wordCloud

Monday, May 11, 2015

Word Cloud and Semantic Analysis.

Four Human Resources Management texts are combined together for this analysis on both individual and co-word frequency. The combination of documents begins with 15,270 total words and through initial processing (described below) reduces to a working version of 8,548 words and bigrams.

```
library(tm)                     ## text management package
library(wordcloud)              ## word cloud addendum
library(plyr)                   ## data manipulation
library(ggplot2)                ## plotting
library(RColorBrewer)           ## plot colors used in wordcloud
library(SnowballC)              ## needed for wordcloud
library(syuzhet)                ## new package for sentiment analysis
library(reshape2)               ## matrix management
```

Section 1. Load data into lists for processing.

We work from the list of files in the directory provided. Each of the files is opened, scanned, parsed by word, and then appended to the end of the existing list of words.

```
setwd("C:/wordanalysis")
filelist <- list.files("C:/wordanalysis")
f = length(filelist)
words <- NULL
bigrams <- NULL

for (i in 1:f) {
      x <- scan(filelist[i],character(0),quote=NULL)
      words <- c(words,x)
}
```

Section 2. Initial document processing.

Most text processing intends to correct unnecessary loss of identical words ("Weather"" vs. "weather", "lesser" vs. "lesser!") as well as removal of non-value words (especially high-volume ones, such as "the","and","if", others).

We process the document as follows:

Remove punctuation, convert all words to lower case, remove low-value words ("stopwords"), removed numbers (low value - this is an HR document).

```
words <- removePunctuation(words)
words <- tolower(words)
words <- removeWords(words,stopwords("english"))
words <- removeNumbers(words)
words <- words[!words==""]
```

Section 3. Detailed processing: frequency analysis of both words and co-words.

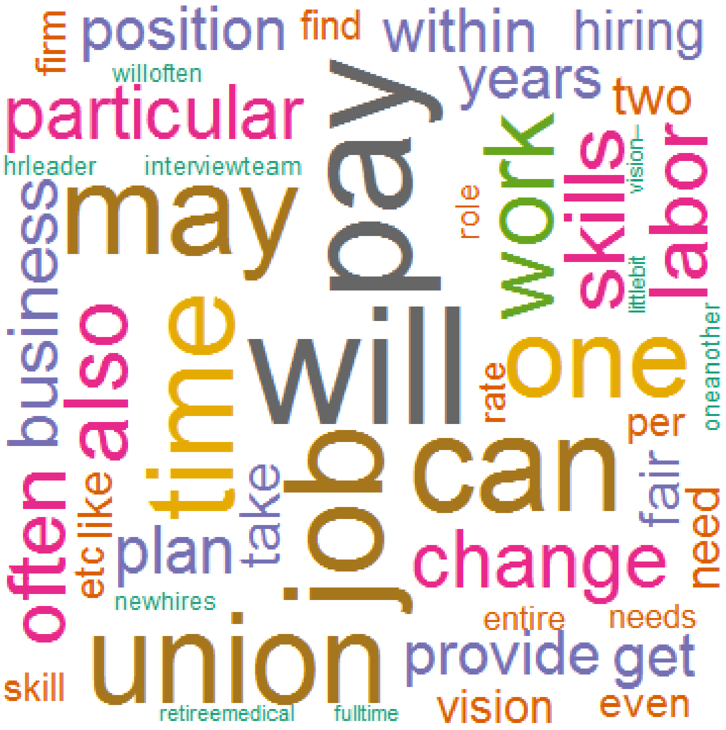
We summarize words by frequency and identify co-words by frequency as well. Higher frequency words and higher frequency co-words are merged into a common list for word cloud and semantic analysis.

```
z <- length(words)
for (b in 1:z) {
      bigrams <- c(bigrams,paste(words[b],words[b+1], sep="-"))
}

dualfreq = count(bigrams)
dualfreq = dualfreq[with(dualfreq,order(-freq)),]

freq = count(words)
freq <- freq[with(freq,order(-freq)),]

wordsForCloud <- freq[freq$freq>=10,]
bigramsForCloud <- dualfreq[dualfreq$freq>=3,]

finalwords <- words[words %in% wordsForCloud$x]
finalwords <- c(finalwords, bigrams[bigrams %in% bigramsForCloud$x])
```

Section 4: Output diagrams.

