Preprocessing text

José de Kruif

19 july 2021

This afternoon

Insight: Cleaning of data and text is a vitally important skill set.

Bag of Words

POStagging

Subjects

- Purposes of text cleaning
- Getting rid of noise and stop words
- Tokenizing and Ngrams
- Stemming and lemmatizing
- Language guesser
- POS tagging!

Install or call relevant libraries/packages

```
library (cld2)          # for language detection
library (corpus)        # for text analysis with support fo
library (dplyr)          # for data manipulation
library (entity)         # for easy NER
library (ggraph)        # for graphs
library (gridExtra)      # for working with grids to obtain
library (hunspell)      # for high-performance stemming
library (igraph)        # for easy graphs
library (janitor)       # for a pretty table
library (lattice)       # for easy charts
library (knitr)         # for manipulating the output of th
library (plyr)          # for data manipulation
library (NLP)           # for natural language processing t
library (openNLP)       # for named entity recognition
library (pander)        # for nice slide output
library (qdap)          # for using parsing tools to prepar
```

Point to the right working directory:

```
#setwd("C:/Noodfolder")
```

Read CSV file:

- Articles on John Maynard Keynes taken from The Times newspaper
- read.csv ("Keynes.csv")

```
Keynes <- read.csv("Keynes.csv", stringsAsFactors = FALSE)
```

Dataframe Times articles

EXAMPLE TEXT FIELD

- Title: Mr. Roosevelt's Experiments
- Publication date: February 27, 1940

```
TEXT <- "HOW TO PAY FOR THE 3WAR. By JOHN\nMAYNARD KEYNES.
\nIn three articles published i6n Tlhe Times last November
Keynes put forward what was expressly a first draft of prop
compulsory savings in  wartime. 3388 His plan  now reappear
of  a pamphlet in which Mr. Keynes elaborates his earlier a
varies it to meet the  comment and suggestion which it has
provoked.The pamphlet includes four appendices on the natio
extent of our resources abroad the cost of family allowance
for the aggregate of deferred pay and direct taxes\nthe\npa
is the subject of a leading article in 344 todays issue"
```

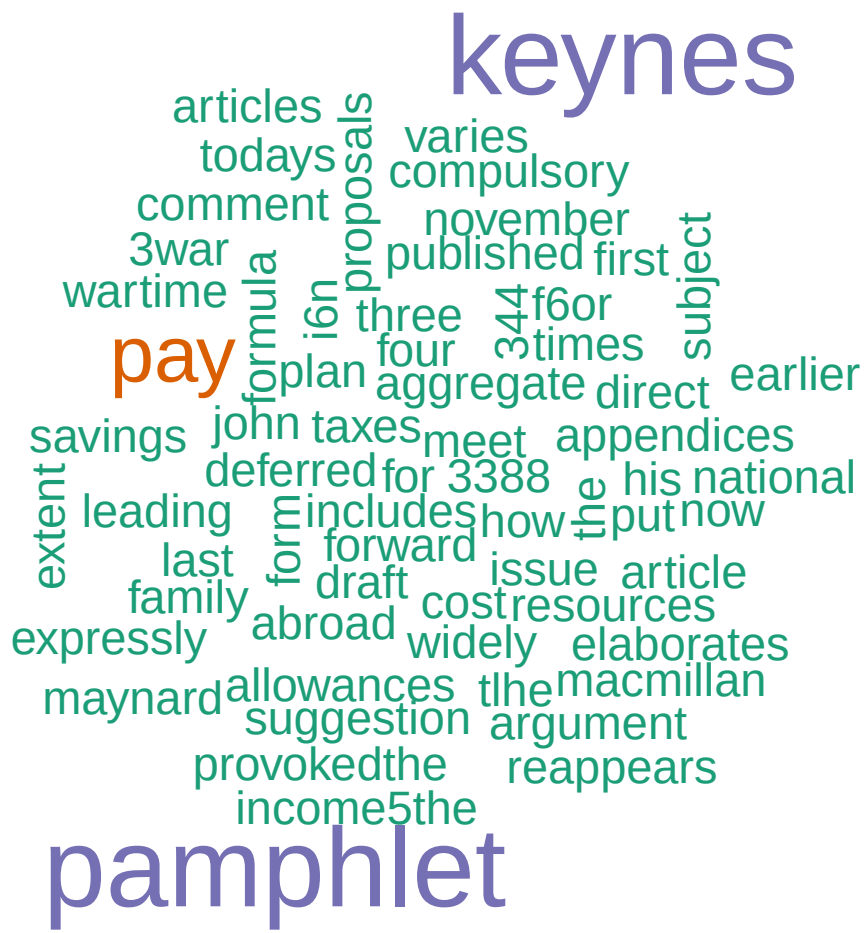
EXAMPLE TEXT FIELD Tokenized

```
TEXTtokens <- str_split(TEXT, " ")
TEXTtokens <- unlist(TEXTtokens)
print (TEXTtokens [1:50])
```

```
##  [1] "HOW"                "TO"                 "PAY"
##  [5] "THE"                "3WAR."              "By"
##  [9] "KEYNES."            "Macmillan."         "Is.\n\nIn"
## [13] "articles"           "published"          "i6n"
## [17] "Times"              "last"               "November"
## [21] "J."                 "M.\nKeynes"         "put"
## [25] "what"               "was"                "expressly"
## [29] "first"              "draft"              "of"
## [33] "f6or\ncompulsory"   "savings"            "in"
## [37] "wartime."           "3388"               "His"
## [41] ""                   "now"                "reappears"
## [45] "the"                "form\nof"           ""
## [49] "pamphlet"           "in"
```

