INSIGHTS FROM R/SKINCAREADDICTION

Fetch reddit comments using package Reddit ExtractoR

Pre-process Text

```
#TEXT MINING
#using the tm pakcage
library(tm)

#cleaning text as a part of preprocess, taking comments and putting into
#separate dataframe
commsposts <- sa_comms$comment

#dumping comments into corpus for preprocressing (using tm package)
sk_corpus <- Corpus(VectorSource(commsposts))

#change to lowercase
sk_corpus  <- tm_map(sk_corpus, content_transformer(tolower))
#remove numbers
sk_corpus  <- tm_map(sk_corpus, removeNumbers)
#remove punctuation
sk_corpus  <- tm_map(sk_corpus , removePunctuation)
#remove stop words and "the" and "and"
sk_corpus  <- tm_map(sk_corpus, removeWords, c("the", "and", stopwords("english")))
#stemming words
sk_corpus  <- tm_map(sk_corpus, stemDocument, language = "english")
#removing white space
sk_corpus  <- tm_map(sk_corpus, stripWhitespace)
```

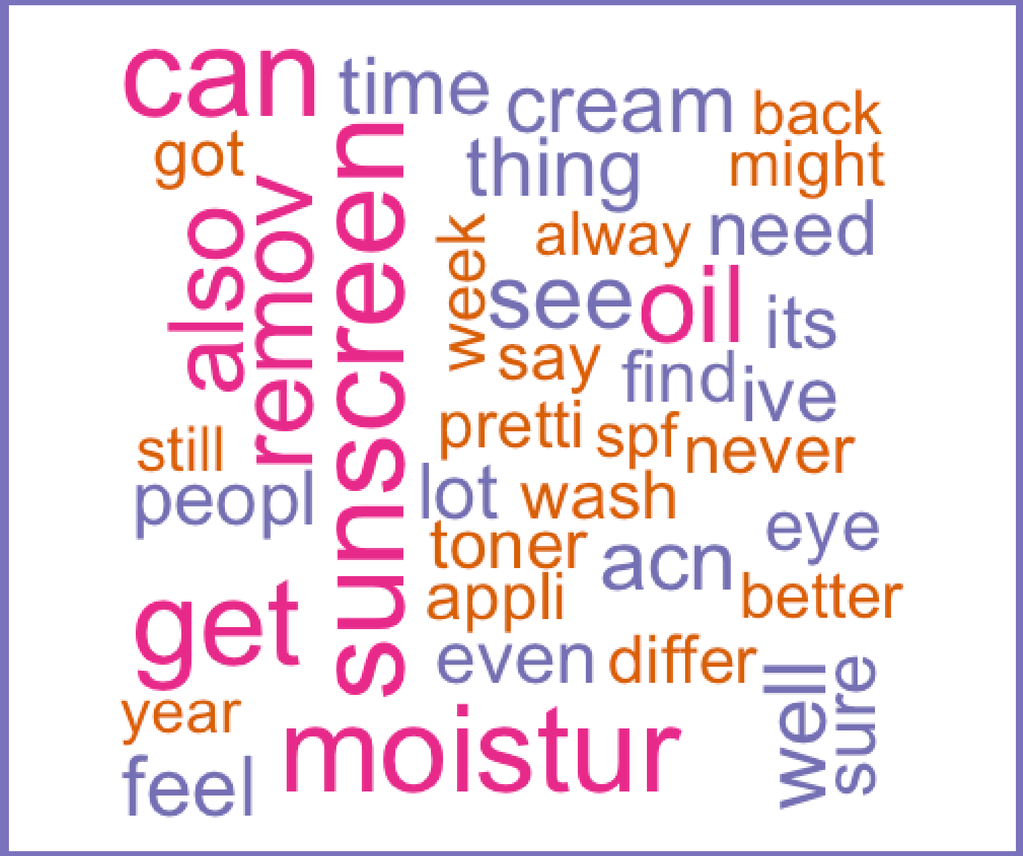
TF-IDF =

TF(t) = (Number of times term appears in a comment) / (Total number of terms in the comment)

