Project1 dataset analysis – Magda Gruszka (ID 10353369)

Data set for analysis contained over 5000 lines of text. After cleaning the data set and saving it as CSV file the remaining 422 lines became a subject of further analysis.

I used RapidMiner to perform additional analysis, filter the attributes of the data and find few interesting statistics about iy.

The attributes of the data are as follows:

* author
* comment
* date
* number\_of\_lines
* revision
* time
* att10
* att11
* att7
* att8
* att9

The main attribute which is the label of the dataset is the author. I decided to drop ‘revision’ as this attribute contained just random numbers designated to each entry. I dropped also all ‘att’ attributes as they appeared to be the following lines of comments and 99%of them was missing. The number\_of\_lines which contained just 1 for all data lines was also dropped in that case.

After that I looked at the authors attribute. There is 10 different ones, and the quantity of entries for each of them is as follows:

A screenshot of a cell phone

Description generated with very high confidence

Author Entries/lines Fraction

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Thomas | 191 | 0.45260663507109006 |
| 2 | Jimmy | 152 | 0.36018957345971564 |
| 3 | Vincent | 26 | 0.061611374407582936 |
| 4 | /OU=Domain Control Validated/CN=svn.company.net | 24 | 0.05687203791469194 |
| 5 | ajon0002 | 9 | 0.02132701421800948 |
| 6 | Freddie | 7 | 0.016587677725118485 |
| 7 | Alan | 5 | 0.011848341232227487 |
| 8 | Nicky | 5 | 0.011848341232227487 |
| 9 | Dave | 2 | 0.004739336492890996 |
| 10 | murari.krishnan | 1 | 0.002369668246445498 |

The next step was to filter the authors and drop all with insignificant numbers of lines/entries. I did this using RapidMiner filter. After that my data set was narrowed to records for Thomas and Jimmy with no missing values in any of the remining attributes.

Jimmy contains 152 entries among 133 days, starting on 13/07/2015 at 11:03:48 with last one on 23/11/2015 at 15:02:18.

There are 38 unique entries, which means throughout all the period 28% days are with entries. 152 entries in 38 days makes 4 entries daily in average.

The time of entries falls between 8:26 and 17:59, 112 out of 152 is between 12 and 17pm with the majority of 94 in between 12 and 14pm – which is 61% of all time entries for Jimmy.

Thomas contains 191 entries among 137 days, starting on 13/07/2015 at 09:21:48 with the last one on 27/11/2015 at 16:57:44.

There are 51 unique entries, which means throughout all the period 37% days are with entries. 191 entries in 51 days gives 3.7 entry per day in average.

The time of entries falls between 8:38 and 16:57, the highest frequency is between 14 – 15pm - 39 entries, 11-12 and 16-17 are both having 29 entries.

After that I’ve copied the comments from RapidMiner results for both Thomas and Jimmy and created two separate text documents. Using and modifying python code from <https://stackoverflow.com/questions/16749635/sort-list-by-frequency-value-in-python>

On my command prompt console, I got the list of the most frequent words occurring in both.

I have copied them to excel and sorted by decrease values, as they all started from number of occurrences it was the easiest way.

The top 19 of my list after deleting all meaningless words (‘from’, ‘and’, etc.) reads as follows:

|  |  |
| --- | --- |
| Thomas word\_count | Jimmy word\_count |
| 42phone | 33phone |
| 35android | 32android |
| 28removed | 28removed |
| 19client | 17client |
| 18frontier | 15frontier |
| 16screen | 14screen |
| 15added | 14merged |
| 14merged | 14added |
| 13unused | 13unused |
| 13notification | 13handset |
| 13handset | 12branch |
| 12branch | 11add |
| 11add | 10support |
| 10support | 10m |
| 10m | 10icon |
| 10icon | 10app |
| 10att | 9resources |
| 10app | 9notification |

The most frequent words are same for both authors, even the frequency of the most used words is nearly this same.

The simplest conclusion would be that Thomas and Jimmy both worked on some similar project in similar period of 2015.