Wordcloud of tokens

```
wordcloud(TEXTtokens, scale = c(4,.5), min.freq =1, colors
```



Clean text: all words in lower case

```
TEXT2 <- tolower(TEXT)
cat(TEXT2)
```

```
## how to pay for the 3war. by john
## maynard keynes. macmillan. is.
##
## in three articles published i6n tlhe times last novembe
## keynes put forward what was expressly a first draft of
## compulsory savings in  wartime. 3388 his plan  now reap
## of  a pamphlet in which mr. keynes elaborates his earli
## varies it to meet the  comment and suggestion which it
## provoked.the pamphlet includes four appendices on the n
## extent of our resources abroad the cost of family allow
## the formula
## for the aggregate of deferred pay and direct taxes
## the
## pamphlet
## is the subject of a leading article in 344 todays issue
```

Remove numbers

```
TEXT2 <- tm::removeNumbers(TEXT2)
cat(TEXT2)
```

```
## how to pay for the war. by john
## maynard keynes. macmillan. is.
##
## in three articles published in tlhe times last november
## keynes put forward what was expressly a first draft of p
## compulsory savings in  wartime.  his plan  now reappears
## of  a pamphlet in which mr. keynes elaborates his earlie
## varies it to meet the  comment and suggestion which it h
## provoked.the pamphlet includes four appendices on the na
## extent of our resources abroad the cost of family allowa
## the formula
## for the aggregate of deferred pay and direct taxes
## the
## pamphlet
## is the subject of a leading article in  todays issue
```

Remove double spaces

```
TEXT2<- str_replace_all(TEXT2, "  ", " ")
cat(TEXT2)
```

```
## how to pay for the war. by john
## maynard keynes. macmillan. is.
##
## in three articles published in tlhe times last november
## keynes put forward what was expressly a first draft of p
## compulsory savings in wartime. his plan now reappears in
## of a pamphlet in which mr. keynes elaborates his earlier
## varies it to meet the comment and suggestion which it ha
## provoked.the pamphlet includes four appendices on the na
## extent of our resources abroad the cost of family allowa
## the formula
## for the aggregate of deferred pay and direct taxes
## the
## pamphlet
## is the subject of a leading article in todays issue
```

Remove punctuation

```
TEXT2<-tm::removePunctuation(TEXT2)
cat(TEXT2)
```

```
## how to pay for the war by john
## maynard keynes macmillan is
##
## in three articles published in tlhe times last november
## keynes put forward what was expressly a first draft of p
## compulsory savings in wartime his plan now reappears in
## of a pamphlet in which mr keynes elaborates his earlier
## varies it to meet the comment and suggestion which it ha
## provokedthe pamphlet includes four appendices on the nat
## extent of our resources abroad the cost of family allowa
## the formula
## for the aggregate of deferred pay and direct taxes
## the
## pamphlet
## is the subject of a leading article in todays issue
```

