# CA4: Transforming large data set to text format/finding insights in data

Name: Alison Lawlor

Student Number: 10363795

## Parsing data and saving into a word file.

Using a python programme, I parsed a data text file of change logs, using slice notation and indexing of specific words, to write to a file for later use.

## Finding insights

I used RapidMiner to conduct text analysis to filter and select relevant, or interesting words only from the ‘update’ column. I used Tableua to create w wordcloud based on the results. I showed the day of the week in a piec hart to reveal the most productive days overall for the team. Finally, I use Tableau.

### First insight: finding most common nouns in update section

The first step was to look for recurring nouns within the update descriptions and commons (excluding common/stopwords), to gain insight into the type of work conducted, and also the common issues being dealt with by the programmers.

I conducted a text analysis of the data, using RapidMiner. As only two columns were needed, ‘changed path’ and ‘comment’ were filtered using the ‘select attributes’ operator.

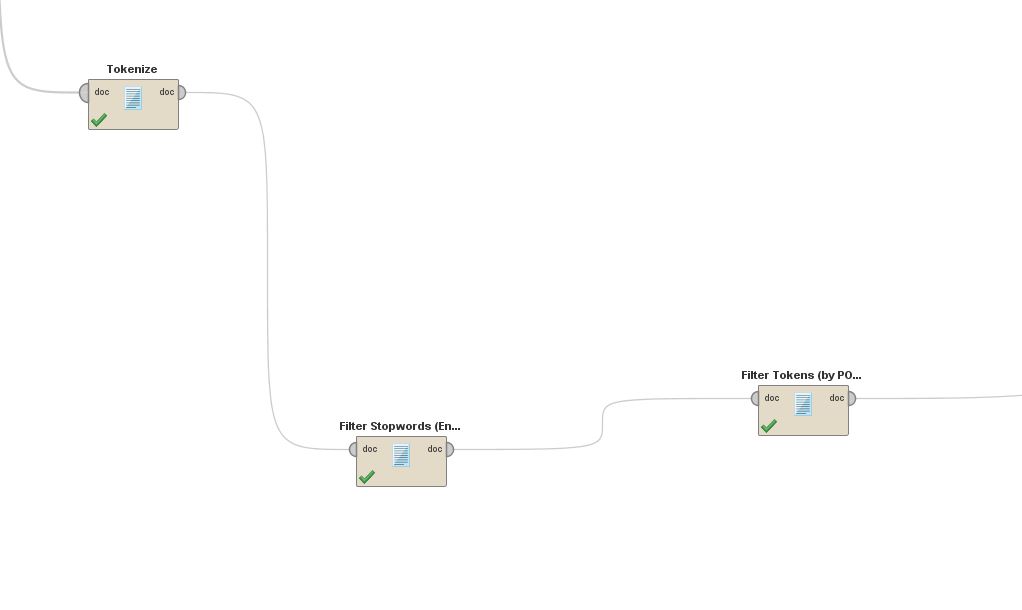
I visualised a word cloud of all nouns in the [bottom 80%](http://blog.minitab.com/blog/adventures-in-statistics-2/using-statistics-to-analyze-words-digging-deeper) of the rankings. The reason for choosing nouns over verbs, is because we are looking for insight into the subject of the data, in this case the comment or user input. Many of the verbs are action words and in this context, indicate a ‘how’ (update, fix, develop) over a ‘why’ or ‘what’ – therefore nouns, words like ‘screen’ and ‘android’ will show signs of the recurring issues. Words with a frequency of 1 or 2 do not provide insight, they rather simply clutter the view.

Words which may call for further investigation are: ATT/Screen (screenshot below and all Tableau and Rapidminer file attached).

