Programming for Big Data

CA- 4, TEXT FILE ANALYSIS – STUDENT ID – 10360474, JAY MONPARA

This is a report for a text file analysis given as a large dataset – over 5000 line of text.

The given file for analysis is as below.

Screen Shot of Text file:

```
------------------------------------------------------------------------
r1551925 | Thomas | 2015-11-27 16:57:44 +0000 (Fri, 27 Nov 2015) | 1 line
Changed paths:
   A /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client/res/drawable-xxxhdpi (from /cloud/personal/clie
   D /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client/res/drawablw-xxxhdpi
   A /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client-bt/res/drawable-xxxhdpi (from /cloud/personal/c
   D /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client-bt/res/drawablw-xxxhdpi

Renamed folder to the correct name
------------------------------------------------------------------------
r1551575 | Thomas | 2015-11-27 09:46:32 +0000 (Fri, 27 Nov 2015) | 1 line
Changed paths:
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/common/mvn-config/res/values/environment.xml
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/mct/build-config/prod.properties
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/mct/build-config/prod_lowbandwidth.properties
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/mct/build-config/qa.properties
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/mct/build-config/qa_lowbandwidth.properties
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/mct/build-config/voucher.properties
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client/build-config/dev.properties
   M /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client/build-config/prod.properties
```

Overview of the file:

By scrubbing the file in python, the following information has been obtained:

Total Number of lines in a file = 5255

Total number of Commits = 422

Each commit contains the following information.

Revision no., Author, Date and Time Stamp, Number of comments and changed path details.

There are three different type of changes are done in the paths, and they are A – Amendments, D – Deletion and M – Modification.

Data Extraction in Excel format:

A python code has been developed to scrub the text file and extract the data out in Excel format for analysis.

File in Excel format:

	A	B	C	D	E	F	G	H	I	J	K
1	Revision	Author	Date	Time	Total_M	Total_A	Total_D	No.Of Lin	Comment		
2	r1551925	Thomas	27/11/2015	16:57:44	0	2	2	1	Renamed folder to the correc		
3	r1551575	Thomas	27/11/2015	09:46:32	27	0	0	1	Removed unused webview.pl		
4	r1551569	Vincent	27/11/2015	09:38:09	1	0	0	1	enable all clients		
5	r1551558	Thomas	27/11/2015	09:13:26	1	0	0	1	Chnaged jira url to htps		
6	r1551504	/OU=Dom	27/11/2015	07:05:41	1	0	0	1	[gradle-release] prepare for n		
7	r1551486	Vincent	27/11/2015	06:10:10	1	0	0	1	SFR-108 : preparing release fo		
8	r1551485	Vincent	27/11/2015	06:06:30	1	0	0	1	SFR-108 : 1.buddy sync remov		
9	r1551375	Vincent	26/11/2015	15:01:51	2	0	0	1	SFR-108 : androidM related st		
10	r1551347	Vincent	26/11/2015	14:35:32	0	0	15	1	SFR-108 : removed unnecessa		
11	r1551334	Vincent	26/11/2015	14:20:12	3	0	0	1	SFR-108 : using WL base & snc		
12	r1551332	Vincent	26/11/2015	14:17:48	2	0	0	4	SFR-108 : Create bilingual Fren		
13	r1551313	Vincent	26/11/2015	13:58:15	1	0	0	4	SFR-108 : Create bilingual Fren		
14	r1551307	Thomas	26/11/2015	13:48:29	0	0	15	1	Removed AMX specific layout		
15	r1551294	Vincent	26/11/2015	13:30:16	3	0	0	4	SFR-108 : Help Foother can be		
16	r1551249	Vincent	26/11/2015	12:16:06	1	0	0	5	SFR-108 : Create bilingual Fren		
17	r1551248	Vincent	26/11/2015	12:15:26	1	0	0	5	SFR-108 : Create bilingual Fren		
18	r1551105	Thomas	26/11/2015	09:17:08	1	0	0	1	FTRPC-393: Frontier - Still rece		
19	r1551061	Vincent	26/11/2015	07:33:33	1	0	0	3	SFR-108 : Create bilingual Fren		
20	r1550863	Thomas	25/11/2015	16:42:06	1	0	0	1	Changed cloudsdkVersion to '		
21	r1550724	Thomas	25/11/2015	14:17:42	1	0	0	1	Changed jira url to https		
22	r1550591	Thomas	25/11/2015	11:29:15	1	0	0	3	Reverted back: FTRPC-500: Fro		
23	r1550580	Thomas	25/11/2015	11:15:27	4	2	11	1	Updated splashscreen to re us		

