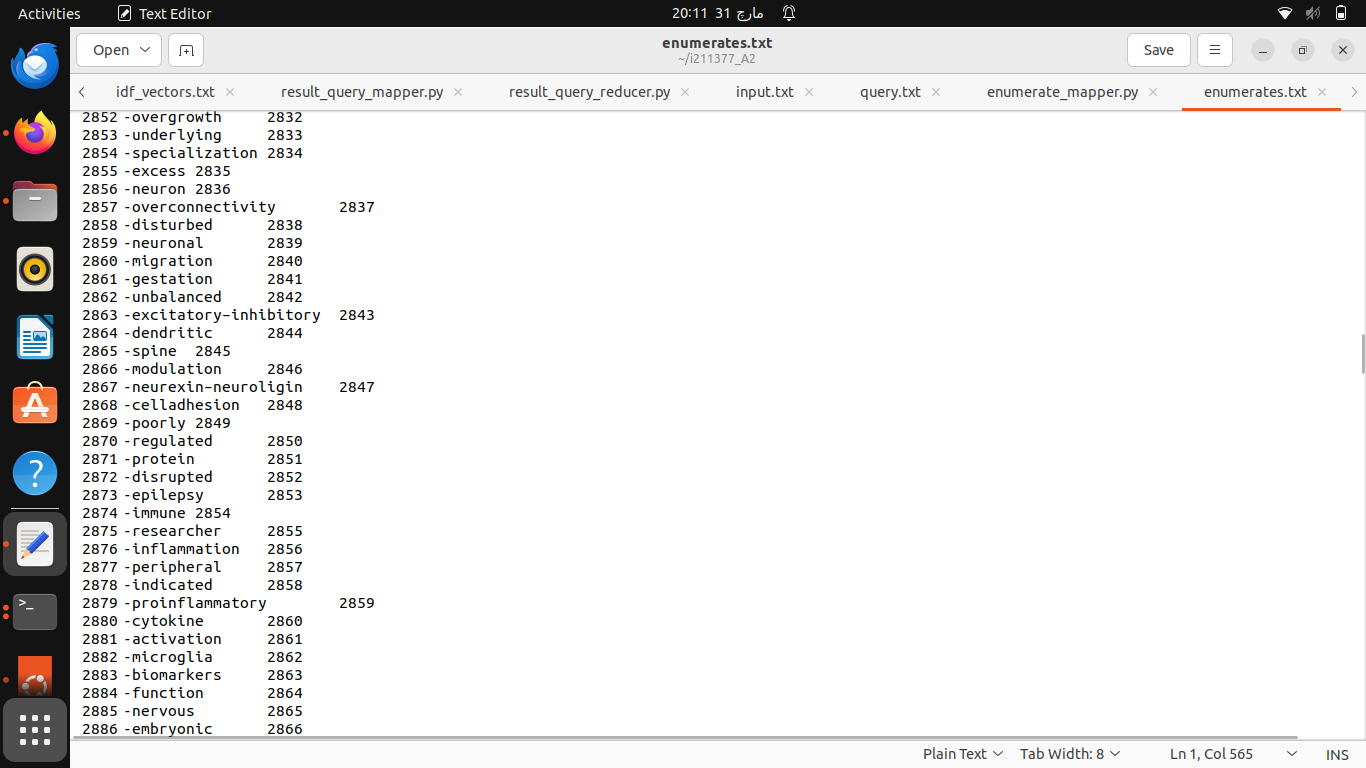
Running mapper and reducer to find word enumerates (unique code for each word):

mapper.py  
  


reducer.py



Output:



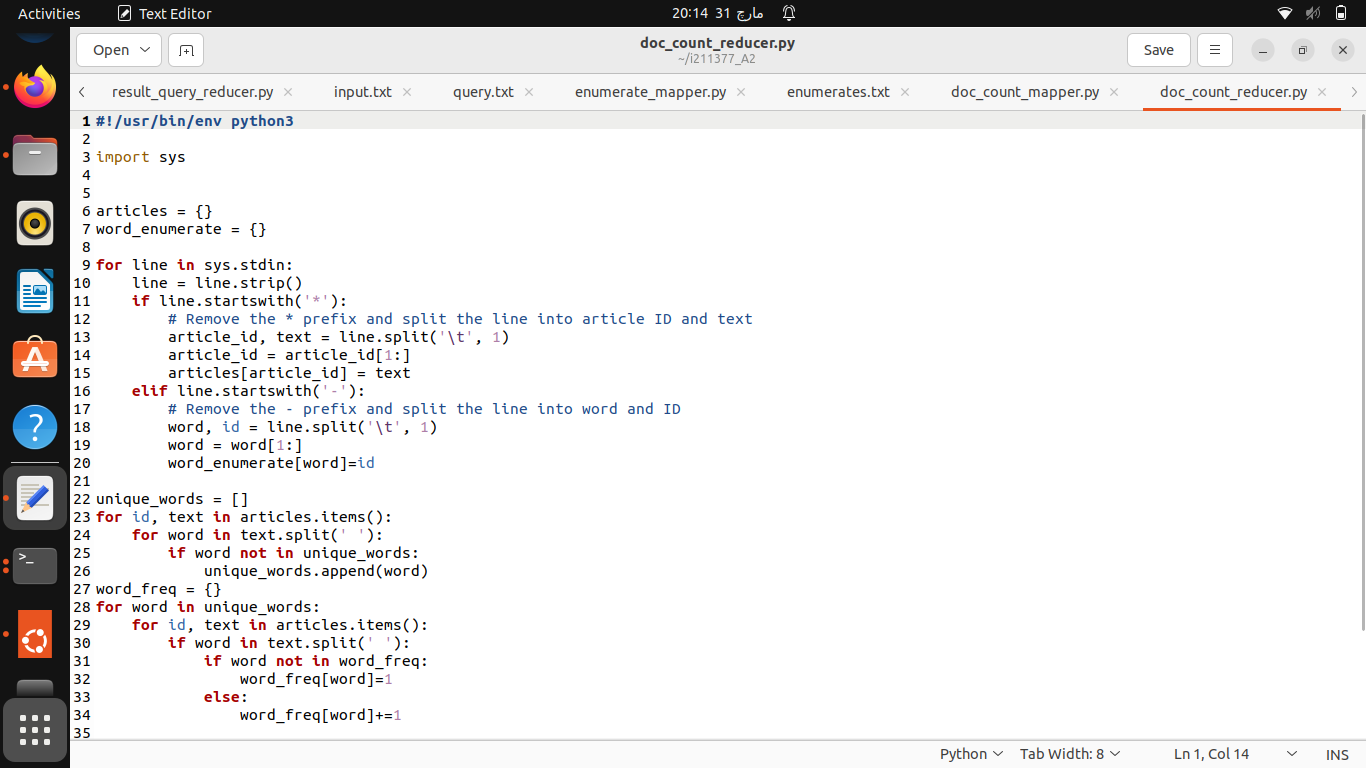
1. ***Document count for each unique term:***