Is the number of tokens reduced?

```
testtokens <- text_ntoken(TEXT)
print(testtokens)
```

```
## [1] 126
```

```
testtokens2 <- text_ntoken(TEXT2)
print(testtokens2)
```

```
## [1] 115
```

CLEAN THE TEXTS IN KEYNES DATAFRAME

```r
#All words in lower case
cleancontent <- tolower(Keynes$content)

## Remove numbers

cleancontent <- tm::removeNumbers(cleancontent)

## Remove double spaces

cleancontent<- str_replace_all(cleancontent, "  ", " ")

## Remove punctuation

cleancontent<-tm::removePunctuation(cleancontent)

## Store results in new column in the dataframe

Keynes$clean_content <- cleancontent
```

Reduction in the number of tokens?

```
testtokens <- text_ntoken(Keynes$content)
print (sum(testtokens))
```

```
## [1] 141015
```

```
testtokens2 <- text_ntoken(Keynes$clean_content)
print (sum(testtokens2))
```

```
## [1] 116667
```

stopwords

Store all too common words in a variable

```
#Stopwordslist

Joseesstopwords <- c("the", "a", "an", "and", "i","in", "ï
                     "him", "should", "or", "we", "no", "o
                     "could", "to", "of", "it", "is", "tha
                     "was", "/'s", "with","/'t", "as", "'t
                     "one", "so", "be","me", "are", "at", "
                     "by", "/", "for", "not", "from", "have
                     "but", "his","than", "their", "were",
                     "can", "what", "will", "would", "been
                     "more", "they", "there", "tibble", "it
                     "mrs", "mr",   "when", "all", "our", "
                     "who")
```

Remove these words:

```
Keynestext <- removeWords(Keynes$clean_content, Joseesstopw
#And store clean text in a new field
Keynes$clean_content <- Keynestext
```

Original content

Clean content

:::

STEMMING AND LEMMATIZATION

Stemming:

SnowballC library uses Porter's word stemming algorithm that collapses words to a common root. It supports many languages:

```
getStemLanguages()
```

```
##  [1] "arabic"     "basque"     "catalan"    "danish"
##  [6] "english"    "finnish"    "french"     "german"
## [11] "hindi"      "hungarian"  "indonesian" "irish"
## [16] "lithuanian" "nepali"     "norwegian"  "porter"
## [21] "romanian"   "russian"    "spanish"    "swedish"
## [26] "turkish"
```

Examples

```
wordStem("compulsary")
```

```
## [1] "compulsari"
```

```
wordStem("articles")
```

```
## [1] "articl"
```

```
wordStem("published")
```

```
## [1] "publish"
```

```
wordStem("November")
```

```
## [1] "Novemb"
```

Stem example text

```
## [1] "how to pay for the war by john\nmaynard keynes macm
```

how to pay for the war by john maynard keynes macmillan Is in three articles published in the times last november mr j m keynes put forward what was expressly a first draft of proposals for compulsory savings in wartime

how to pay for the war by john maynard keyn macmillan is in three articl publish in the time last novemb mr j m keyn put forward what was expressli a first draft of propos for compulsori save in wartim

Lemmatization:

Make lemma dictionary

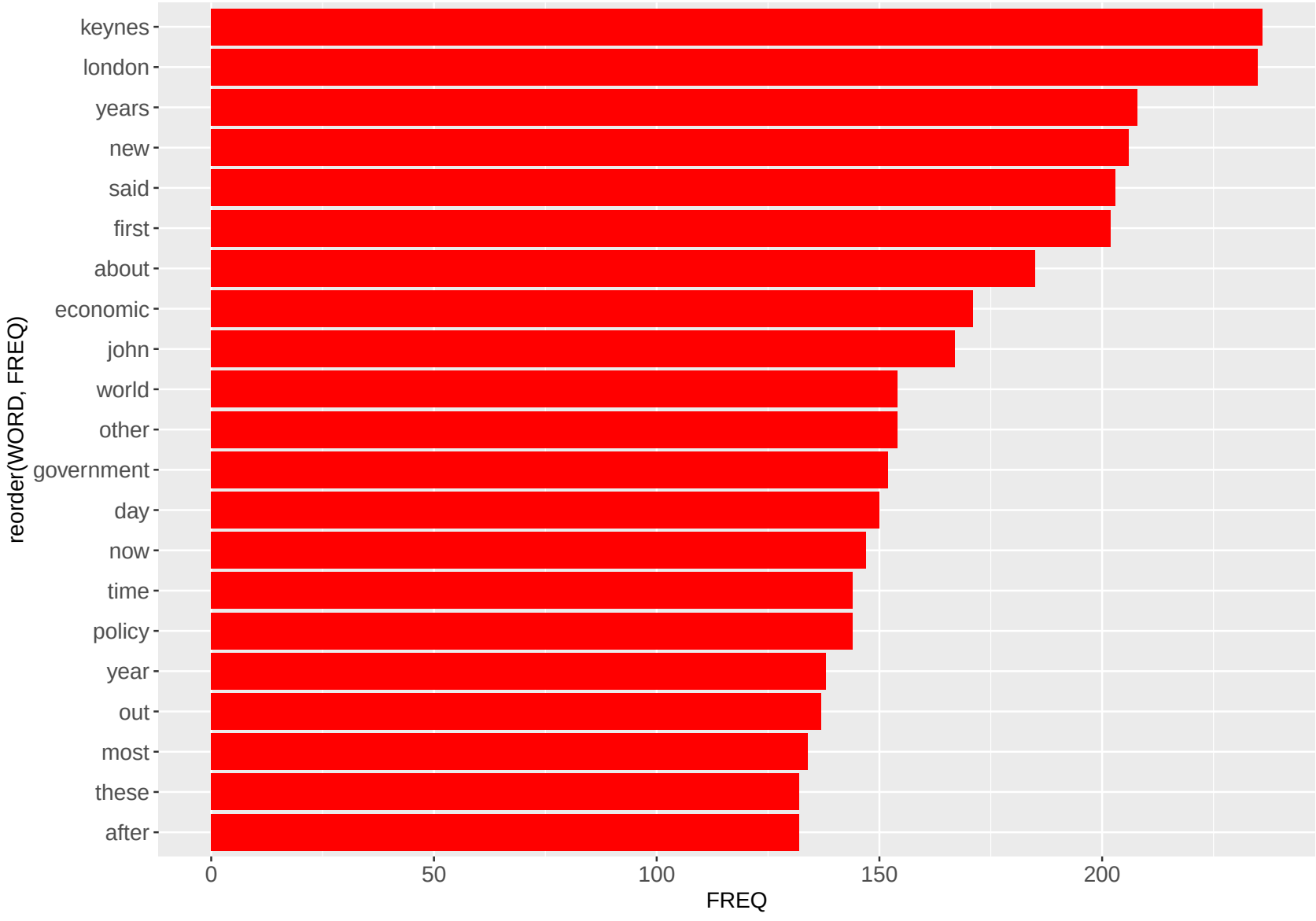
```
make_lemma_dictionary(TEXT2)
```

```
##          token     lemma
## 1           is         i
## 2     articles   article
## 3    published   publish
## 4        times      time
## 5    expressly   express
## 6    proposals  proposal
## 7      savings      save
## 8          his        hi
## 9    reappears    appear
## 10 elaborates elaborate
## 11     earlier     early
## 12      varies      vary
## 13     comment       com
## 14         has        ha
## 15      widely      wide
```

Lemmatize (cleaned) example text

```
Lemmatext <- lemmatize_strings(TEXT2, dictionary = lexicon
```

Keynes Top 20 of (meaningful) words



Language detection

Several language detectors available

- Package cld = Compact Language Detector (Google)
- several versions (2, 3)

Use CLD2 to detect language in Good Reads reviews

```
GRreviewssc <- read.csv("CLDdemo2.csv", sep = ";")
GRreviewssc$langdtcd <- detect_language(GRreviewssc$Text,
plain_text = TRUE, lang_code = TRUE)
```

POSTAGGING

Example text

```
## how to pay for the war by john
## maynard keynes macmillan is
##
## in three articles published in tlhe times last november
## keynes put forward what was expressly a first draft of p
## compulsory savings in wartime his plan now reappears in
## of a pamphlet in which mr keynes elaborates his earlier
## varies it to meet the comment and suggestion which it ha
## provokedthe pamphlet includes four appendices on the nat
## extent of our resources abroad the cost of family allowa
## the formula
## for the aggregate of deferred pay and direct taxes
## the
## pamphlet
## is the subject of a leading article in todays issue
```