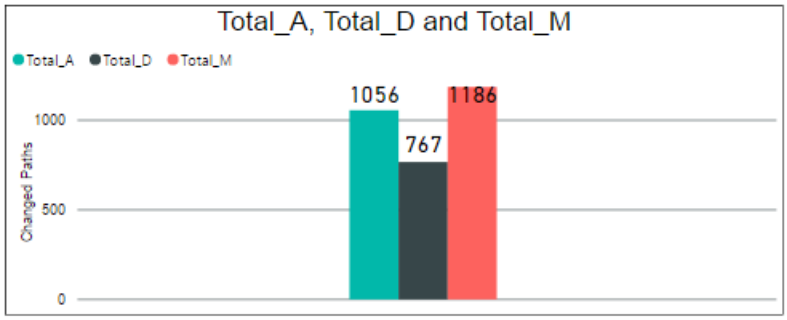
Note: Please find the attached python code for extracting the data from text file into Excel file.

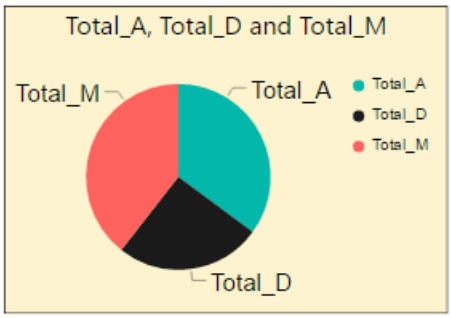
File Analysis:

Some interesting information has been obtained by analyzing the file.

Changed Paths:

The chart below shows that the total number of Amendments are **1056**, total number of Deletions are **767** and total number of Modifications done as per the data file is **1186**.

