TextAnalysis - MYP reports S1 2014-2015

FdR

3 March 2015

After downloading the .csv file from Managebac about the semester 1 report for the 2014-2015 academic year, we have copied-paste all the comments into a .txt file ready for text analysis (aka text mining).

Because text analysis is case sensitive, we put everything into lower case, we then remove all punctuation signs, we then also remove all the **stopwords**. Stopwords are all the common words like: is, he, she, I , and, etc. Finally we also remove all the extra space that are unecessary.

```
## Loading required package: NLP
```

This already strip the comments from a lots of noise.
Although reading them doesn't flow as well, the comments now look like this

```
## <<VCorpus (documents: 3, metadata (corpus/indexed): 0/0)>>
##
## [[1]]
## <<PlainTextDocument (metadata: 7)>>
## alexandra greatly benefit participating actively speaking activities give opportunity practise creat
##
## [[2]]
## <<PlainTextDocument (metadata: 7)>>
## fabio friendly member group regularly contributes class discussion participates actively group inves
##
## [[3]]
## <<PlainTextDocument (metadata: 7)>>
## zena progressing well ideas original works hard edit hand revised pieces narrative well organised eng
```

Let's see the 30 most frequently used words in this S1 reports comments. On top is the word, and the number below it is how many timed that word appeared in the comments.

```
##          class           work           well         skills           time
##            901            794            503            359            319
##           good        improve          ideas    assignments          needs
##            317            276            272            269            248
##           will          group        working    discussions  understanding
##            247            246            210            197            193
##            can           also          peers    participant       learning
##            188            187            184            181            177
##     activities            new           next           help        student
##            174            168            158            152            152
##         effort         others        writing        benefit           unit
##            149            149            149            147            147
```