Tagging of content field of Keynes dataframe (package: Udpipe)

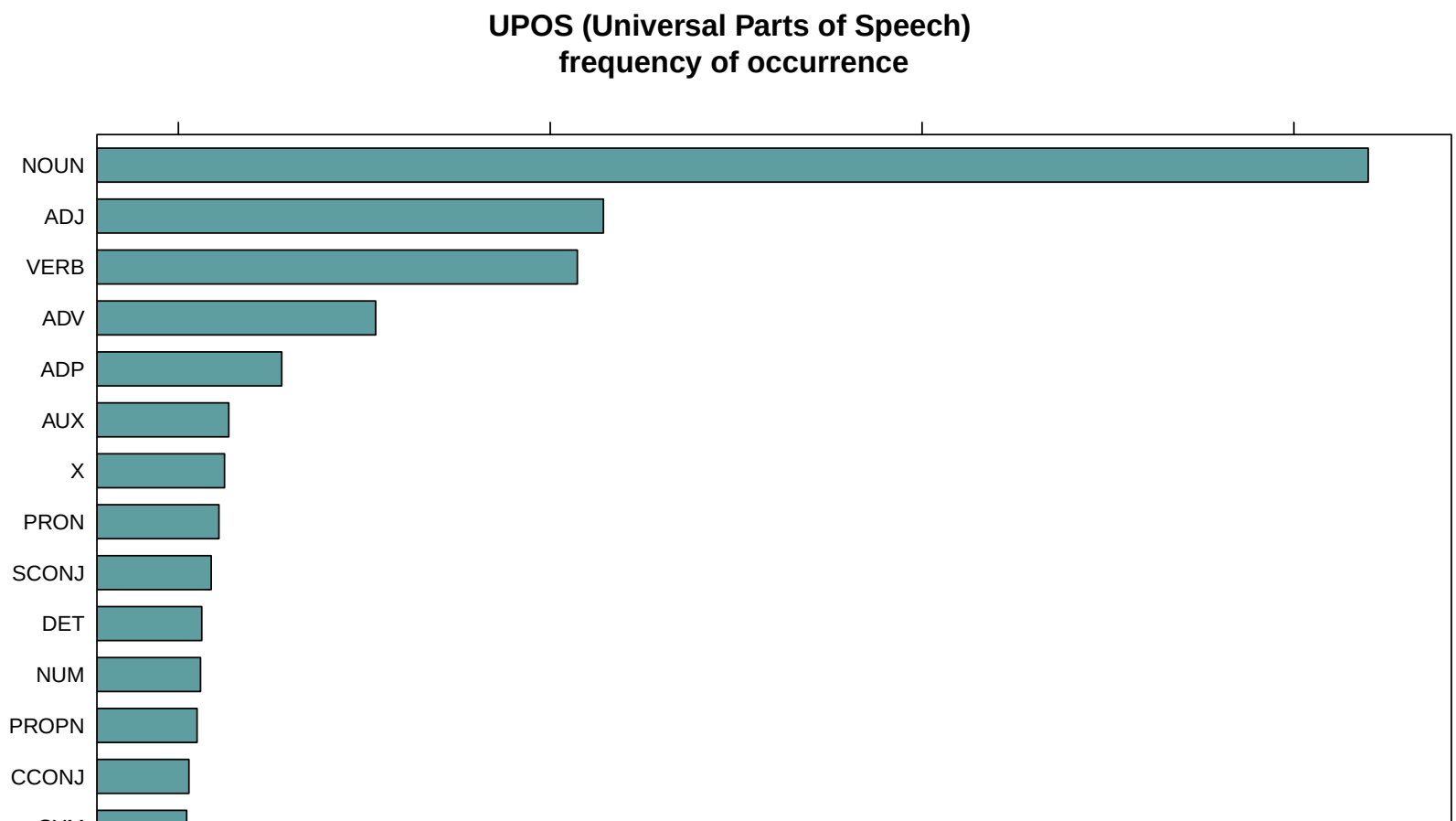
```
ud_model <- udpipe_download_model(language = "english")
```