#### Tableau wordcloud

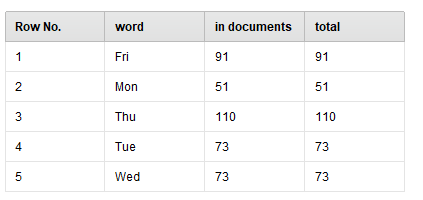


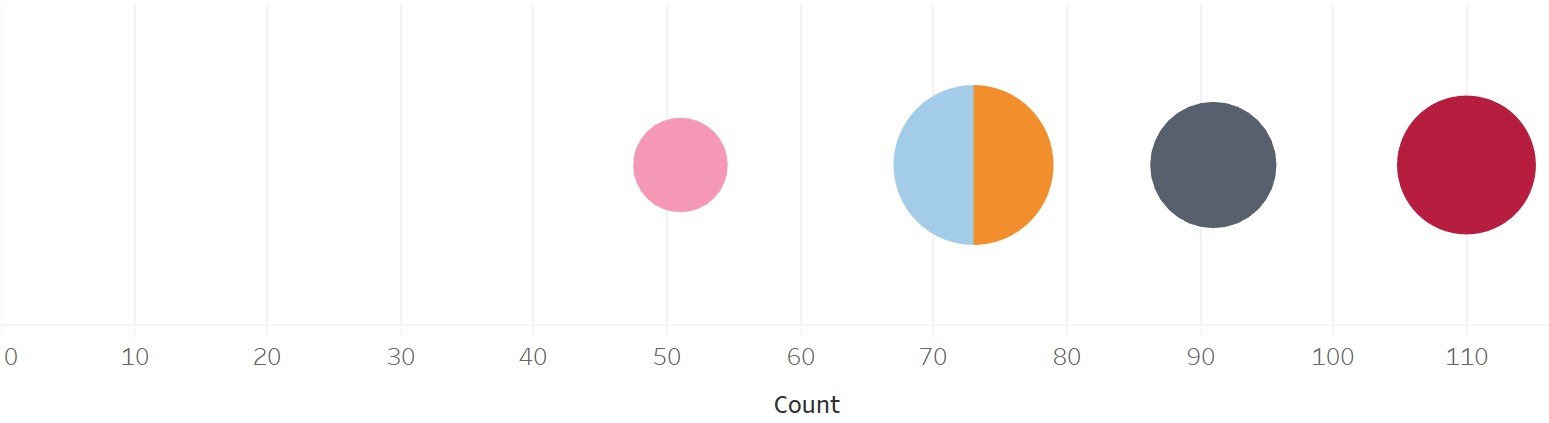
#### Using Rapidminer to tokenize, filter stopwords, and filter by ‘parts of speech’ to depict nouns only.



## Second insight

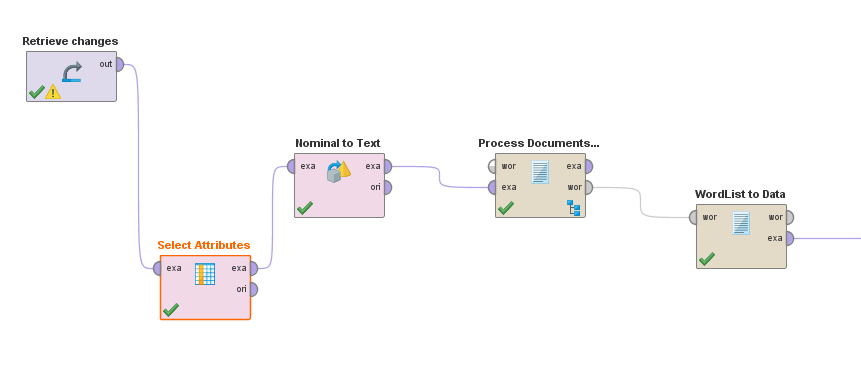
The second insight again used Rapidminer to conduct a wordcount of days of the week, to reveal the most productive days and least productive days overall. The data was visualised using a pie chart. Thursday was the most productive day, with 110 submits, and Monday (unsurprisingly) was the least productive day, with 52 submits.





### Process

#### Create text analysis workflow in Rapidminer



## Insight three

The last insight was to discover the employee submit frequency, filtered by days of the week, to determine if there were any outliers or inconsistencies for some workers.

Two employees, Jimmy and Vincent, skew the average result with a frequency far above any of the others. The max days worked is 55, on a Friday by Vincent – far above the average number of days worked, which is 6.8. Vincent deserves a very good yearly review. See graph below.