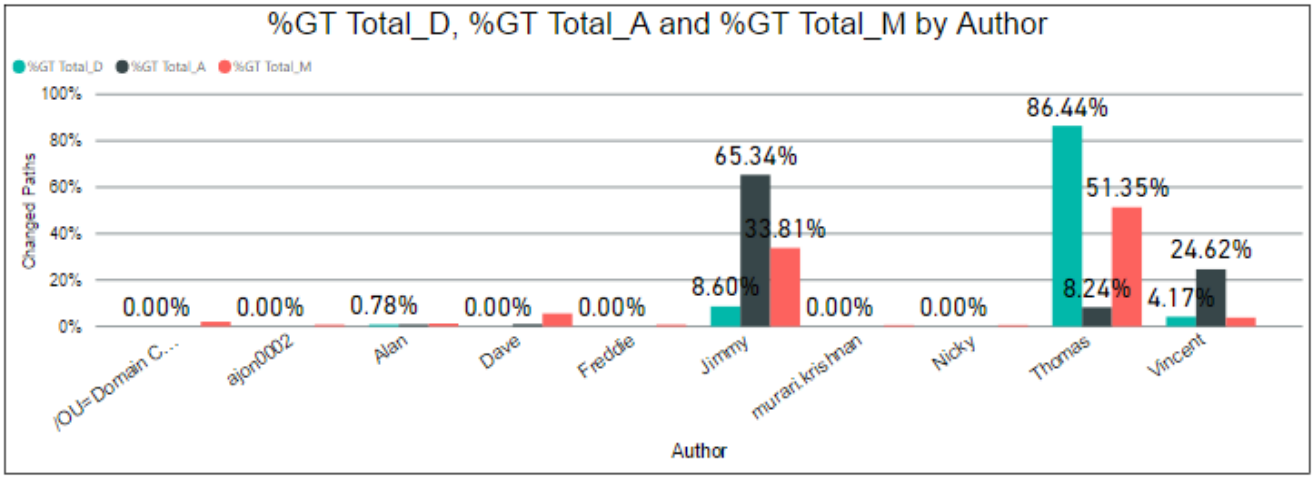




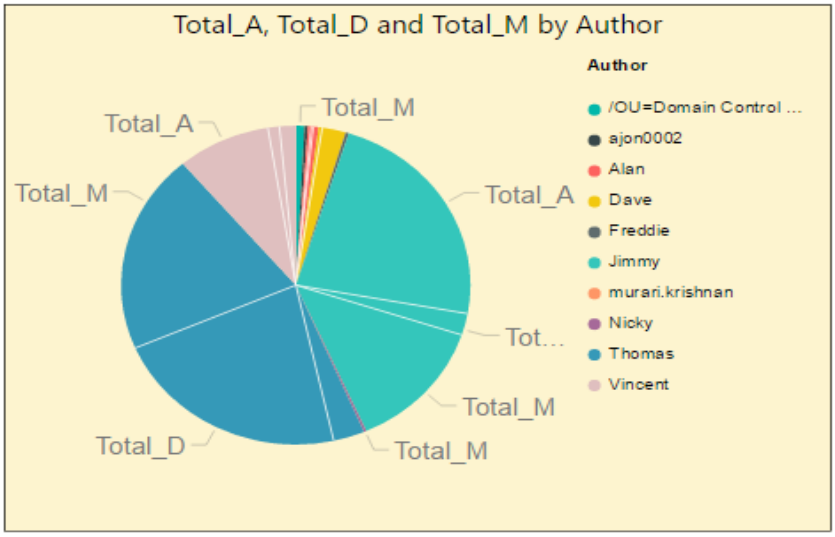
Authors & changed paths by Authors:

There are **10 authors** in total who have made any changes in the file.

The following bar chart shows the % of changes made by different authors.



Pie Chart:



As per the above chart, we can say that authors Jimmy, Thomas and Vincent are very active authors and have made most of the changes in the paths.

Amendments

The highest number of Amendments done by **Jimmy - 65.34%** followed by **Vincent - 24.64%** and **Thomas - 8.24%.** That is 98.2% of the total Amendments in the paths are done by these 3 authors.

Deletions

The highest number of deletions are done by **Thomas – 86.44%** followed by **Jimmy - 8.60%** and **Vincent – 4.17%.** Thomas has done almost 80% more deletions than the next author Jimmy. Rest of the authors total contribution in deletion is less than 1%

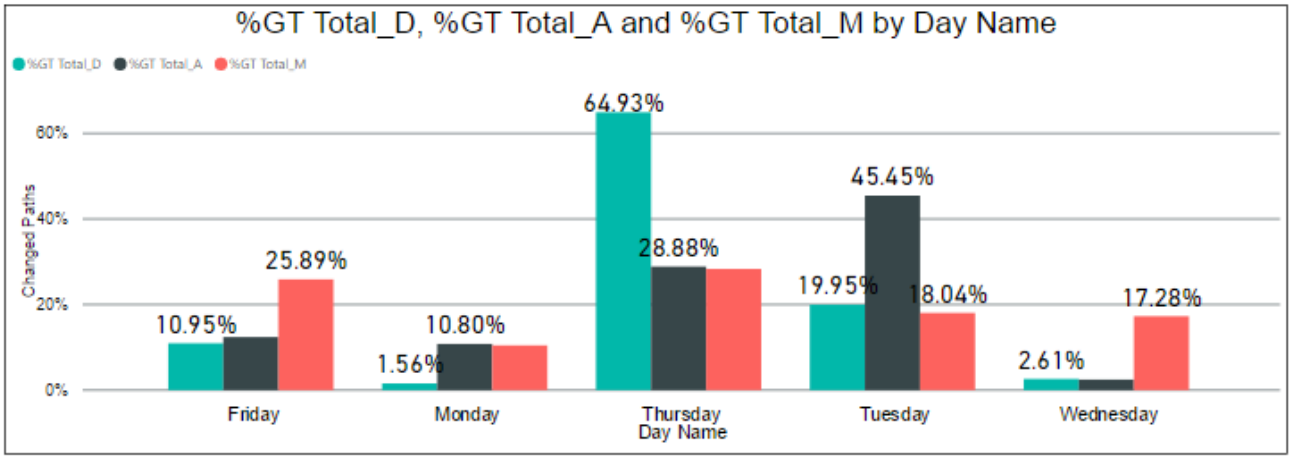
Modifications:

Most of the modifications are done **Thomas – 51.35%** (609 in total) followed by **Jimmy – 33.81%**. Remaining about 15% Modifications are done other 8 authors.

