Student Name: Jonathan Fuller

Student Number: 10352792

Module Code: B8IT105

Module Title: B8IT105 Programming for Big Data

CA: CA4 - Perform Analysis on a 5000 line dataset

The dataset is a log of all the commits made to a particular remote repository over a period of time. Each commit log stores the following information 'revision', 'author', 'date', 'line count, ‘changes‘ and 'comment'. The 3 interesting statistical pieces of information I found with regard to dataset were the author, comments and dates.

**Author**

Many authors were working on this project, my analysis showed 10 different users committed changes to the repository. The names of authors were ‘Thomas’, 'Vincent' '/OU=Domain Control Validated/CN=svn.company.net',

'Jimmy’, 'Freddie’, 'Dave’, 'ajon0002’, 'murari. krishnan’, ‘Nicky’, and 'Alan’.

The author who made the most changes to the repository was Thomas with a total of 191 commits. The author with the least amount of changes was Murari.krishnan with only 1 commit. There is a very large difference between the amount of commits that Thomas made and Maurice. Krishnan made. As there is a total difference of 190 commits.

**Comments**

The identity of the person with the most number of Commits with most changes was user 1495579. This user had the highest number of changes per commit with a total of 297 changes made.

The average amount of changes per commit was 7.13, that was a total difference of 289.87 changes made between the maximum and the average.

**Dates**

The total number of different dates were changes were made is 76.

Most number of commits made on a single day was19 and the date that most changes were made committed was 2015, 8, 4. There seems to be no real reason for why this date was important for the authors to make changes.

**Results**

Type of changes made to repository: ['add' 'delete' 'modify']

Total number of file changes: 3009

Number of different files that were altered: 2031

Break down of change types: type

add 1056

delete 767

modify 1186

dtype: int64

Most times a file was changed 51

File most changed:/cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client-att/res/values/strings.xml

Commit with most changes: 1495579

The highest number of changes per commit: 297

Average changes per commit 7.13033175355

Number of different dates: 76

Most number of commits made on a single day: 19

Date most changes were committed: (2015, 8, 4)

Number of authors: 10

List of all authors: ['Thomas' 'Vincent' '/OU=Domain Control Validated/CN=svn.company.net'

'Jimmy' 'Freddie' 'Dave' 'ajon0002' 'murari.krishnan' 'Nicky' 'Alan']

Authors and the number of commits author

/OU=Domain Control Validated/CN=svn.company.net 24

Alan 5

Dave 2

Freddie 7

Jimmy 152

Nicky 5

Thomas 191

Vincent 26

ajon0002 9

murari. Krishnan 1

dtype: int64

Author with the least commits: murari.krishnan

Author with most commits: Thomas

Highest number of commits by author: 191