**Alexander Brown**

**Student Number: 10357791**

**B8IT105 Programming for Big Data**



**Project: CA4 - GitHub Analysis**

**November 2017**

Contents

[1. Data/Approach 3](#_Toc499459339)

[2. Cleaning the Data 3](#_Toc499459340)

[3. Test Script 3](#_Toc499459341)

[4. Python Analysis 4](#_Toc499459342)

[4.1 Commits by Author 4](#_Toc499459343)

[4.2 Comments per author and average comments per commit 5](#_Toc499459344)

[4.3 TimeOfDay 6](#_Toc499459345)

[4.4 Posting Day for Developers 7](#_Toc499459346)

[4.5 Posting Week 9](#_Toc499459347)

[4.6 Most common comment 10](#_Toc499459348)

[5. Conclusion 11](#_Toc499459349)

# Data/Approach

We have a data-file of over 5000 text lines which is made up of Github updates from various users. The high level approach to analysing this data will be as follows;

1. Use python to read in file and clean the data
2. Run a separate python test script to test the various functions I will be using on the data
3. My code will output a csv. file
4. Complete visualization of these interesting observations using a variety of tools such as Excel ToolPak and Power BI etc.

# Cleaning the Data

As per my code I cleaned the data and placed it into the following objects;

* Review
* Author
* Date
* Time
* Number
* Comment

There were 422 different sets of commit objects. An initial review of these objects shows that one of the ‘author’s in 24 records is ‘/OU=Domain Control Validated/CN=svn.company.net’. I have made an assumption that this is bad data and have changed the chars to ‘Thomas’.

# Test Script

To ensure that my extract.py code is doing its job I run a test code against it, testing the following;

* That the total number of lines in the source file is 5255
* That the total number of commit objects in the output file is 422
* That the index commits[0] of the commits returns the correct data
* That the authors in commits [0] and commits [420] are ‘Thomas’ and ‘Jimmy’ respectively.

# Python Analysis

## 4.1 Commits by Author

This chart shows the number of commits by author. We can see that Thomas has the highest number of commits followed closely by Jimmy. The other authors have a fairly low amount of commits compared to the two leaders.

This could indicate a couple of things;

* Could be because of their specific roles within a given project i.e. bug fixes, continuous updates or testing tranches etc.
* That there are serious issues with their work and there needs to be constant fixes and updates

## 4.2 Comments per author and average comments per commit

We know that Thomas generated the largest number of commits as we can see from the Power BI pie-chart below

|  |
| --- |
|  |

However, when we look at the average number of lines per author we see that Vincent has the highest number at 3.08 and Thomas has 1.20.

|  |
| --- |
|  |

## 4.3 TimeOfDay

Here, I have applied four time values to the time commits were made by authors;

* Morning1 (00.00-09.59)
* Morning2 (10.00-11.59)
* Afternoon (12.00-13.59)
* Evening1 (14.00-16.59)
* Evening2(17.00-23.59)

We can see from the visualizations below that 14.00-16.59 is the most popular time of the day to make commits with some activity in late morning and afternoon.

|  |
| --- |
|  |

This could because the bulk of work is done in the early hours of the working day and the result is saved down via commits later in the day. It could also represent a lack of productivity in the mornings.

We can see from further visualisation that only two authors generate more comments in the morning (Alan% Freddie) while the rest of the authors are more creative in the evening

|  |
| --- |
|  |
|  |
|  |

## 4.4 Posting Day for Developers

I used the ‘WEEKDAY’ function in excel to find the dates associated day and created the below chart. From this we can see that the author ‘ajon0002’ only posts on Monday, while Dave and Murari only post on Thursdays. The two highest authors post consistently across the week, starting with relatively lower volumes on Monday, Tuesday and Wednesday and then higher volumes on Thursday/Friday for Thomas. Jimmy tends to post lower on the first day and slightly higher on the second day and so on. This could demonstrate a few things;

* Thomas and Jimmy, while both posting high volumes of commits, have different working styles and this could result in one of them (Thomas) having a higher productivity
* ajon0002, Dave and Murari only post on a single day which could mean they only work this day at and also at low volumes.
* Thursday has the most commits for the week. This could indicate a problem as a development team is expected to make daily submissions in relatively equal amounts

From a Sankey chart in Power Bi we can confirm that two authors only post on Thursdays (Dave & Murari) while Ajon0002 only posts on a Monday.

|  |
| --- |
|  |

## 4.5 Posting Week

Using the WEEKNUM function in excel I produced the following chart. This shows that week 29 (mid-July) has the most commits and is nearly double the average week. It may indicate a flurry of activity before summer holidays by team members or a project delivery date.

If we look at popular comments for just week 29 we see that the comment ‘prepare for next development iteration’ occurring 3 times and ‘partial code clean-up’ occurring 6 times, indicating some flurry of activity.

## 4.6 Most common comment

I looked at the two high commit authors Thomas and Jimmy and see if they have any comments that are regularly used i.e. occur more than once. Thomas has the following comment 24 times *(“[gradle-release] prepare for next development iteration”*) and Jimmy has the following comment 10 times (“*Lint Fix*”)

I used the WordMap app on Power BI to compare the above. While not pulling back strings it does give us an insight into what exactly the authors are collaborating on – in this case it seems to be the development of a phone/software when we see high volume words such as ‘phone’, ’android’, ’screen’, ’homescreen’ etc.

|  |
| --- |
|  |

# Conclusion

From the above information and associated observations we can say the following about the data;

* Thomas is the most active author i.e. the largest number of commits objects belong to him with Jimmy coming in second. However, while Thomas has the highest number of commits he does not have the highest average of ‘comments’ per commit. This could be because of the nature of each author’s role within the team or a productivity issue.
* Authors are more active in the late afternoon/evening in terms of number of commits and comments created. Two authors only post on Thursdays and one author only posts on a Monday. We can see that the largest generation on commits occurs on a Thursday and week 29 (mid-July) was the busiest week. All of this would need to be looked at further if one wanted to analyse the team’s productivity and performance but the above conclusions would be an ideal starting point.
* The most common comments might indicate the nature of the role Thomas and Jimmy work at while the wordmap visualization indicate the nature of what project the team are working on