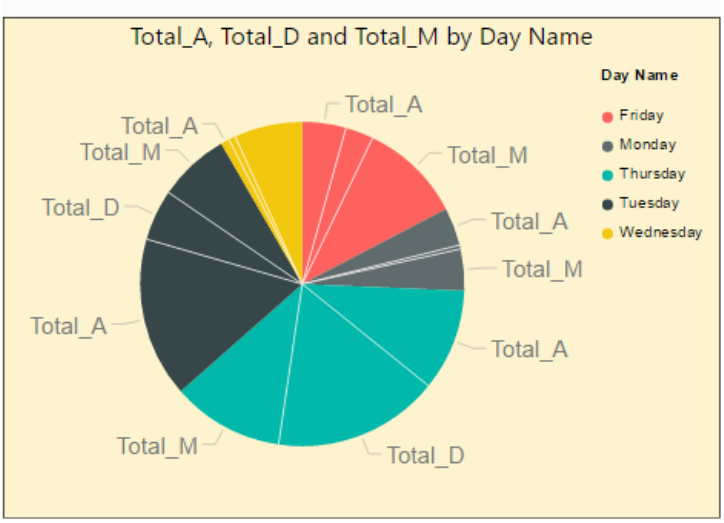
Changed Paths by Days of the week:

By drilling down the amount of path changes happened by days of week, some interesting facts are found out.

Bar chart:



Pie Chart



It is seen in the above charts that Thursday and Tuesday are busiest days of the week when significant amount of changes in paths are done. The least number of changes are done on Mondays and Wednesdays.

Deletions:

It is interesting to know that 64.93% Deletions are done on Thursdays of any week, and only 1.56% and 2.61% Deletions are on Mondays and Wednesdays.

Amendments:

The maximum number of amendments are happening on Tuesday – 45.45% followed by 28.88% on Thursdays. The minimum number of Amendments are done on Wednesdays.

Modifications:

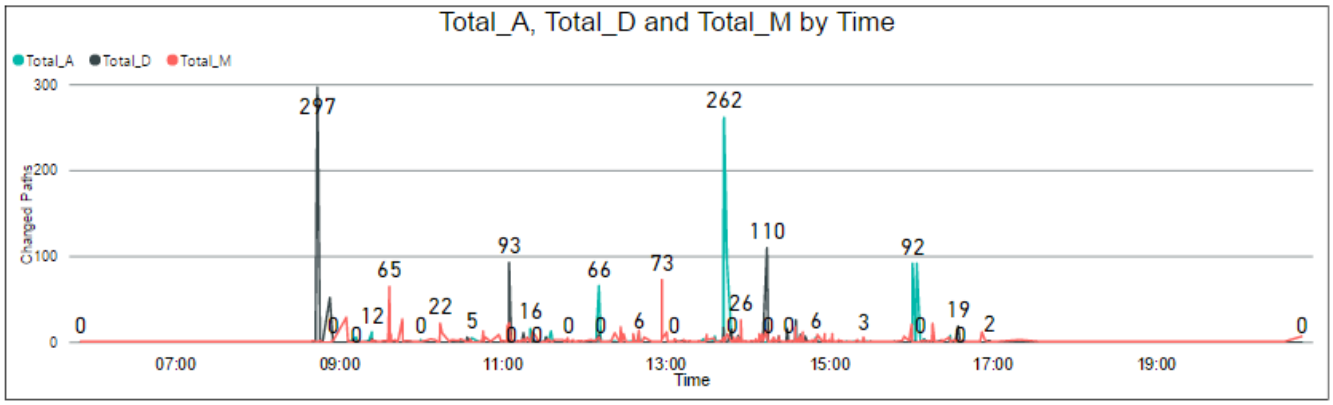
Maximum 28.88% of Modifications are done on Thursday followed by 25.89% on Friday. The lowest number of Modifications are done on Mondays.

The charts show that Modifications are happening more or less every day of the week.

<u>Path Changes by Time:</u>

Here is the line chart for times of the day vs path changes being done.