Here we are checking i how many documents each unique word is occuring:

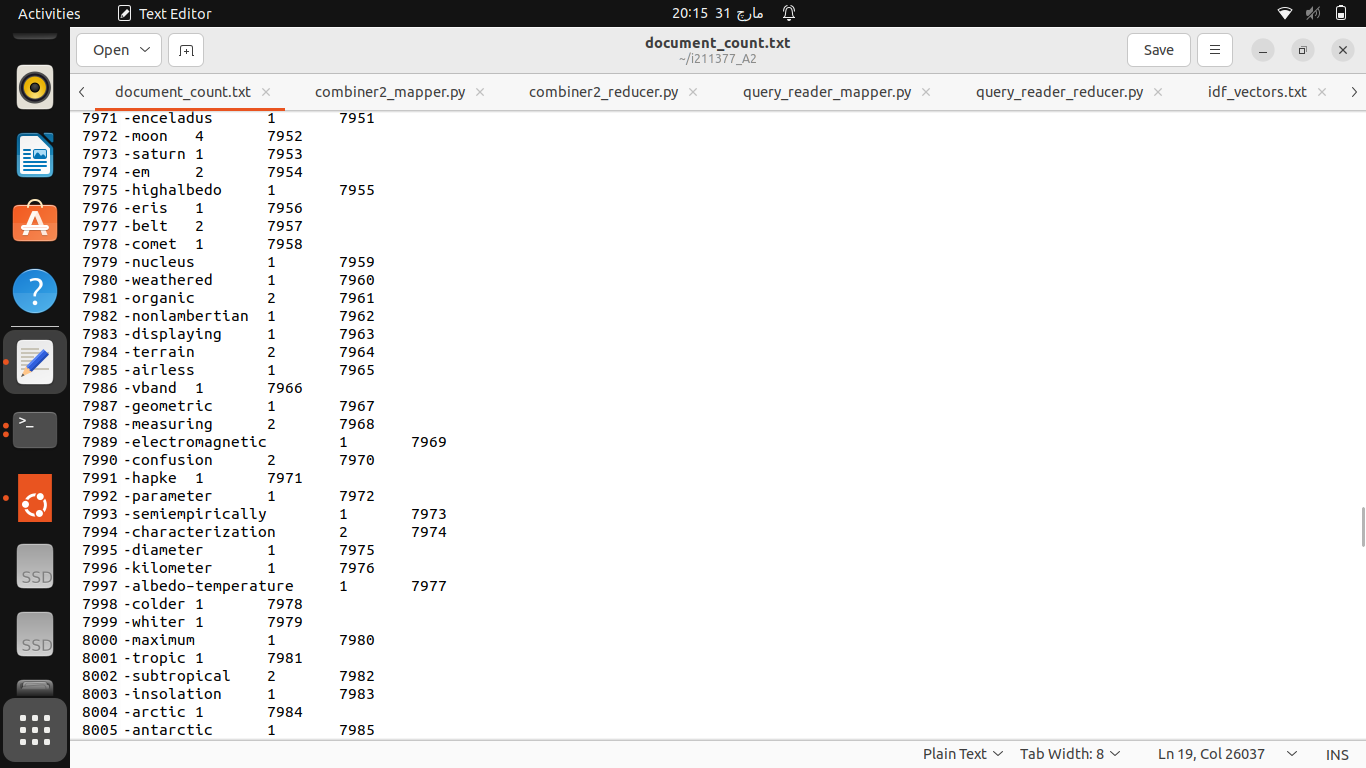
mapper.py



reducer.py



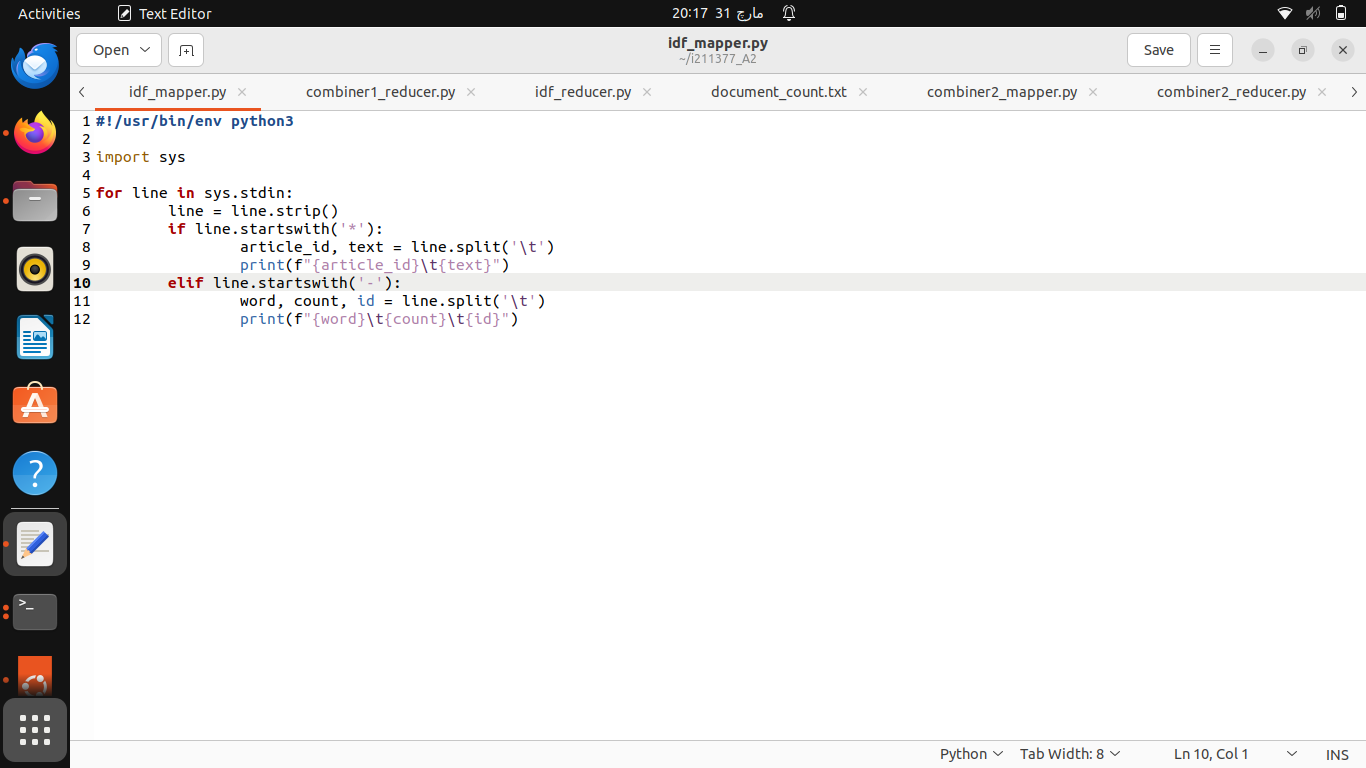
Output:



