Assigment 2

Pavel Linder, Nikita Brancatisano

12/26/2019

0. Read input

```
train = read.table(file = 'train.tsv', sep = '\t', header = TRUE, stringsAsFactors = FALSE)
test = read.table(file = 'test.tsv', sep = '\t', header = TRUE)
length(which(!complete.cases(train)))
```

```
## [1] 0
```

```
train$text_a[1:3]
```

```
## [1] "Xanax was her death blow. \\\\xc2\\\\xa0That stuff is totally dangerous because you
## [2] "you are both morons and that is never happening"
## [3] "you are just an idiot blabbermouth that is gonna get stopped HARD one day! You W
```

1. Cleaning data

Remove punctuation and stopwords

```
train$text_a = as.character(train$text_a)
train$text_a = tm::removePunctuation(train$text_a)
train$text_a = tm::removeWords(x = train$text_a, stopwords(kind = "SMART"))
train$text_a = tm::stripWhitespace(train$text_a)
train$text_a[1:3]
```

```
## [1] "Xanax death blow xc2xa0That stuff totally dangerous build tolerance quickly stop abruptly xc2xa
## [2] " morons happening"
## [3] " idiot blabbermouth gonna stopped HARD day You WILL NOT saved"
```

Anonymize proper nouns

Remove unknown symbols (non UTF-8 characters)

```
train$text_a = str_replace_all(train$text_a, "[^[:alnum:],[:blank:]/\\-]", "")
train$text_a[1:3]
```

```
## [1] "Xanax death blow xc2xa0That stuff totally dangerous build tolerance quickly stop abruptly xc2xa
## [2] " morons happening"
## [3] " idiot blabbermouth gonna stopped HARD day You WILL NOT saved"
```

2. Exploration

Plot the frequency of words