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 creating language spontaneously receive constructive feedback future improvement also review complex grammar points regular basis seek help encounters difficulties  
##   
## [[2]]  
## <<PlainTextDocument (metadata: 7)>>  
## fabio friendly member group regularly contributes class discussion participates actively group investigations lines reasoning concise logical complete able select apply appropriate inquiry mathematical problem solving techniques familiar unfamiliar situations   
##   
## [[3]]  
## <<PlainTextDocument (metadata: 7)>>  
## zena progressing well ideas original works hard edit hand revised pieces narrative well organised engaging read discursive piece lacked detail facts inaccurate continues work improving analysis vocabulary writers techniques

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

Now, with the wordle (which displays over 100 words), one can see that some words that are quite similar in meaning are appearing more than once; for instance: *make*, *making* or *improves*, *improving*, or again *discuss* and *discussion*. There is a function in text mining to stem the words. By stemming our corpus, we can now redo our working our high frequency words.

## work class improv well skill time   
## 1205 907 576 503 414 376   
## particip activ need good assign idea   
## 333 329 322 317 301 286   
## will complet consist group understand continu   
## 282 268 267 261 258 257   
## focus task discuss learn perform peer   
## 249 244 227 212 204 196   
## help effort student can show also   
## 194 192 191 188 188 187

Now we can really see education related concepts on top of that list.

Another thing one can do with text mining is to check how one concept is related (associated) to other concepts. I have chosen the top 3 concepts for this:

## $work  
## allevi anxieti hard test day allot narrat finish   
## 0.25 0.25 0.23 0.21 0.20 0.19 0.19 0.18   
## group analys complet edit hand school humor manner   
## 0.18 0.17 0.17 0.17 0.17 0.17 0.16 0.16   
## principl   
## 0.16   
##   
## $class  
## particip pride discuss music contribut often creativ   
## 0.30 0.21 0.20 0.20 0.19 0.18 0.17   
## encourag french within   
## 0.16 0.16 0.16   
##   
## $improv  
## can forward look drill fit   
## 0.42 0.37 0.37 0.31 0.30   
## second calendar success overal human   
## 0.29 0.27 0.27 0.26 0.24   
## adequ deadlin forth begin continu   
## 0.23 0.23 0.23 0.22 0.21   
## first particular activitiesskil domain especi   
## 0.21 0.21 0.20 0.20 0.20   
## managebac sinc involv check cours   
## 0.20 0.20 0.19 0.18 0.18   
## manag one strongest classmat load   
## 0.18 0.18 0.18 0.17 0.17   
## managabac simpli two decis near   
## 0.17 0.17 0.17 0.16 0.16   
## sit upcom   
## 0.16 0.16

Now finding association on report comments is a really skew process due to the fact that many teacher copy paste almost the same comments for many students. An example of this is the concepts associated with the word *student*.

## student  
## humor 0.34  
## principl 0.34  
## openmind 0.31  
## demeanor 0.29  
## pleasant 0.21  
## leader 0.17  
## appreci 0.16  
## capabl 0.16  
## best 0.15  
## contributor 0.15  
## alway 0.14  
## dedic 0.14  
## exampl 0.14  
## readili 0.14  
## solver 0.14  
## fellow 0.13  
## request 0.13  
## wellmotiv 0.13  
## bit 0.12  
## cooper 0.12  
## fun 0.11  
## includ 0.11  
## much 0.11  
## obvious 0.11  
## talk 0.11

The first 3 words *humor*, *principl* and *openmind* seems to have the highest association with the concept of *student*, but at looking at the reports, it all comes from one particular teacher who copied-pasted the same phrases over and over.