Wordcloud

Sébastien Guyader

2018 November 28

```
library(knitr)
opts_chunk$set(dev.args=list(pointsize=10))
```

Load the required librairies, and load extra fonts.

```
library(wordcloud)
library(tm)
library(tidyverse)
library(cowplot)
library(extrafont)

#font_import() # run this command the first time only to import the
              # True Type fonts from the operating system
loadfonts(device="pdf", quiet=F) # device="win" or "pdf" possible
```

Load text files to be analysed, into a corpus.

```
#set the path to the folder containing the documents
source <- DirSource("./wordcloud_text")

#load in txt documents
docs <- VCorpus(source, readerControl=list(reader=readPlain,
                                           language="en-EN"))
docs
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
```

```
summary(docs)
```

```
##          Length Class             Mode
## text.txt 2      PlainTextDocument list
```

```
inspect(docs[1])
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
##
## [[1]]
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 13942
```

Filter out undesirable words.

```
skipwords = c(stopwords("SMART"), "fig", "allowed", "showed", "lwd", "wpa")
kb.tf <- list(weighting = weightTf, stopwords  = skipwords,
              removePunctuation = TRUE,
              tolower = TRUE,
              minWordLength = 3,
              removeNumbers = TRUE, stripWhitespace = TRUE,
              stemDocument= TRUE)
```

Generate the term-document matrix and convert it to a matrix.

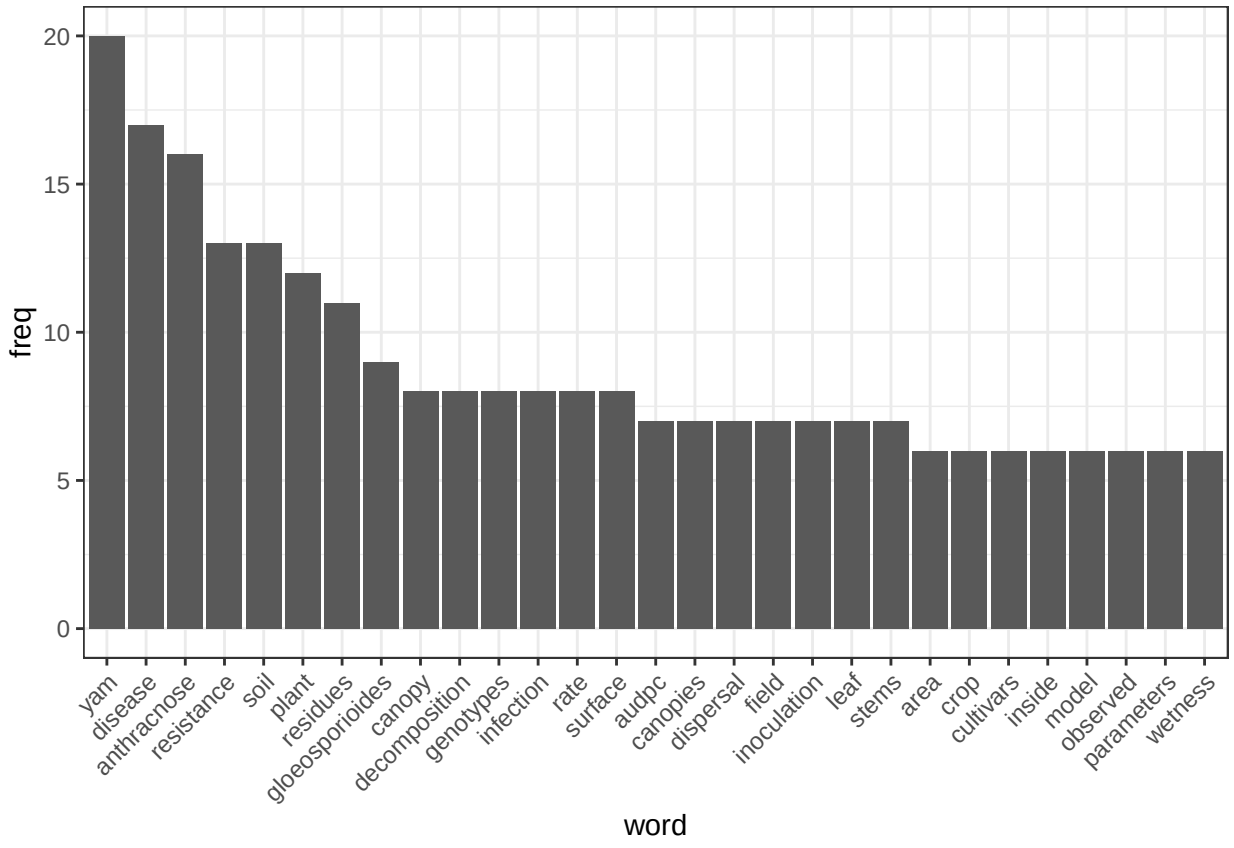
```
tdm <- TermDocumentMatrix(docs, control = kb.tf)
tdm <- as.matrix(tdm)
```

