Challenge 2

Problems created by Pablo Barbera, Dan Cervone « Text Analysis Module

Write commands that help you answer the following questions about the bullying dataset.

1. What are the most popular words in the dataset, other than bullying words?

```
library(tm)
```

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

```
df.tweets <- read.csv("bullying.csv", header = TRUE, stringsAsFactors = FALSE)

corpus <- VCorpus(VectorSource(df.tweets$text))
corpus <- tm_map(corpus, content_transformer(tolower))
corpus   <- tm_map(corpus, removeWords, stopwords("english"))
corpus   <- tm_map(corpus, removePunctuation)
corpus   <- tm_map(corpus, stripWhitespace)
corpus   <- tm_map(corpus, stemDocument)
dtm <- DocumentTermMatrix(corpus)
findFreqTerms(dtm, 50)
```

2. Create a wordcloud comparison between bullying tweets and non-bullying tweets.

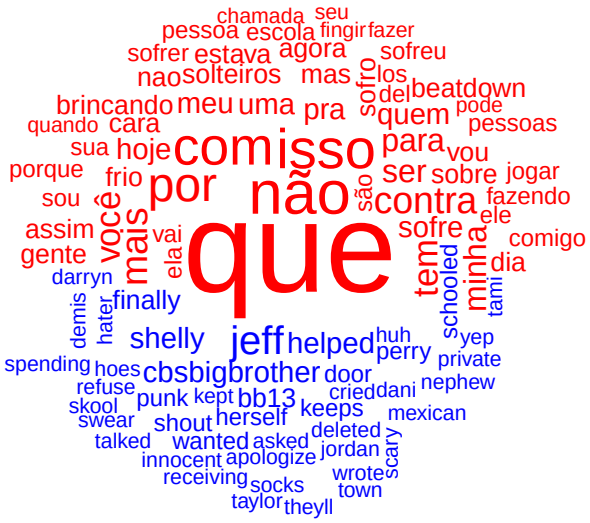
```
# Identify posts with and without bullying traces and create large documents
no_bullying <- paste(df.tweets$text[df.tweets$bullying_traces=="n"], collapse=" ")
yes_bullying <- paste(df.tweets$text[df.tweets$bullying_traces=="y"], collapse=" ")

# Create DTM and preprocess
groups <- VCorpus(VectorSource(c("No bullying" = no_bullying, "Yes bullying" = yes_bullying)))
groups <- tm_map(groups, content_transformer(tolower))
groups <- tm_map(groups, removePunctuation)
groups <- tm_map(groups, stripWhitespace)
dtm <- DocumentTermMatrix(groups)

## Label the two groups
dtm$dimnames$Docs = c("No bullying", "Yes bullying")
## Transpose matrix so that we can use it with comparison.cloud
tdm <- t(dtm)
## Compute TF-IDF transformation
tdm <- as.matrix(weightTfIdf(tdm))
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
comparison.cloud(tdm, max.words=100, colors=c("red", "blue"))
```

No bullying



Yes bullying