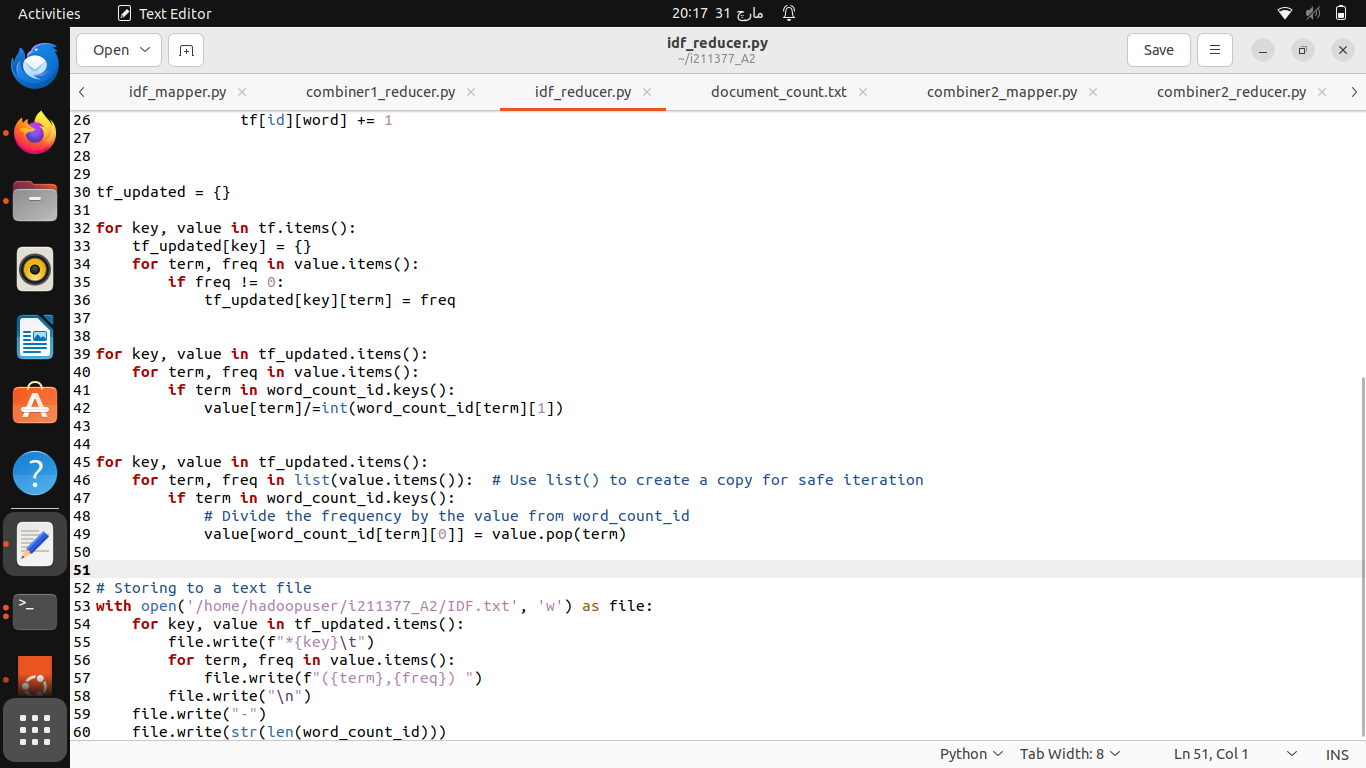
1. ***FINDING IDF :***

Here we are calculating IDF for the non zero values in all documents:

mapper.py:



reducer.py:

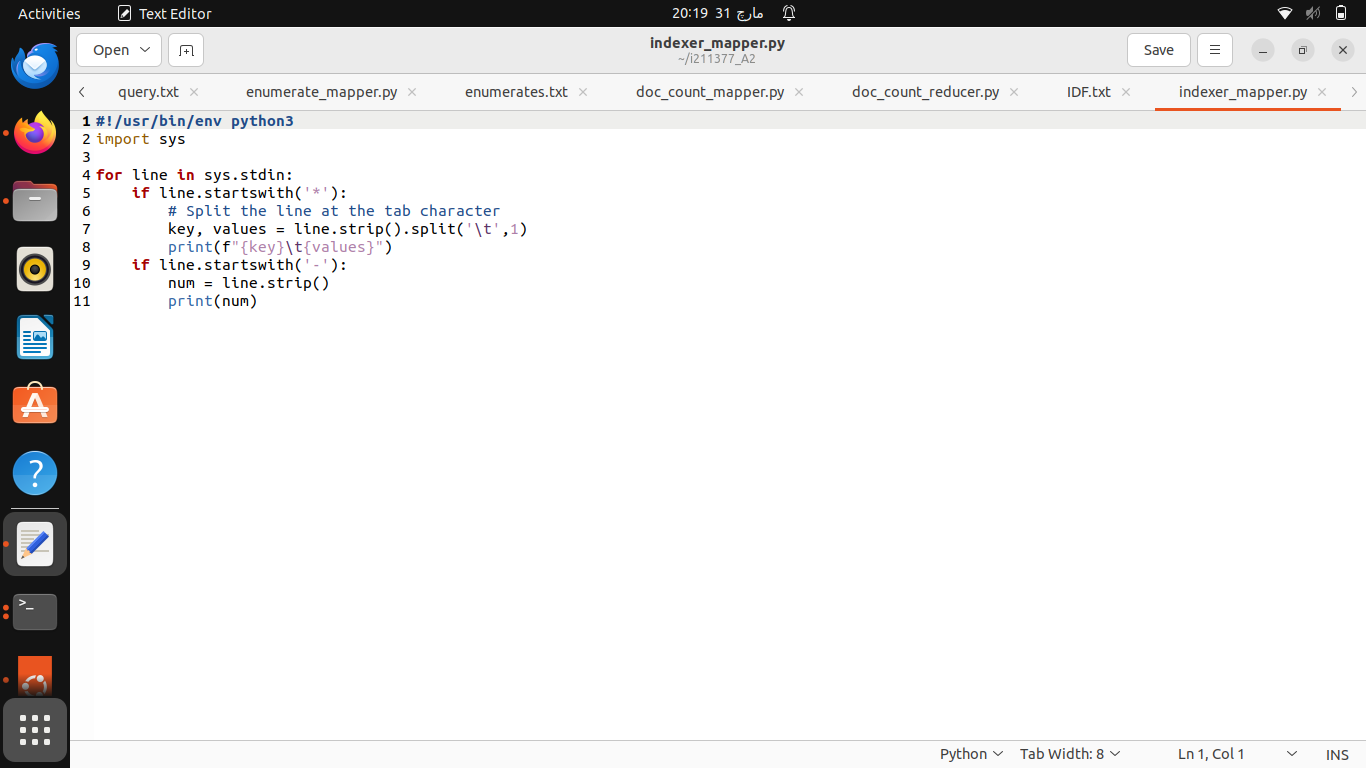


Output:

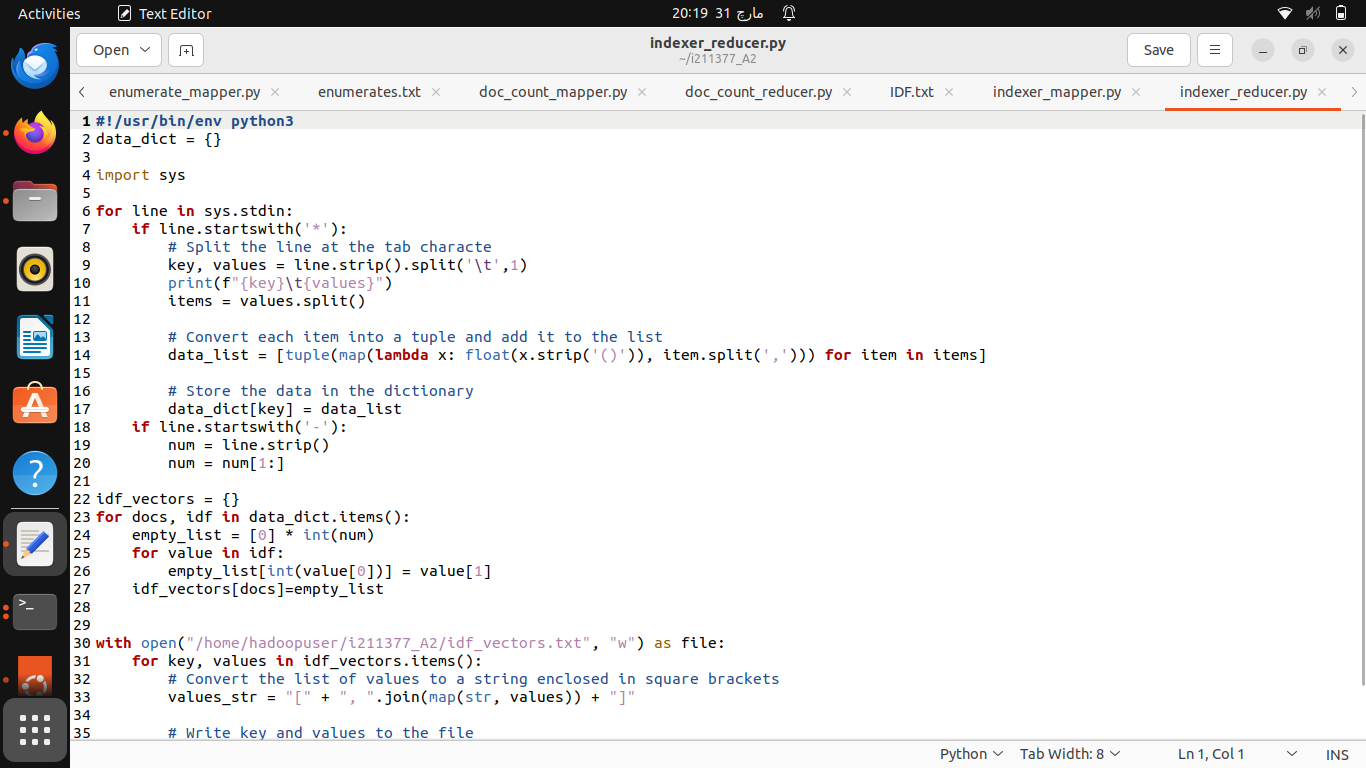


1. ***Indexer (Creating idf\_vectors for all docs)***

mapper.py:



reducer.py:



Output:



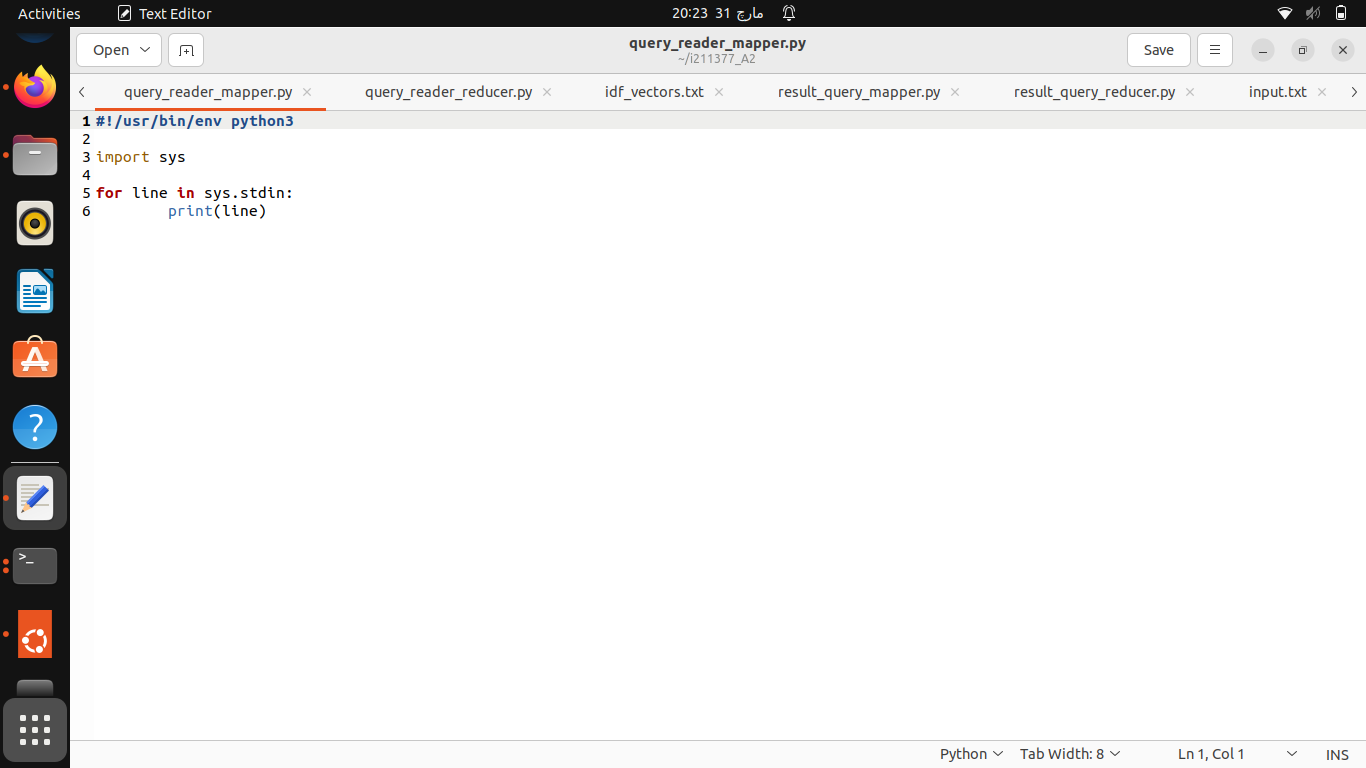
—------------------------------------------------------------------------------------------------------------------

