

```
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
```

LOAD:

```
> filepath <- "retail_transaction_seurvey_data.csv"
> text<-readLines(filepath)
> docs <- Corpus(VectorSource(text))

> docs <- Corpus(VectorSource(text))
> toSpace <-content_transformer(function (x , pattern)gsub(pattern, "
", x))
> docs <- tm_map(docs, toSpace, "/")
> docs <- tm_map(docs, toSpace, "@")
> docs <- tm_map(docs, toSpace, "\\|")
```

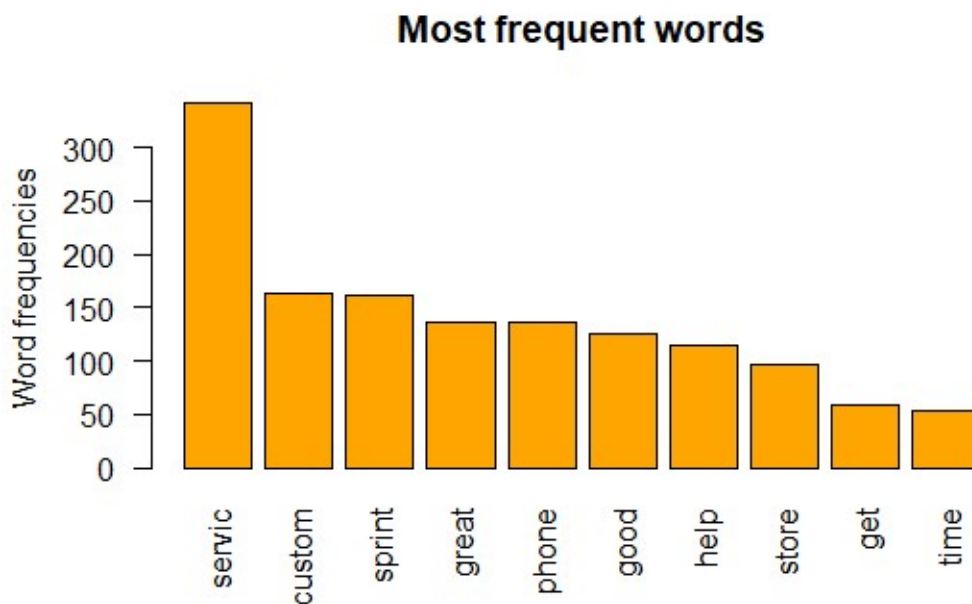
Clean text:

```
> docs <- tm_map(docs, content_transformer(tolower))
> docs <- tm_map(docs, removeNumbers)
> docs <- tm_map(docs, removewords, stopwords("english"))
> docs <- tm_map(docs, removewords, c("blabla1", "blabla2"))
> docs <- tm_map(docs, stripWhitespace)
> docs <- tm_map(docs, stemDocument)

> set.seed(1234)
> wordcloud(words = d$word, freq = d$freq, c(8,.3), 2,, min.freq = 1,
max.words = 200, random.order = FALSE, rot.per = 0.35,
colors=brewer.pal(8, "Dark2"))
```



```
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,
+       col = "orange", main = "Most frequent words",
+       ylab = "word frequencies")
```



```
> findFreqTerms(dtm, lowfreq=100)
[1] "servic" "help" "good" "phone" "sprint" "custom" "great"
```

only 7 words above occur atleast 100 times.

```
> dtm <- TermDocumentMatrix(docs)
> m <- as.matrix(dtm)
> v <- sort(rowSums(m),decreasing = TRUE)
> d <- data.frame(word = names(v),freq=v)
> head(d, 10)
```

	word	freq
servic	servic	341
custom	custom	164
sprint	sprint	161
great	great	137
phone	phone	136
good	good	125
help	help	115
store	store	97
get	get	58
time	time	54

Based on the most frequent occurring words, it looks like this text document is a customer service (word service occurring 341 times and word custom probably stemmed from the word customer occurring 164 times) **review** (words great, good and help clearly indicate a kind of review) **for sprint**(which occurs 161 times). **It is also clear from the word “phone” and “store” that it is store review of the phone service.** **And the words “great”, “help” and “good” indicate that customer mostly have positive review on the service provided.**