R Notebook

```
library("tm") # text mining
```

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

```
library(stringr)
library(wordcloud)
```

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

```
options(warn=-1)
```

```
o.jd <- readLines("jds.txt") # each line is a vector
str(o.jd)
```

```
##  chr [1:199] "Data Scientist (ML) skills:" "" ...
```

```
t.jd <- paste(o.jd, collapse = " ") # convert to a single vector
t.jd <- gsub(pattern = "\\W", replace = " ", t.jd) # replace punctuation with spaces
t.jd <- tolower(t.jd) # to lower case
t.jd <- removeWords(t.jd, stopwords("english")) # remove stopwords
# t.jd <- gsub(pattern = "\\b[A-z]\\b{1}", replace = " ") # replace words of length 1
t.jd <- tm::stripWhitespace(t.jd) # cleanup white spaces from previous transformations

# t.jd
```

```
s.jd <- str_split(t.jd, pattern = "\\s+") # split string
s.jd <- unlist(s.jd) # list to a character vector\

pal2 <- brewer.pal(5,"Dark2")
wordcloud(s.jd, min.freq = 3, random.order = FALSE, scale = c(4,0.7), color = pal2, rot.per=.1)
```


```
ta.jd <- sort(table(s.jd), decreasing = TRUE)
ta.jd <- as.data.frame(ta.jd)
names(ta.jd)[1] <- "Word"
ta.jd[ta.jd$Freq > 10, ] # top words
```

```
##           Word Freq
## 1   experience   43
## 2         data   42
## 3     learning   34
## 4      machine   25
## 5      science   20
## 6       skills   17
## 7       models   15
## 8     computer   14
## 9    technical   14
## 10 engineering   13
## 11        deep   11
## 12     product   11
## 13      python   11
## 14     related   11
```