**Query part:**

1. ***Query Reader map reduce job***

We have a file named as query.txt in which we wrote a query ;

mapper.py:

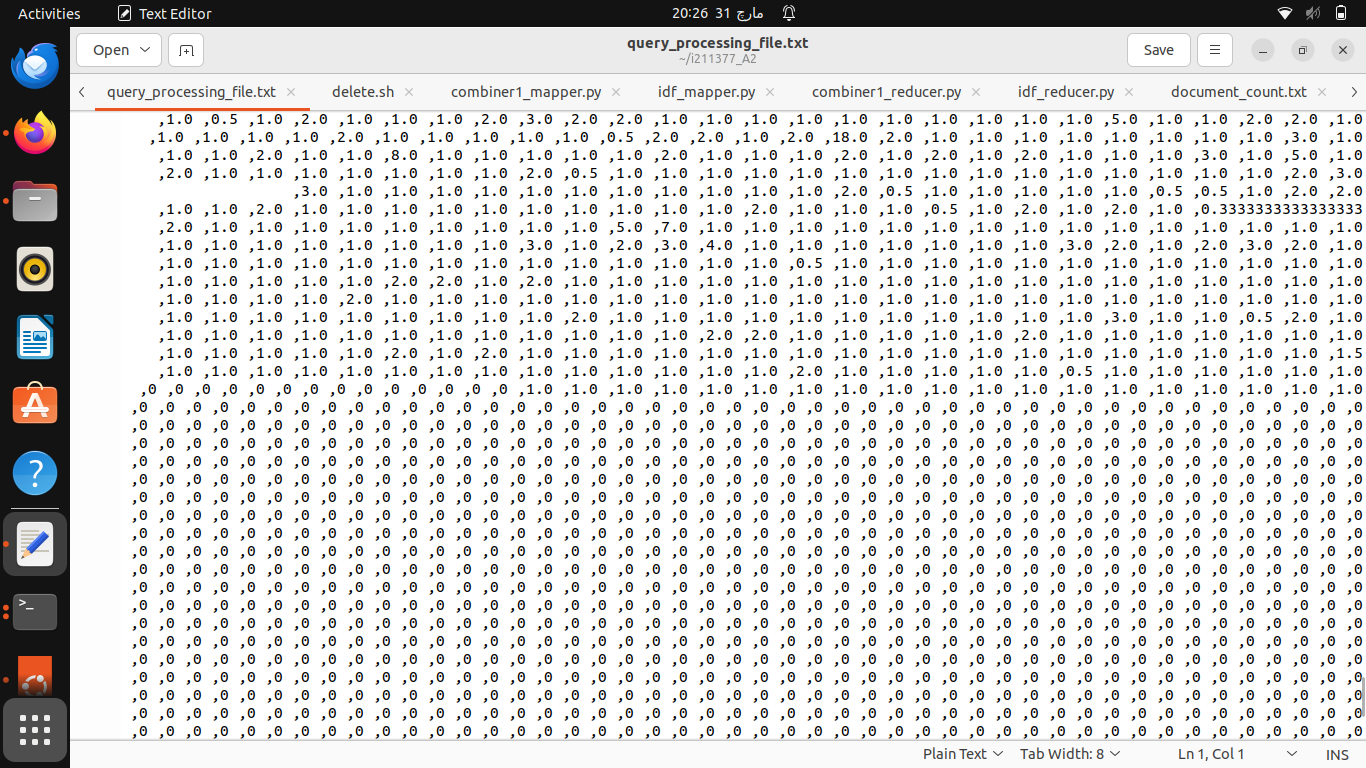


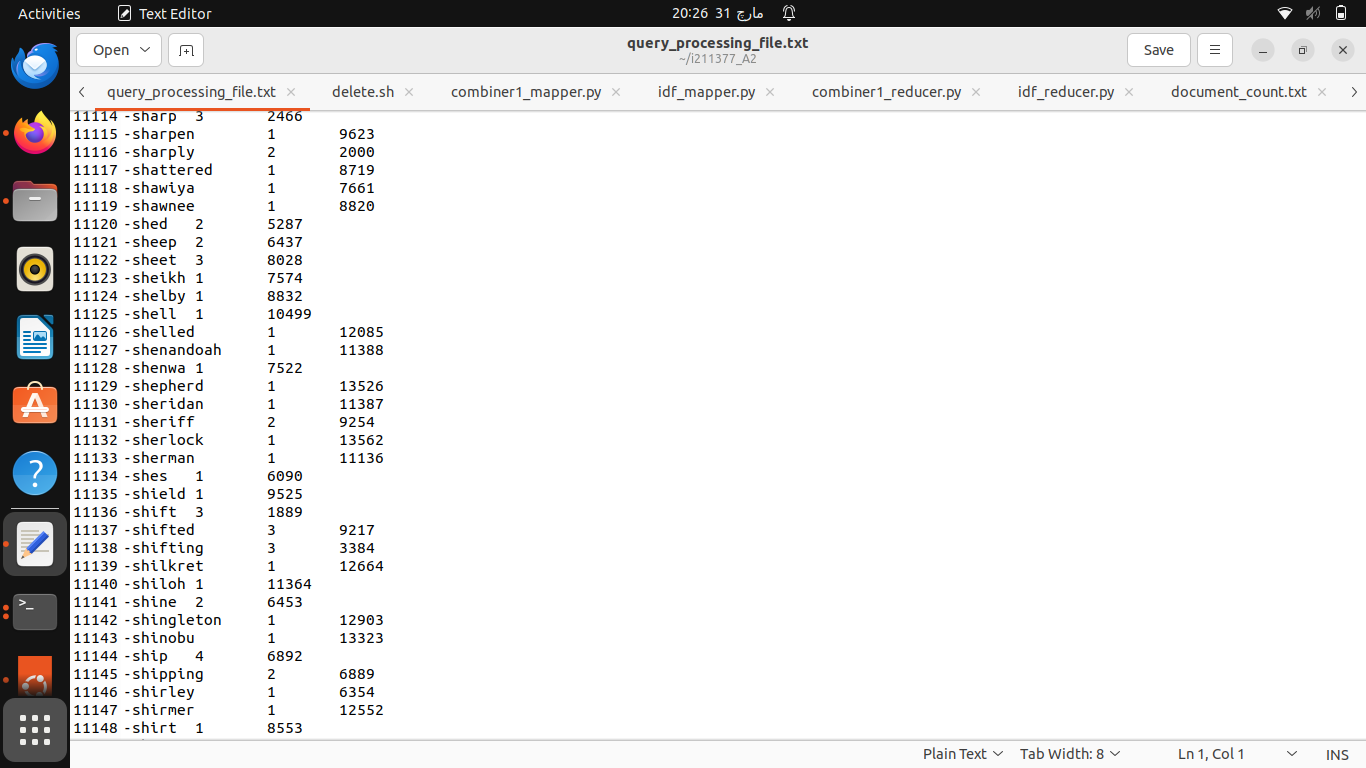