A wordle would be quite nice to display in a visually more appealing way these frequent words. So let's build one for all words that appears 60 or more times on all the comments.

```
## Loading required package: RColorBrewer
```
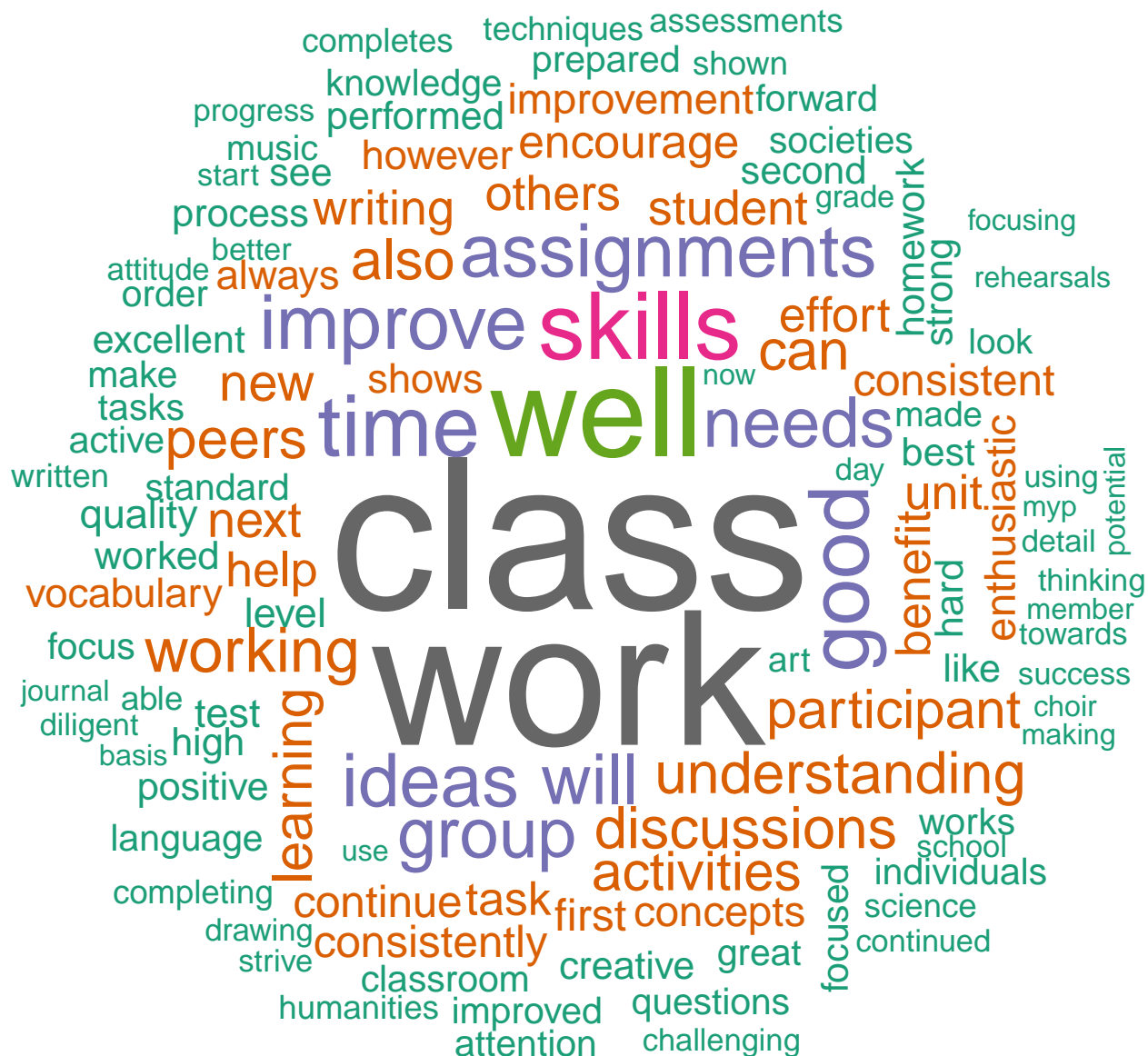
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##     speed           dist
## Min.   : 4.0   Min.   :  2.00
## 1st Qu.:12.0   1st Qu.: 26.00
## Median :15.0   Median : 36.00
## Mean   :15.4   Mean   : 42.98
## 3rd Qu.:19.0   3rd Qu.: 56.00
## Max.   :25.0   Max.   :120.00
```

You can also embed plots, for example:
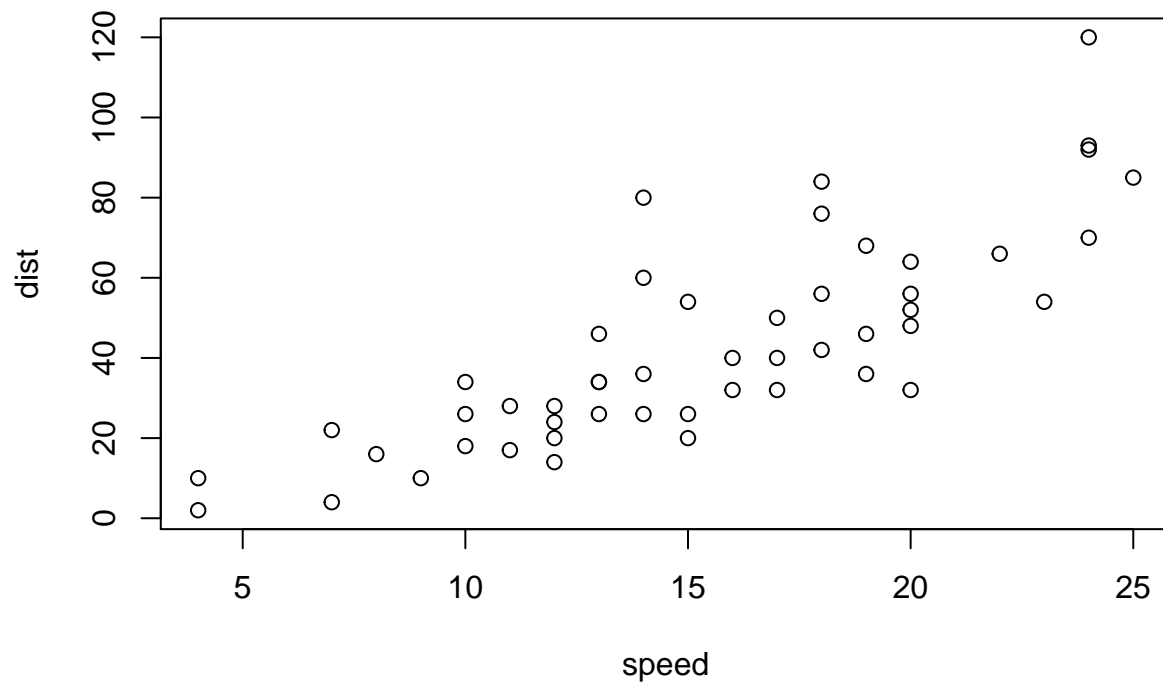
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.