Scraping a soccer fan forum

Marie Mortensen

10/30/2020

## Packages

pacman::p\_load(tidyverse, # General-purpose data wrangling  
 rvest, # Parsing of HTML/XML files   
 stringr,# String manipulation  
 rebus, # Verbose regular expressions  
 lubridate, # Eases DateTime manipulation  
 tidytext)

## Defining stopwords and functions for scraping

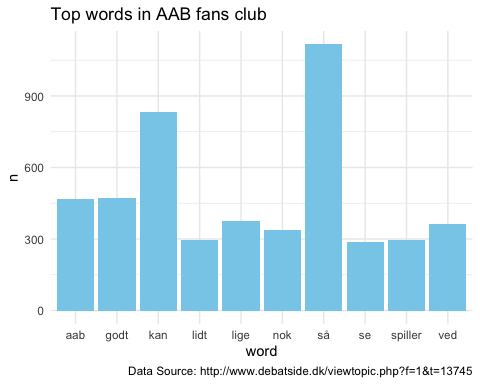
#reading stopwords list to remove uninteresting words  
dk\_stop <- data.frame(word = stopwords::stopwords("danish"))  
  
#pasting site where fanforum is  
url\_aab <- "http://www.debatside.dk/viewtopic.php?f=1&t=13745"   
# it looks like it says &start=20 for second page page and then +20 for every page  
  
aab <- read\_html(url\_aab)  
aab\_text <- aab %>%   
 html\_nodes('.content') %>% #defining what to extract from the website  
 html\_text() %>% #making it text  
 str\_trim() %>% #removing whitespaces  
 unlist() %>%   
 as.data.frame()  
  
all\_pages <- aab\_text  
pages <- seq(from =20, to =1980, by=20) #making a sequence from 20 to 1980 with breaks of 20 in between to define the different urls and pages to extract text from  
  
for (i in pages) { #this takes all the numbers in the list pages and performs the actions below  
 new\_url <- paste("http://www.debatside.dk/viewtopic.php?f=1&t=13745&start", i, sep = "=")  
   
 text <- new\_url %>%   
 read\_html() %>%   
 html\_nodes('.content') %>%   
 html\_text() %>%   
 str\_trim() %>%   
 unlist() %>%   
 as.data.frame()  
  
all\_pages <- rbind(all\_pages, text)  
}  
  
colnames(all\_pages) <- "words"  
all\_pages <- all\_pages %>% mutate(words = as.character(words))

## Unigram - inspecting what words are most frequent

sentences <- all\_pages %>%  
 str\_remove\_all("[/;:?.-]") %>%   
 str\_to\_lower() #lowercase

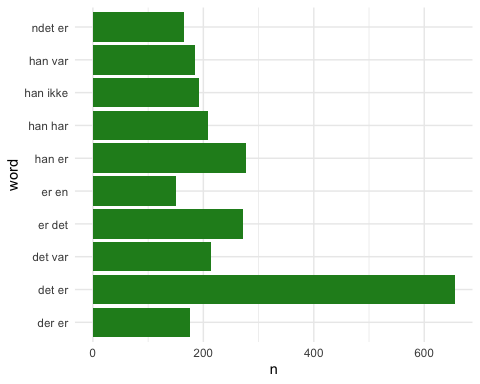
## Warning in stri\_replace\_all\_regex(string, pattern,  
## fix\_replacement(replacement), : argument is not an atomic vector; coercing

words <- sentences %>%   
 strsplit(" ") %>% #split every time there is whitespace  
 unlist() %>%   
 as.data.frame() %>%  
 na\_if("") %>% # recode empty strings "" by NAs  
 na.omit() # remove NAs  
  
colnames(words) <- "word"  
words <- words %>%   
 anti\_join(dk\_stop, by = "word") #removing stopwords  
  
words %>%   
 filter(word !='onclick=\\"windowopen(thishref)return') %>%   
 count(word, sort = TRUE) %>%   
 slice(1:10) %>%   
 ggplot() +   
 geom\_bar(aes(word, n),   
 stat = "identity",   
 fill = "skyblue") +  
 theme\_minimal() +  
 labs(title = "Top words in AAB fans club",  
 caption = "Data Source: http://www.debatside.dk/viewtopic.php?f=1&t=13745")



## Bigrams - which to words often occur together?

sentences <- as.character(sentences) %>% as.data.frame()  
colnames(sentences) <- "txt"   
  
 sentences %>%   
 # filter(txt != "") %>%   
 unnest\_tokens(word, txt, token = "ngrams", n = 2) %>% #split column into rows with two words per row  
 #below we test whether the words occur in the stoplists   
 separate(word, c("word1", "word2"), sep = " ") %>%   
 filter(!word1 %in% stop\_words$word) %>%  
 filter(!word2 %in% stop\_words$word) %>%   
 unite(word,word1, word2, sep = " ") %>%   
 filter(word != "windowopen thishref" &  
 word != "onclick windowopen" &   
 word != "return false" &   
 word !="thishref return" &  
 word !="www.bold.dk nyt") %>%   
 count(word, sort = TRUE) %>%   
 slice(1:10) %>% #take the 10 most frequent and plot  
 ggplot() + geom\_bar(aes(word, n), stat = "identity", fill = "forestgreen") +  
 theme\_minimal() +  
 coord\_flip()

 It does not seem like there is anything exciting on the fans site so far. It looks like there is a lot more preprocessing and removal of more stopwords needed (such as så, ved etc) before more interesting results can be seen.