reducer.py:



1. ***COMBINER 1 and 2 which combines the values we need nito a same text file (query\_preprocessing\_file.txt)***

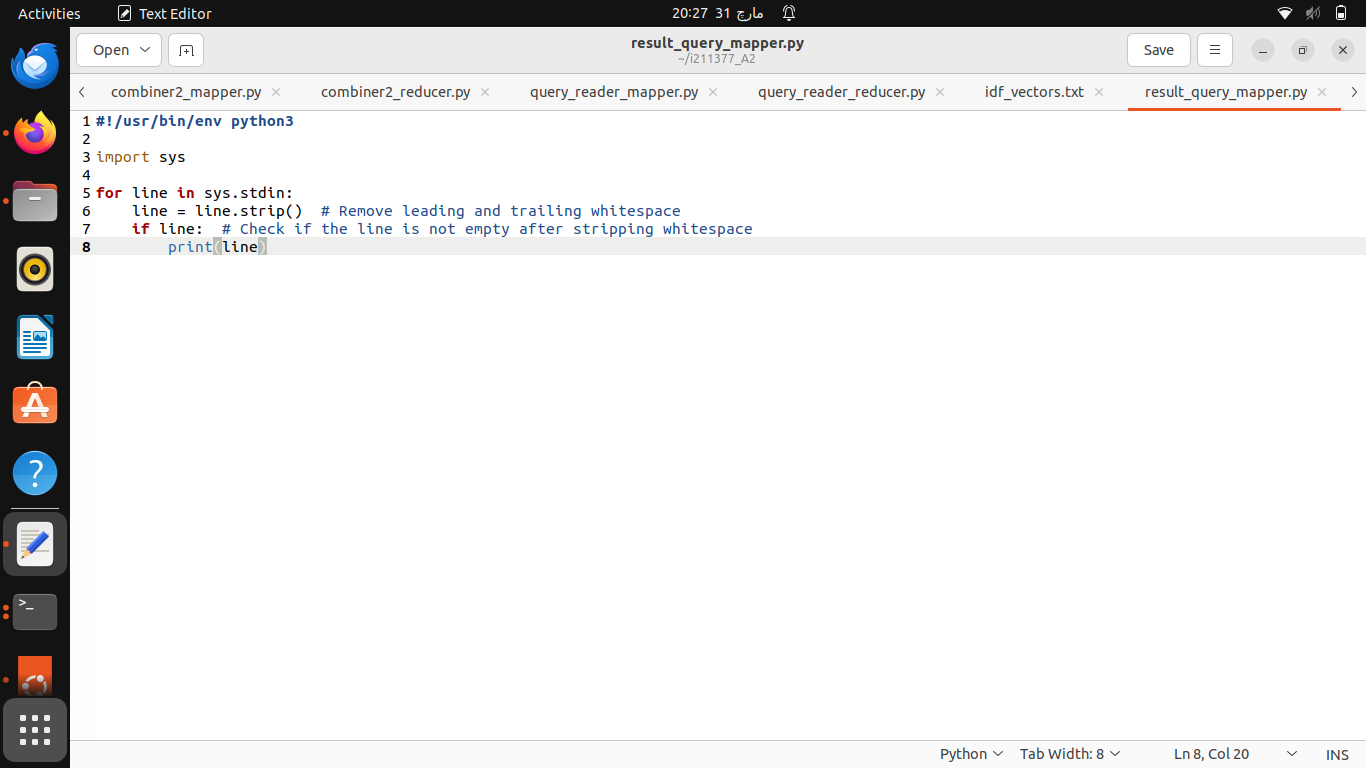
Output:



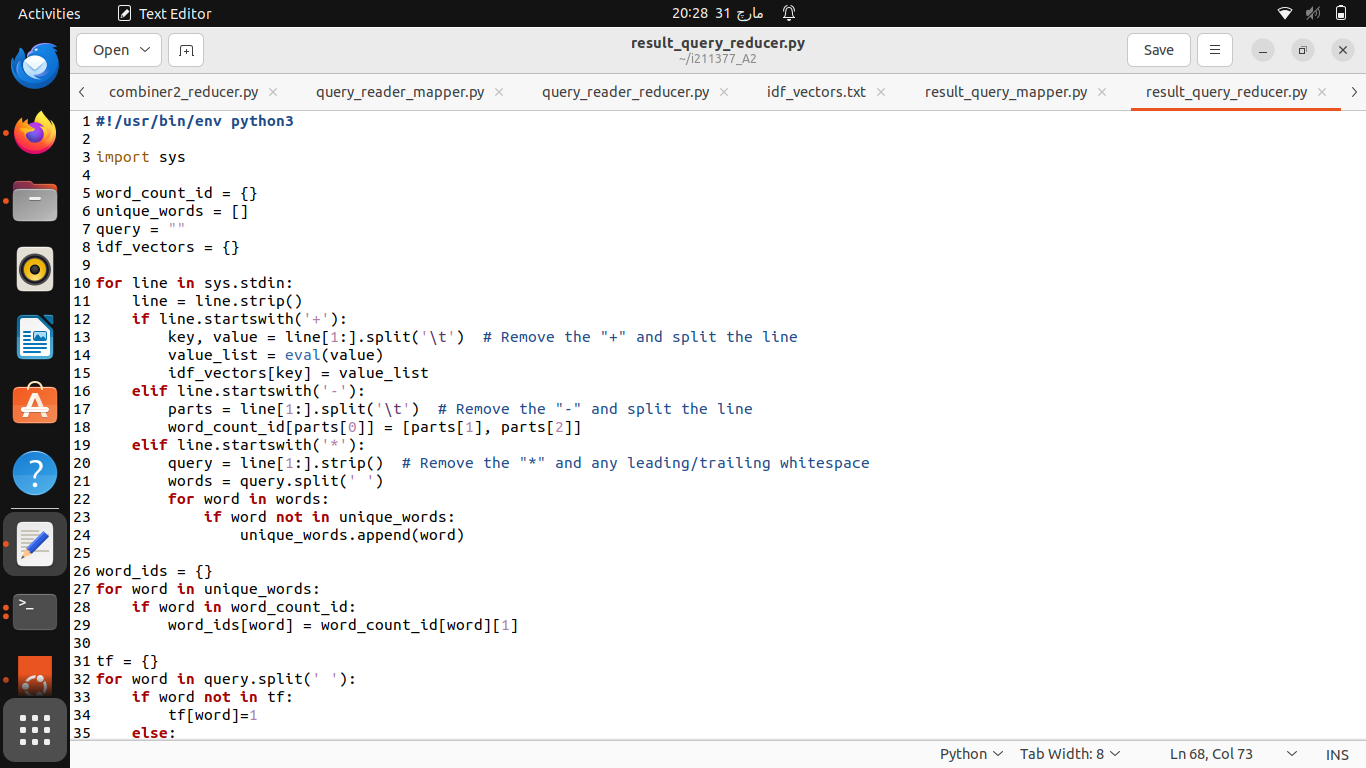


1. ***Query vectorizer and the cross product finder***

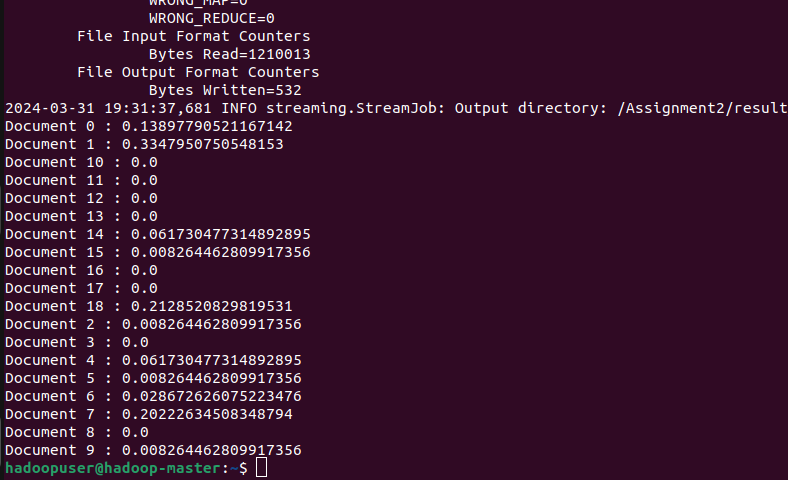
mapper.py:



reducer.py:



Our final output after having cross product with all docs is:



clustering:

