Student Number: 10352221

Module Title: Programming for Big Data

Lecturer Name: Darren Redmond

Module Code: B8IT105

Assignment Title: Perform Analysis on a 5000-line dataset

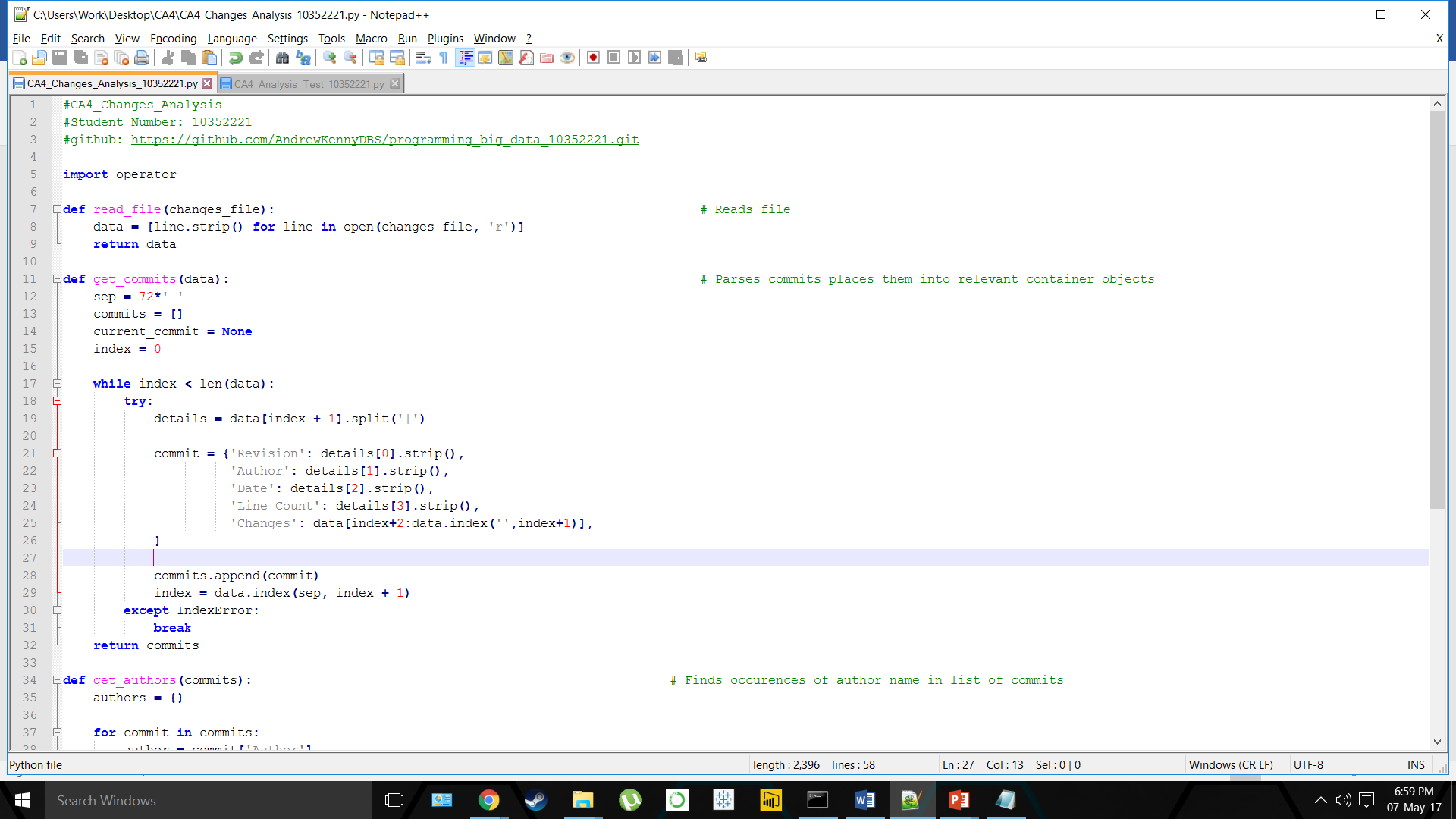
# Code for Calculating Analysis

The dataset chosen for investigation was *change\_python.log*, a log of commits which include a revision number, the author name, date and time of change, number of lines, changed paths and author comment for each. The three items chosen to investigate in this report include:

1. The total number of commits made
2. The total number of authors
3. The number of commits made by author

## **Total Number of Commits**

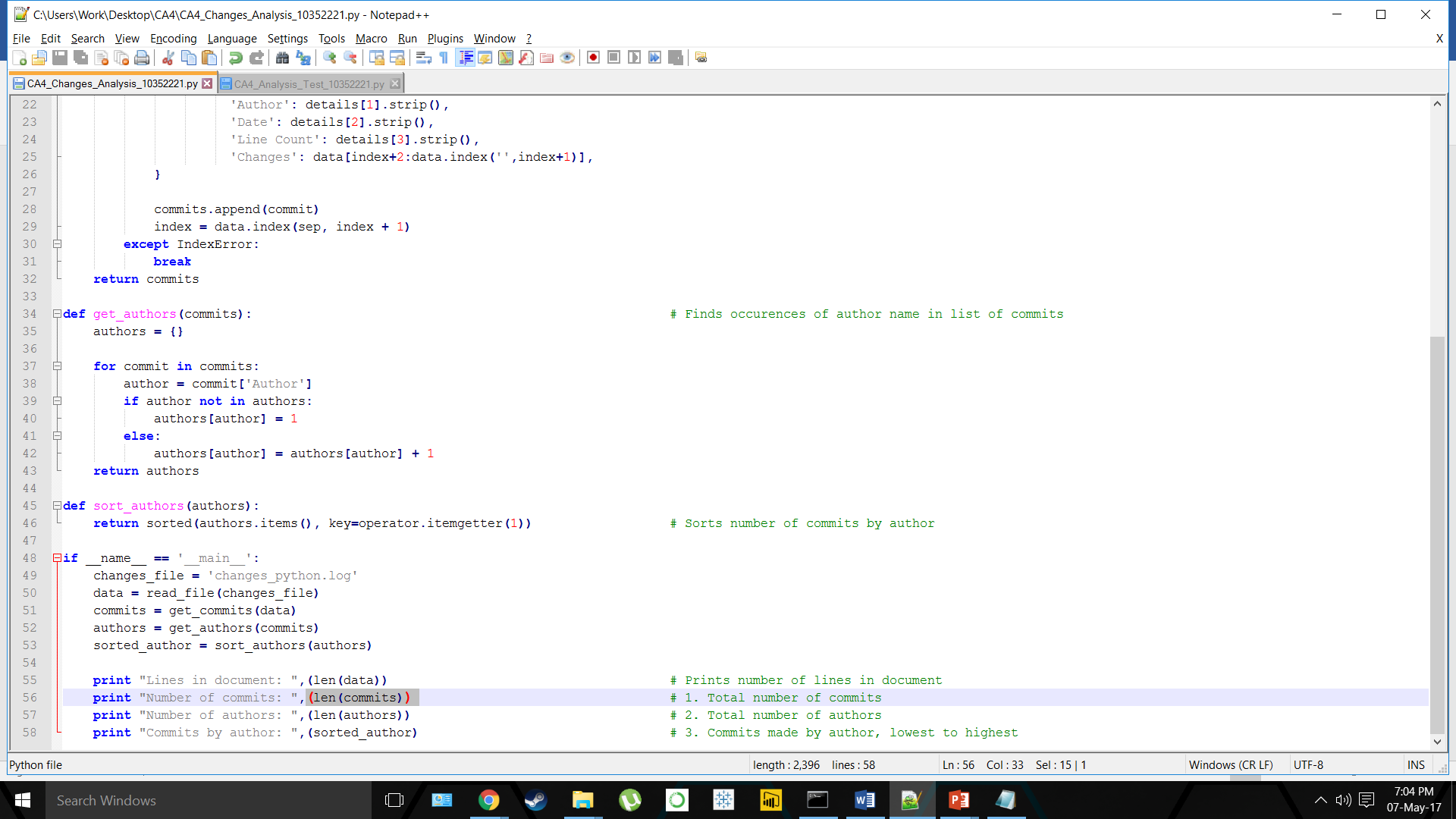
To find the total number of commits in *change\_python.log,* code is written to read the file and separate it on to new lines. Next the data is divided under the headings of *'Revision', 'Author,' 'Date', 'Line Count'* and *'Changes'* and then separated into individual commits. To find the total number of commits, the *len* function is used.