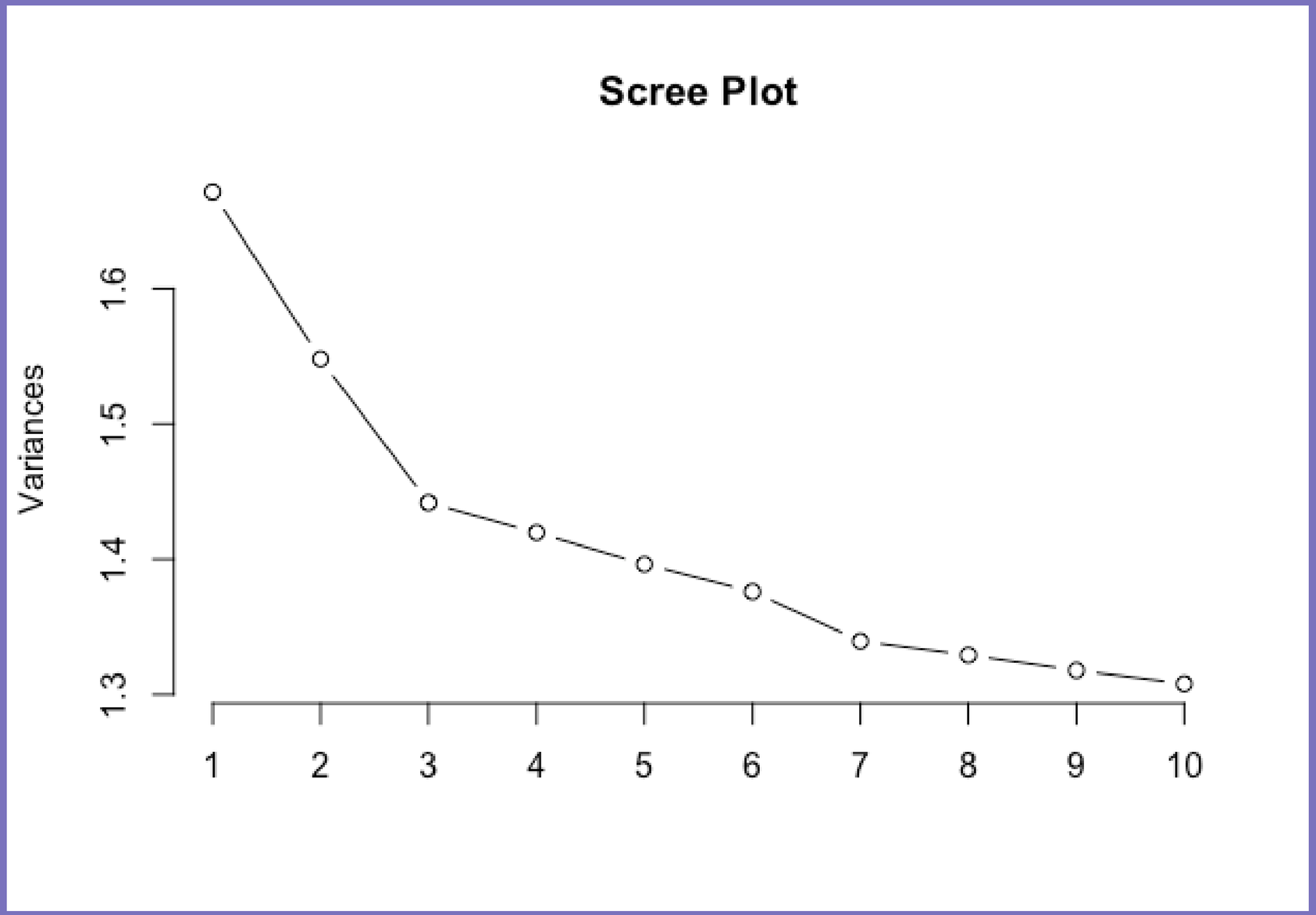
*

IDF(t) = log(Total number of comments / Number of comments with term t in it)

