Data Visualization and Technique

Saurabh Yelne and Joseph Vele

Nov 8, 2017

```
ama_con <- read_excel("~/Amazon.xlsx")
str(ama_con)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    400 obs. of  2 variables:
##  $ Sr.no          : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Employee Review: chr  "Constantly on your feet" "Lack of structure (workwise)" "Can be a lot of w
```

Isolating Cons reviews'

```
ama_con<-ama_con$`Employee Review`
```

Making a vector source

```
ama_con_vec<-VectorSource(ama_con)
```

Making a VCorpus

```
ama_con_corpus<-VCorpus(ama_con_vec)
```

Creating Clean_Corpus Function using tm package functions

```
clean_corpus <- function(corpus){
  corpus <- tm_map(corpus, stripWhitespace)
  corpus <- tm_map(corpus, removePunctuation)
  corpus <- tm_map(corpus, removeNumbers)
  corpus <- tm_map(corpus, tolower)
  corpus <- tm_map(corpus, PlainTextDocument)
  corpus <- tm_map(corpus, removeWords, c('amazon','company',stopwords("en")))
            tm_map(corpus, stemDocument, language = "english")
  return(corpus)
}
```

Applying clean_corpus to Amazon Cons reviews corpus

```
clean_ama_con_corpus<-clean_corpus(ama_con_corpus)
```

Note the difference between orinal text and the cleaned

```
clean_ama_con_corpus[[7]][1]
```

```
## $content
## [1] "huge   dont always feel   individual work   steering  ship"
```

```
ama_con_vec$content[7]
```

```
## [1] "HUGE company so don't always feel that my individual work is ' steering the ship'."
```

Creating a bigram tokenizer function

```
BigramTokenizer <-
  function(x)
    unlist(lapply(ngrams(words(x), 2), paste, collapse = " "), use.names = FALSE)
```

Create term-document matrix (TDM) from our amazon cons clean corpus

```
amazon_con_tdm <-TermDocumentMatrix(clean_ama_con_corpus, control = list(tokenize=BigramTokenizer))
dim(amazon_con_tdm)
```

```
## [1] 4113  400
```

Convert the amazon_con_tdm to matrix

```
amazon_con_m<- as.matrix(amazon_con_tdm)
dim(amazon_con_m)
```

```
## [1] 4113  400
```

```
amazon_con_m[1000:1005,6:10]
```

```
##                          Docs
## Terms                     character(0) character(0) character(0)
##   environment ability                0            0            0
##   environment aggressive             0            0            0
##   environment aspects                0            0            0
##   environment can                    0            0            0
##   environment competitive            0            0            0
##   environment creating               0            0            0
##                          Docs
## Terms                     character(0) character(0)
##   environment ability                0            0
##   environment aggressive             0            0
##   environment aspects                0            0
##   environment can                    0            0
##   environment competitive            0            0
```

```
##   environment creating               0            0
```

```
ama_c_freq <- sort(rowSums(amazon_con_m),decreasing=T)
```

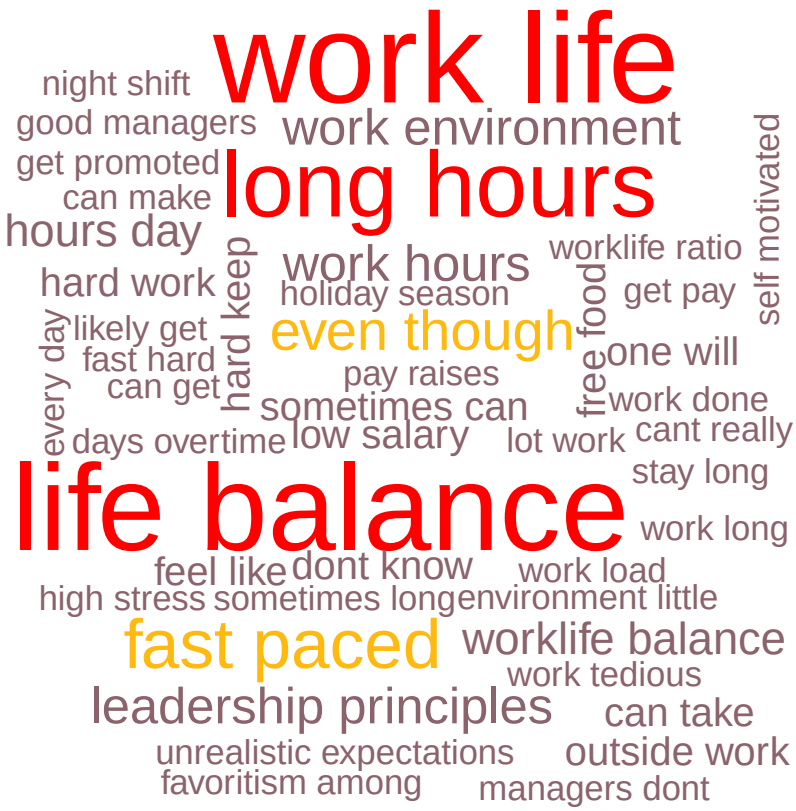
Creating a word cloud of 25 negative words in employee reviews

```
wordcloud(names(ama_c_freq),ama_c_freq,max.words=50,colors=c('pink4','darkgoldenrod1',"red"))
```

```
## Warning in wordcloud(names(ama_c_freq), ama_c_freq, max.words = 50, colors
## = c("pink4", : can stressful could not be fit on page. It will not be
## plotted.
```

```
## Warning in wordcloud(names(ama_c_freq), ama_c_freq, max.words = 50, colors
## = c("pink4", : short breaks could not be fit on page. It will not be
## plotted.
```

```
## Warning in wordcloud(names(ama_c_freq), ama_c_freq, max.words = 50, colors
## = c("pink4", : among management could not be fit on page. It will not be
## plotted.
```



Converting Amazon Frequency to a data frame

```
ama_c_freqdf<-as.data.frame(ama_c_freq)
ama_c_freqdf<-setDT(ama_c_freqdf,keep.rownames = T)[ ]
colnames(ama_c_freqdf)<-c('Bigram','Freq')
ama_c_freqdfbar<-ama_c_freqdf[1:10,]
```

Bar Graph of top 10 bigrams

```
pl<-ggplot(ama_c_freqdfbar,aes(x=reorder(Bigram,-Freq),y=Freq))
pl2<-pl+geom_bar(stat = 'identity',fill='tan2')+xlab('Top 10 Bigrams')+ggtitle('Frequency of top 10 bigr
print(pl2)
```