Part of speech tagging of the clean content field of the Keynes dataframe.

```
ud_english <- udpipe_load_model(ud_model$file_model)
cleanfieldPOS <- udpipe_annotate(ud_english, x = Keynes$cle
cleanfieldPOS <- as.data.frame(cleanfieldPOS)
```

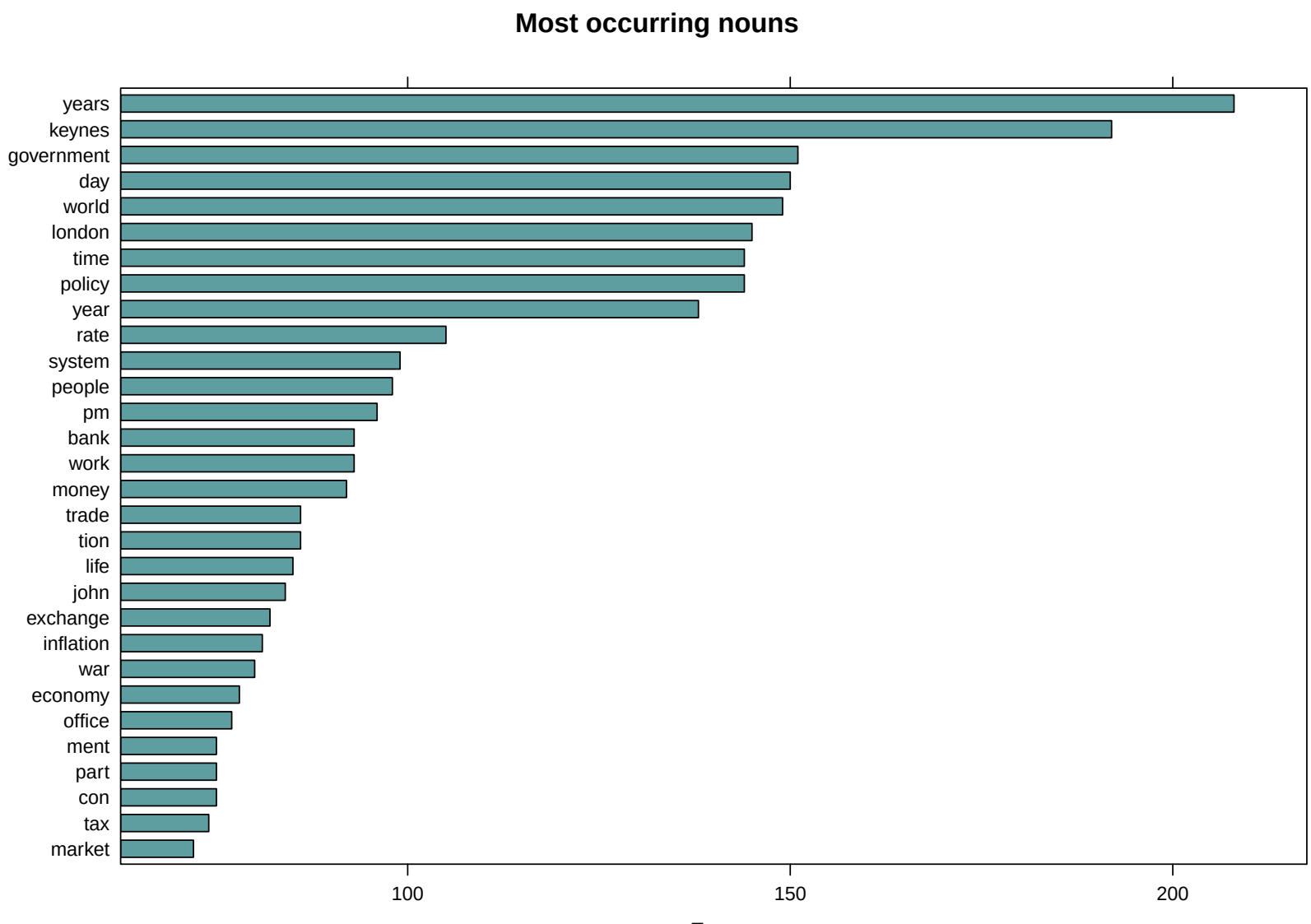
Results barchart

```
statsclean <- txt_freq(cleanfieldPOS$upos)
statsclean$key <- factor(statsclean$key, levels = rev(stats
barchart(key ~ freq, data = statsclean, col = "cadetblue",
main = "UPOS (Universal Parts of Speech)\n frequency of occ
 xlab = "Frequency")
```