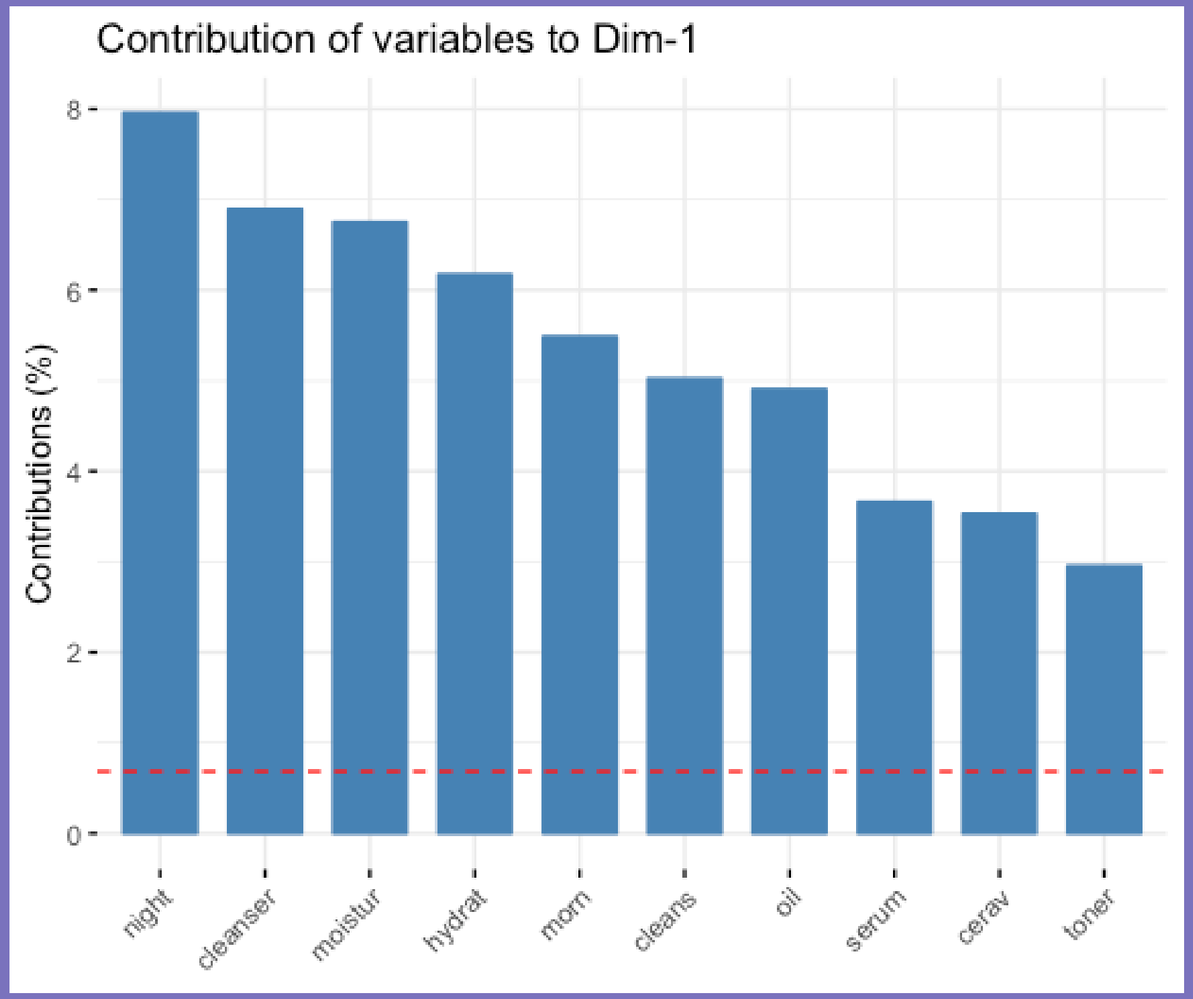
```r
#ranking of terms
termsdec <- data.frame(sort(colSums(as.matrix(tfidfdf_skdf)), decreasing=TRUE))
termsdec #show decreasing list of tfidf words


library(wordcloud)
#now making wordcloud
wordcloud(rownames(termsdec), termsdec[,1], max.words=60, colors=brewer.pal(4, "Dark2"))
#word cloud shows that sunscreen, moisture, oil are high tf-idf terms
```