*Figure 1.1 Parsing data in change\_python.log*

## **Total Number of Authors**

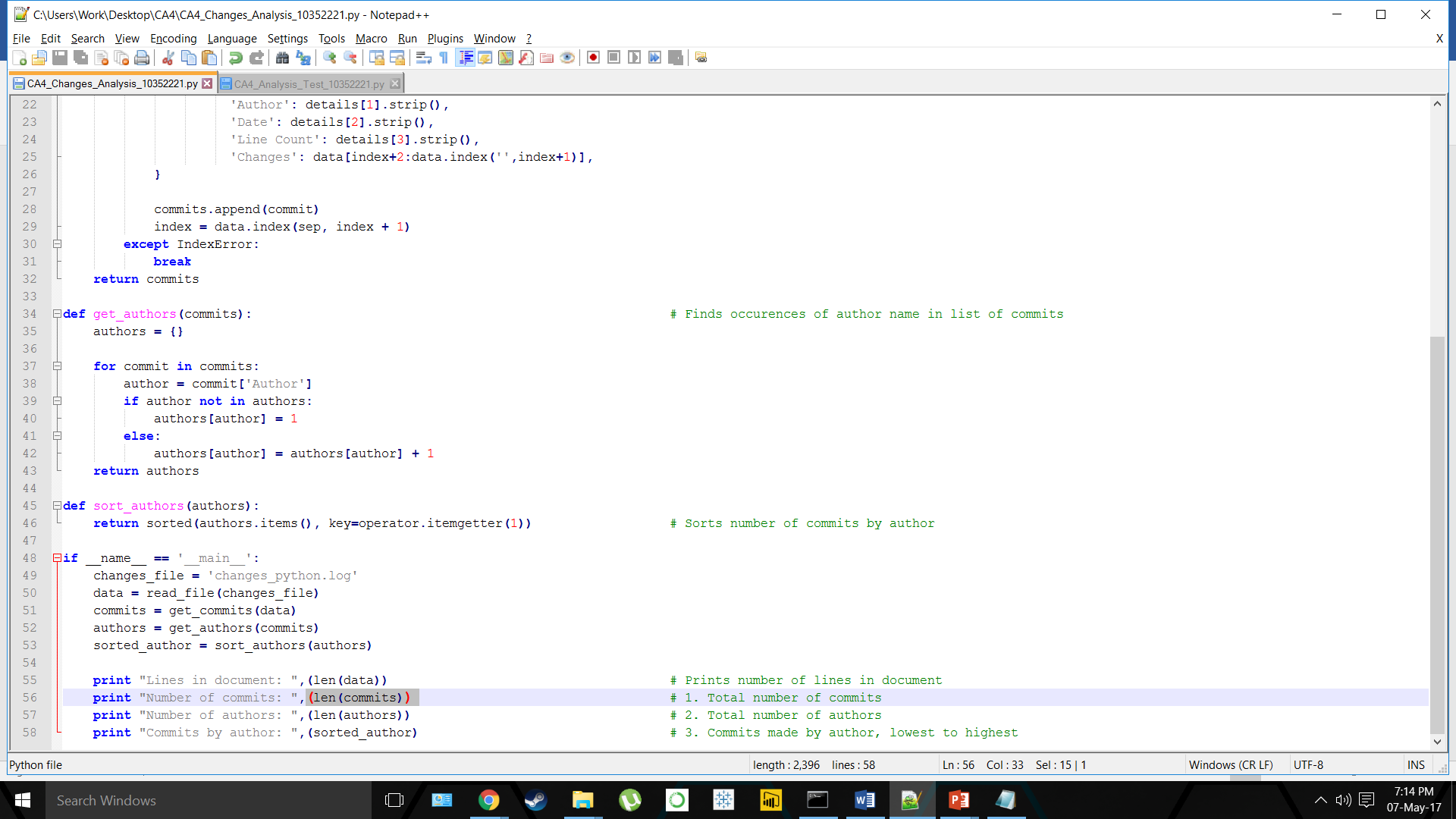
*get\_authors* is defined which searches through the commits to find names of authors, counts the instances and then stores the results. The *len* function is then used again to find the total number of authors.