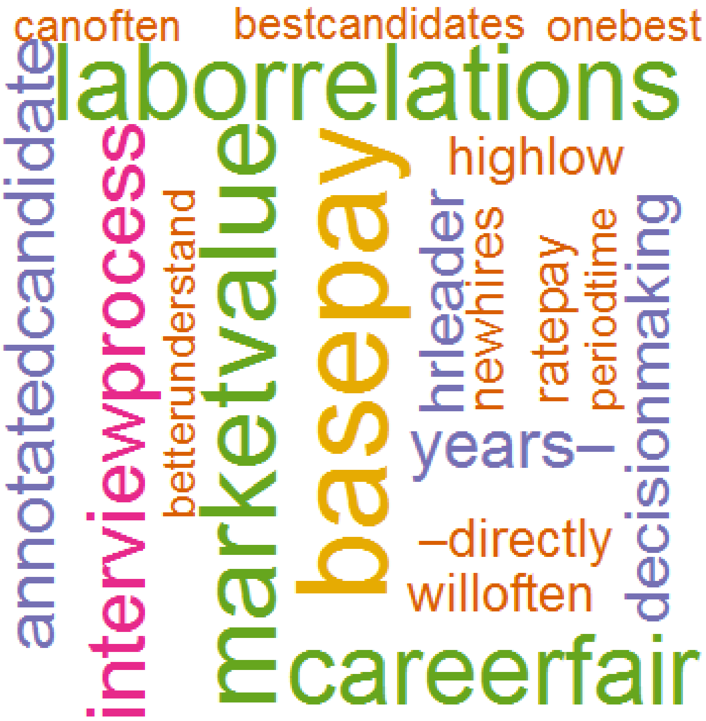
Three primary outputs: Word cloud of the aggregated list of words and co-words, a word cloud of co-words only (since co-word frequency is significantly lower than individual words), and the semantic analysis of the aggregated list.

```
wordcloud(finalwords,
          scale=c(5,0.5),
          max.words=250,
          random.order=FALSE,
          rot.per=0.35,
          use.r.layout=FALSE,
          colors=brewer.pal(8,"Dark2"))
```



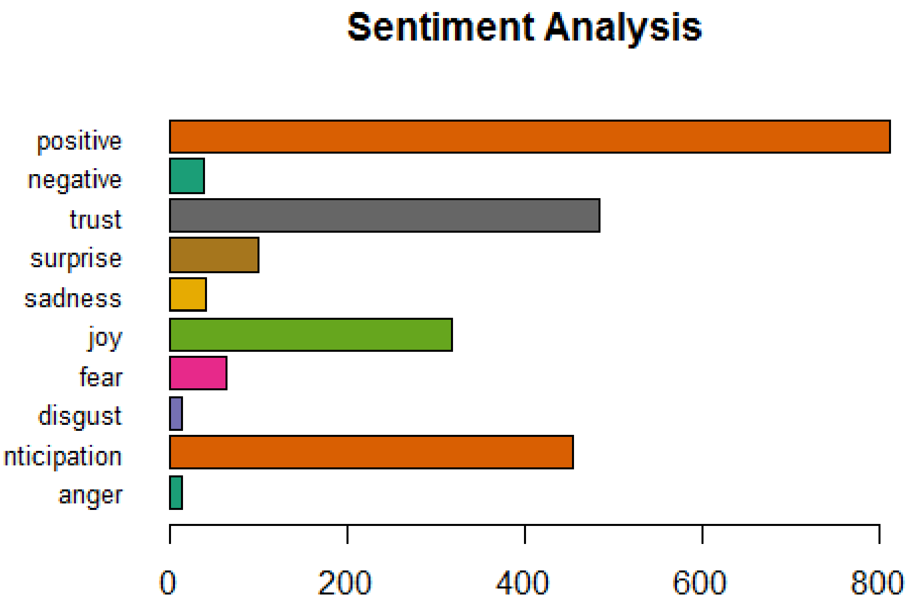
Word cloud of combined individual and co-words

```
wordcloud(bigrams,
          scale=c(5,0.5),
          max.words = 100,
          random.order = FALSE,
          rot.per = 0.35,
          use.r.layout = FALSE,
          colors=brewer.pal(8,"Dark2"))
```



Word cloud, co-words only

```
sentmt <- get_nrc_sentiment(finalwords)
n <- colSums(sentmt)


barplot(n,
        col=brewer.pal(8,"Dark2"),
        cex.names = 0.80,
        horiz = TRUE,
        main = "Sentiment Analysis",
        mar = c(6,4,4,2) + 0.1,
        las = 1)
```



Barplot of semantic analysis, combined words and co-words