Line Chart



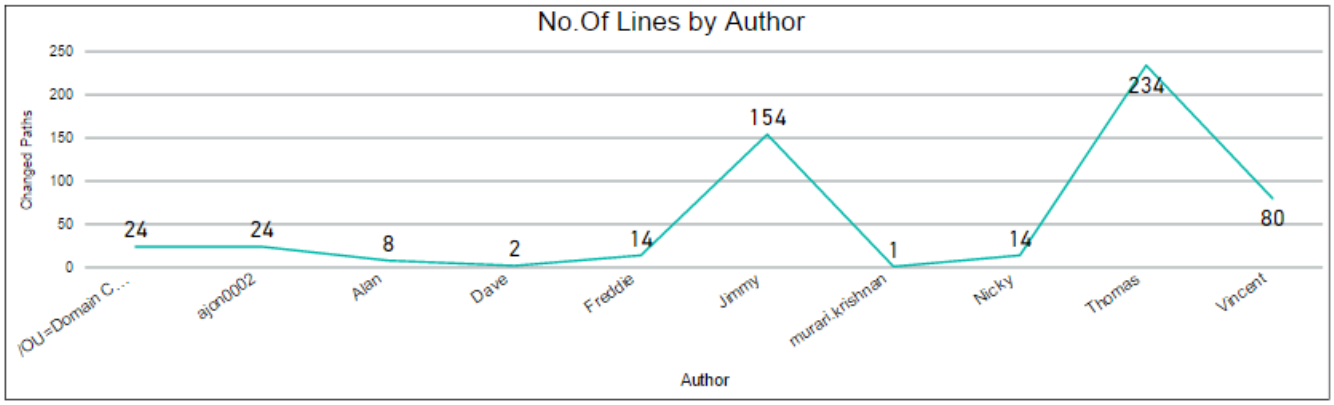
From the above chart we can see that, most of the Deletions are happening at the start of any day at around 09:00 am in the morning and most of the Amendments are done around 14:00 PM.

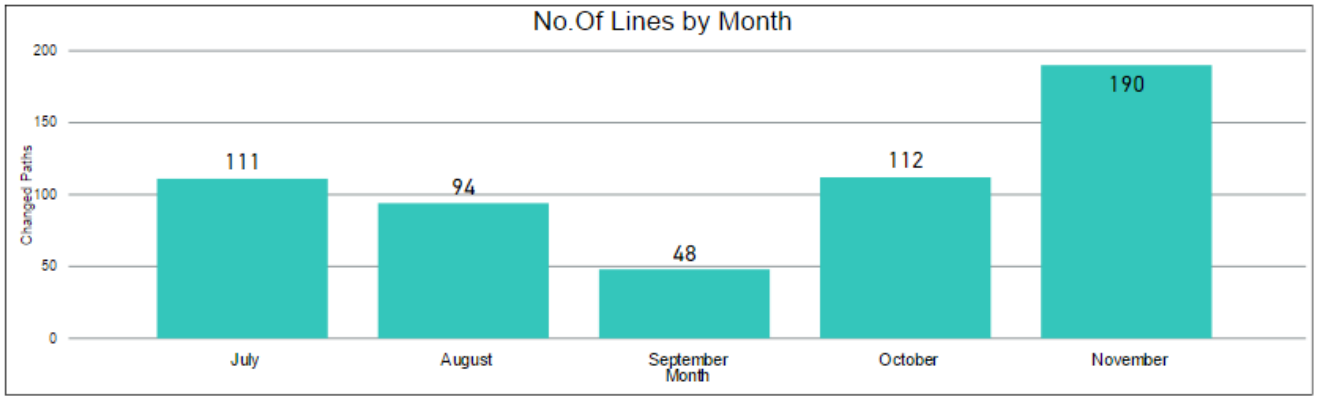
Modifications are done more or less through out the day. However, higher number of modifications are done before the lunch time on any day of the week.

Comments by Authors:

The charts below show that the maximum number of comments are made by Thomas – 234 followed by Jimmy – 154. Only 1 is made by author murari.krishnan, and only 2 comments are made by Ajan.

Line chart





If we see the chart for comments made by the month of the year, it's clear that the highest number of comments are made in November – 190 comments. And lowest number of comments are made in September – 48 comments only.

Summary:

In summary, with the available data, we can say that authors Jimmy and Thomas are most active authors in terms of changing.

Most of the path changes happened on Tuesday and Thursday and significant number of changes are made around 9:00 am and around 14:00 PM.

Maximum number of comments are made in November; however, the rest of the months also have fair amount comments registered.