Get word counts in decreasing order and create a data frame with words and their frequencies. It is important to set the words variable as ordered factor, so that ggplot2 will not sort them alphabetically.

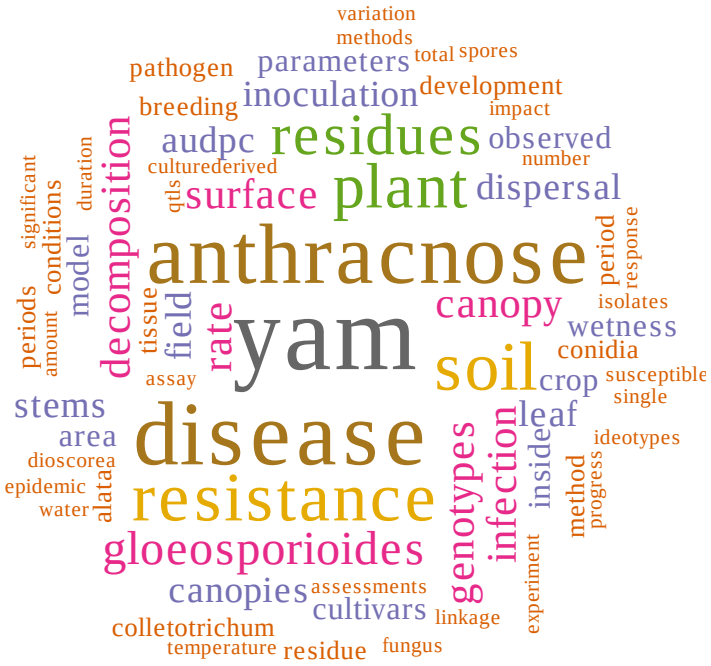
```
word_freqs <- sort(rowSums(tdm), decreasing=TRUE)
dm <- data_frame(word=names(word_freqs), freq=word_freqs)
dm$word <- factor(dm$word, levels=dm$word)
```

Plot Histogram of word frequencies (frequencies > 10 for this example), and then generate the wordcloud.

```
library(cowplot)
dm %>% filter(freq > 5) %>% ggplot(aes(word, freq)) +
  geom_bar(stat="identity") +
  #my_ggplot_theme() +
  theme_bw() +
  theme(axis.text.x=element_text(angle=45, hjust=1))
```




```
wordcloud(dm$word, dm$freq, random.order=F,
          colors=brewer.pal(8, "Dark2"),min.freq=4,
          scale=c(4,.03),rot.per=.15,max.words=100,
          family="Impact")
```



Some interesting fonts to try for wordclouds are: *Arvo, Asea, EB Garamond 12, EB Garamond 12 All SC, Swiss921 BT*

For Windows: *Showcard Gothic, Impact, Haettenschweiler, Gill Sans Ultra Bold Condensed, Franklin Gothic Heavy, Bernard MT Condensed, Agency FB, Coaster, Coaster Shadow, Spicy Rice*