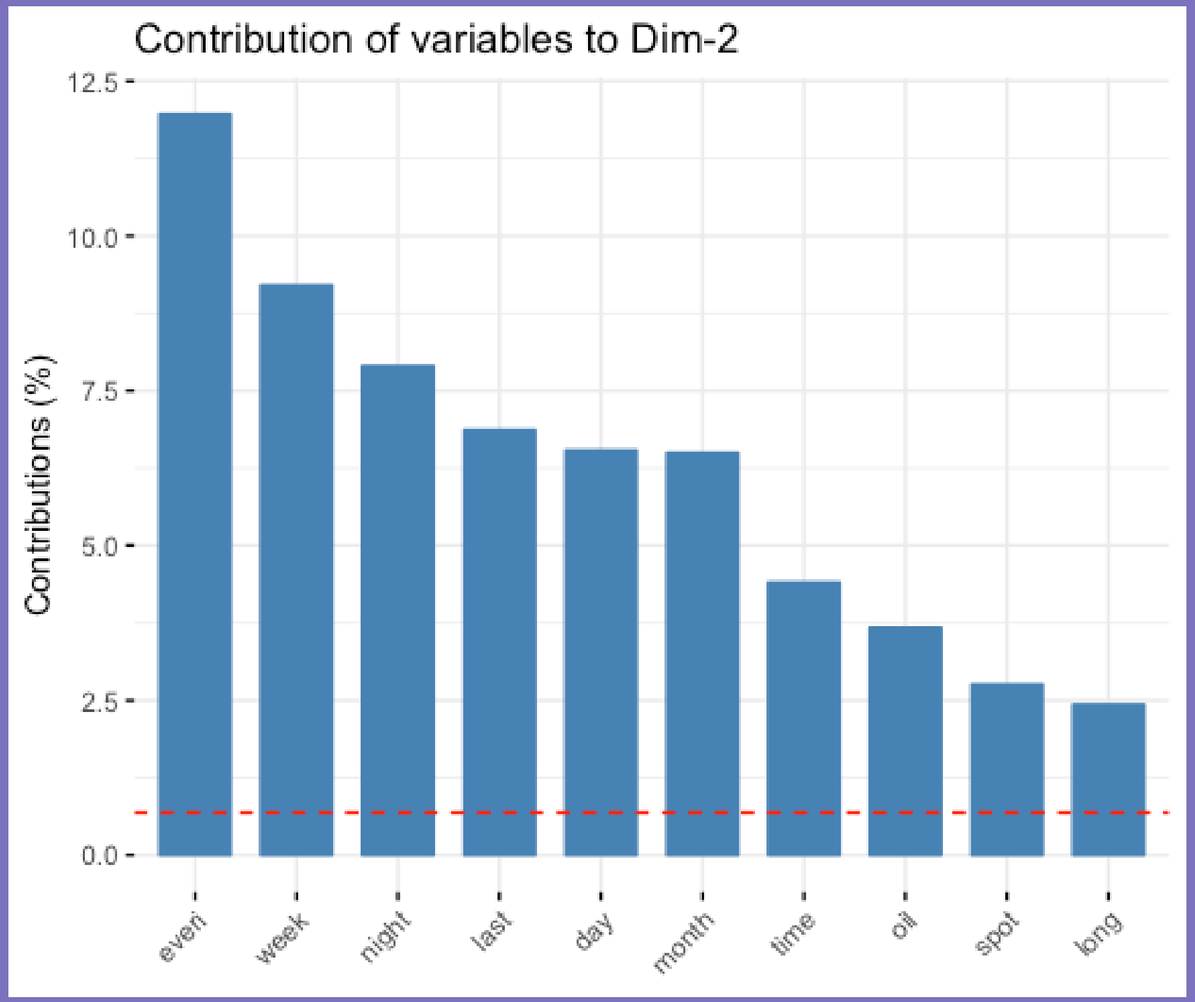


PCA



PCA


Contribution of variables to Dim-1
Contributions (%)
8
6
4
2
0
night
cleanser
moistur
hydrat
morn
cleans
oil
serum
cerav
toner


Contribution of variables to Dim-2
Contributions (%)
12.5
10.0
7.5
5.0
2.5
0.0
even
week
night
last
day
month
time
oil
spot
long