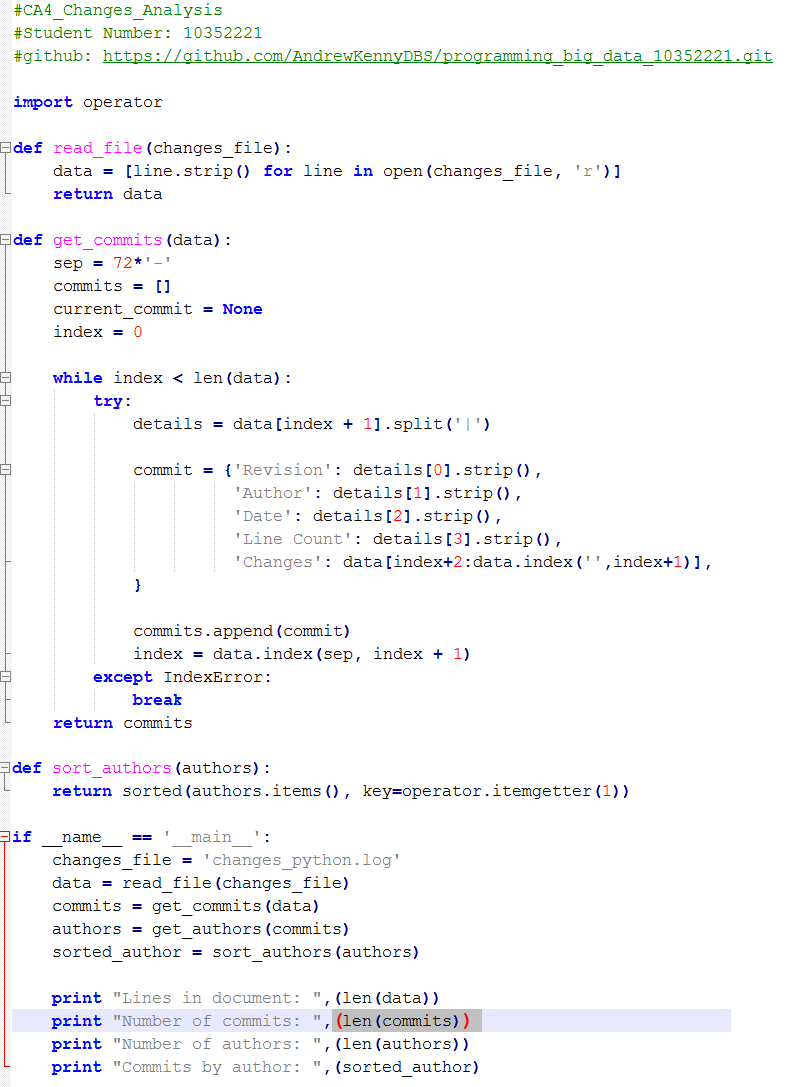
*Figure 2.1 Finding authors in change\_python.log*

## **Total Number Commits by Author**

To see the number of commits made by author, python’s built-in *sort* method is used, having imported *operator* earlier in the code. This arranges the author in order of number of commits, from lowest to highest.



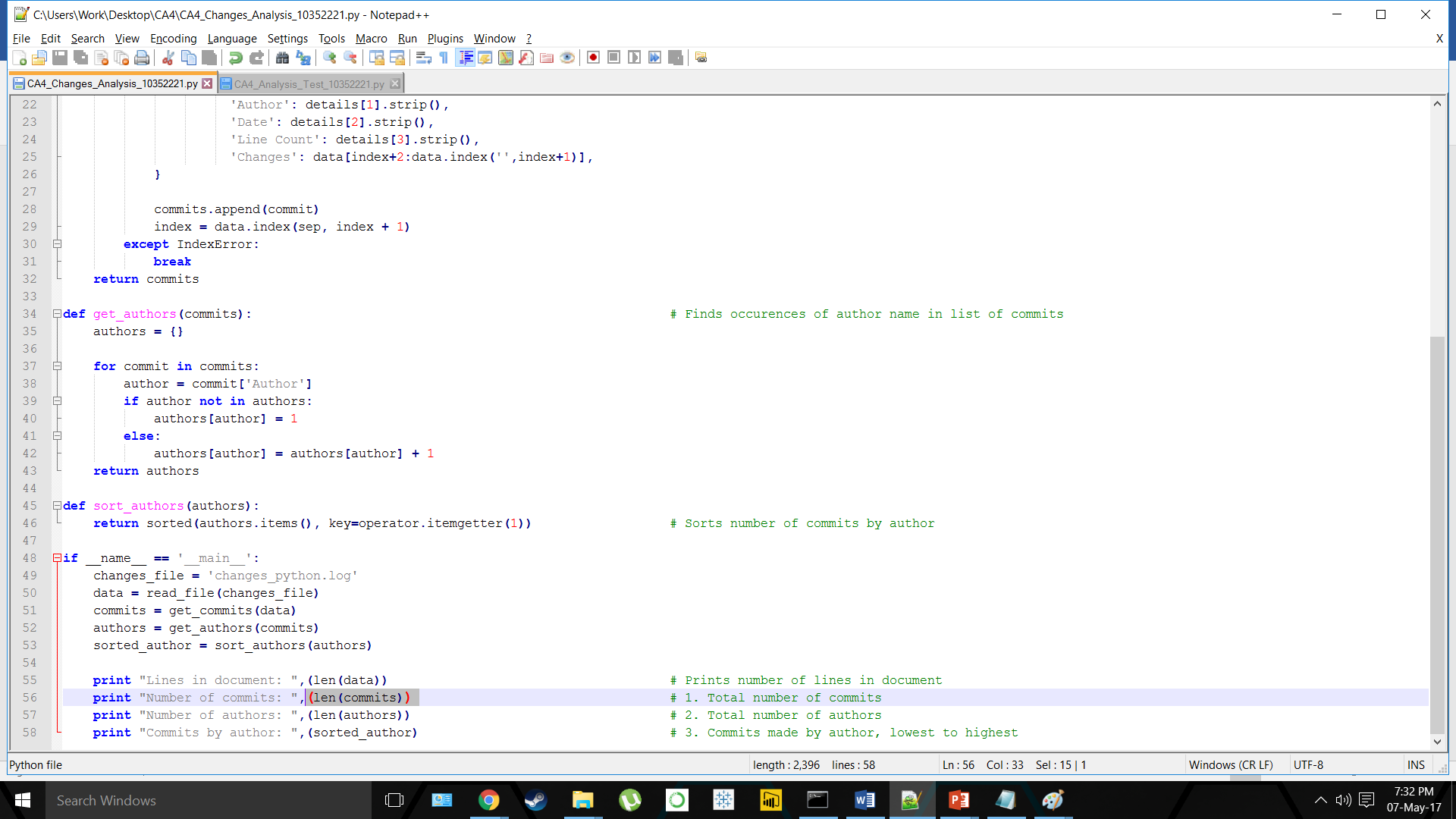
*Figure 3.1 Sorting by commits made per author*

**

*Figure 3.2- All code (without comments)*

# Analysis Results & Conclusions

New code is implemented to print the results of the earlier code (*Figure 4.1*), which can be seen in *Figure 4.2* below. This code is then tested to ensure the correct results.



*Figure 4.1- Code to print results of Commits analysis*



*Figure 4.2- Summary of printed results*

From the analysis undertaken to find 3 interesting statistical pieces of information for this dataset, the results show that:

1. A total of 422 commits have been made
2. A total of 10 authors were involved in these commits
3. The number of commits made by author (from lowest to highest) found is: *murari.krishnan (1), Dave (2), Alan (5), Nicky (5), Freddie (7), ajon0002 (9), Vincent (26), Jimmy (152) and Thomas (191).*

However, due to an error in the code, a user named */OU=Domain Control Validated/CN=svn.company.net* is also returned in the results with 24 commits, skewing the overall findings.

Despite this error, the information that can be gleamed from the dataset is that a total of 10 authors were responsible for 422 total commits over the period recorded. From these 10 authors, *Thomas* is the most active of the authors, responsible for 191 of the 422 commits, or just over 45% in total; *Jimmy* was the next most active with 152 commits (36%). *murari.krishnan* on the other hand was the least active with only 1 commit.

The conclusion that can be taken from this information is that the workload is very unevenly balanced across the different authors. The top two most active authors are responsible for almost 82% of total commits, four times the combined number of the other 8 authors.