PERFORM ANALYSIS ON A 5000 LINE DATASET

CA 4

MAY 6, 2017

DEREK BAKER
10353830

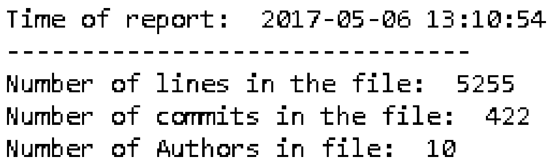
Requirements

For my assignment, I am required to look at log file and analyse it for 3 interesting statistical pieces of information. I created a report to read in the log file to scrub (clean) the data and placed it into lists, dictionaries, functions, and class and class objects .and parsed the data to pull the interesting statistics. The 3 statistics I will look at are:

- File Details – A top level look at the file.
- Authors – A look at who made changes to the log file.
- Dates – A look at the dates and the number of commits.

File Details

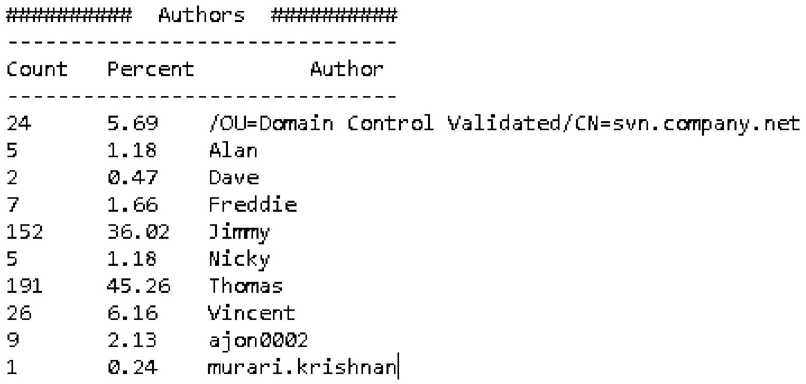
I have garnered from the changes_python.log file, at a first look, (by using simple ‘len’ code) that counts the the number of lines, number of commits and the number of authors in the file. I used code to show the time that the report is run at:



```
Time of report:  2017-05-06 13:10:54
-------------------------------
Number of lines in the file:  5255
Number of commits in the file:  422
Number of Authors in file:  10
```

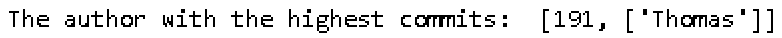
Authors

I feel that this statistical output for the authors in the log file would be the type of information sought in a report of this kind. I created a dictionary for the authors and a function to aggregate the dictionary to get qualitative and inferential information regarding the frequency of commits by the authors and returned the following output. I tried to work out how to get the results in order of count descending but was unable to:



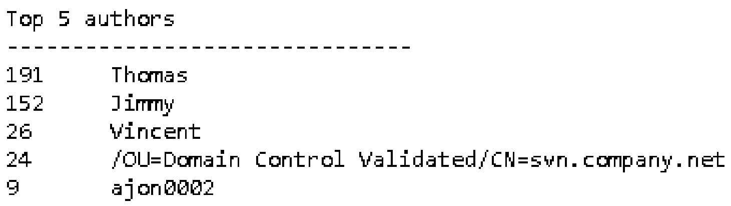
```
##########  Authors  ##########
-------------------------------
Count   Percent         Author
-------------------------------
24      5.69    /OU=Domain Control Validated/CN=svn.company.net
5       1.18    Alan
2       0.47    Dave
7       1.66    Freddie
152     36.02   Jimmy
5       1.18    Nicky
191     45.26   Thomas
26      6.16    Vincent
9       2.13    ajon0002
1       0.24    murari.krishnan
```

I then felt that the top author being highlighted is relevant and I looked for the author with the highest commits and created a function which looped through the authors dictionary looking for the author with the highest number of commits:



```
The author with the highest commits:  [191, ['Thomas']]
```

After creating that function, I felt that it would be more valuable to find qualitive information for the top 5 authors:

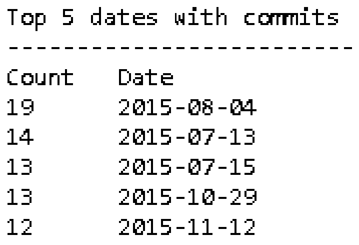


```
Top 5 authors
------------------------------
191     Thomas
152     Jimmy
26      Vincent
24      /OU=Domain Control Validated/CN=svn.company.net
9       ajon0002
```

Dates

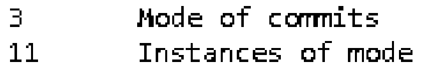
The 3rd interesting piece of statistical information regards the commits and the frequency of dates.

While it would be beneficial to show a histogram for the dates by commits I went with the simpler task of creating a function similar to the top 5 authors, showing qualitive values for the top 5 dates with most commits:



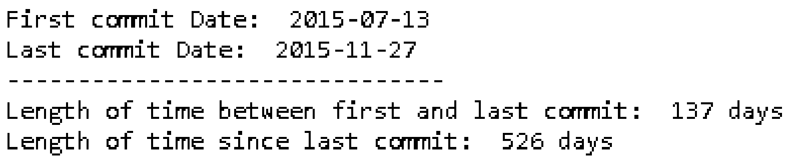
```
Top 5 dates with commits
-------------------------
Count    Date
19       2015-08-04
14       2015-07-13
13       2015-07-15
13       2015-10-29
12       2015-11-12
```

I then pulled the mode of commits and the number of instances of the mode from the same function as above: (not I am not able to run this in the command line but it works in Spider)



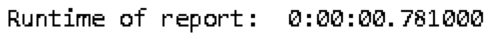
```
3        Mode of commits
11       Instances of mode
```

Then I ran code that returns time series data for the first and last commit dates which may be determined as project duration - the length of time between first and last commit and how long since last update - the length of time since last commit:



```
First commit Date:  2015-07-13
Last commit Date:  2015-11-27
------------------------------
Length of time between first and last commit:  137 days
Length of time since last commit:  526 days
```

I felt that the last commit revision number would be of use for reporting and finally, I created code to identify the runtime of the report to evaluate if the report is affected by code change or other hardware or software issues.



```
Runtime of report:  0:00:00.781000
```

Conclusion

The statistical evidence I have submitted shows that the report I have created would be beneficial to determine who had input to the advance of the project and who had the biggest impact. The timeframe of development and key dates in the project.