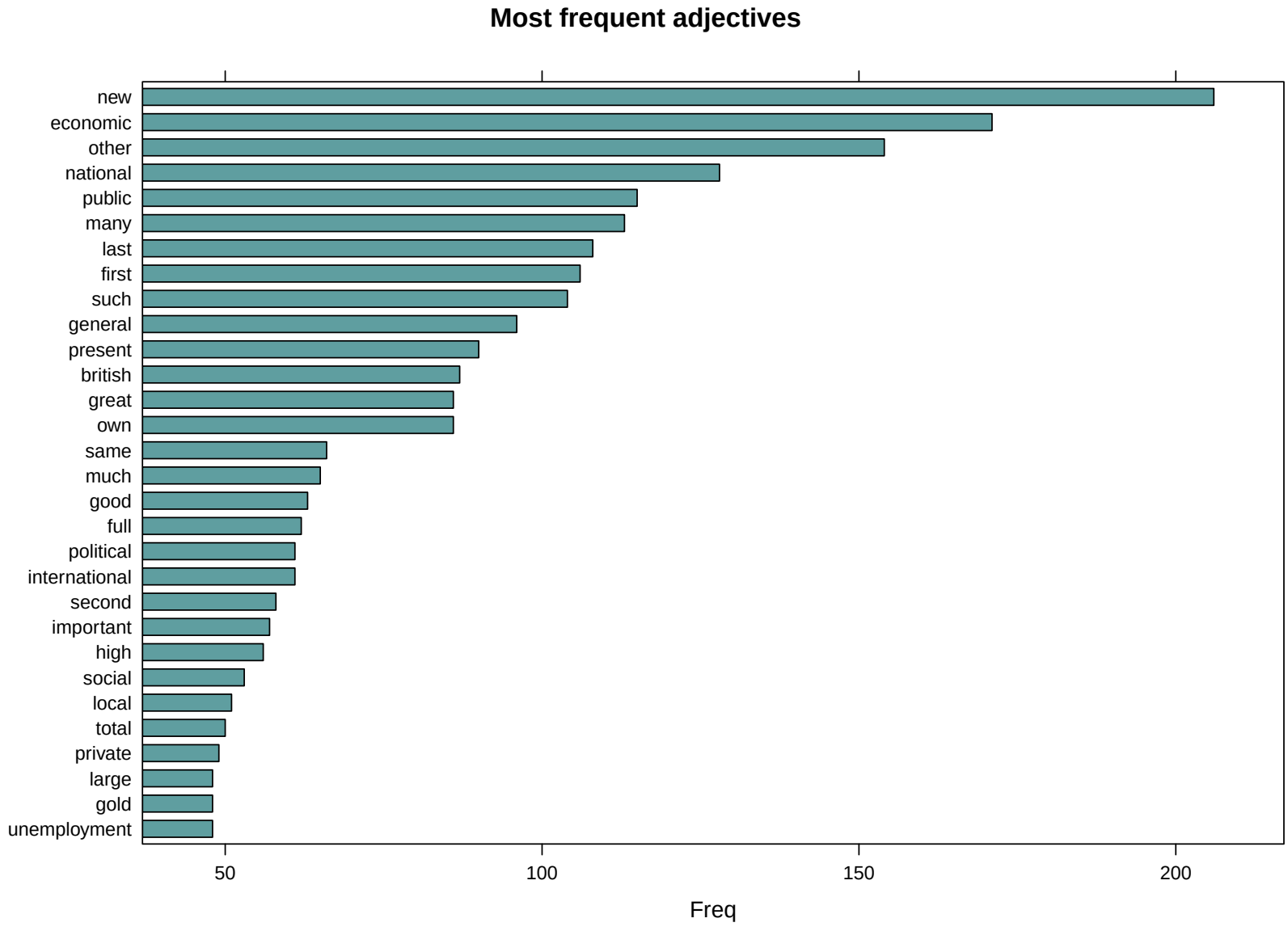


Which nouns are most important?

```
nounstat <- subset(cleanfieldPOS, upos %in% c("NOUN"))
nounstat <- txt_freq(nounstat$token)
```



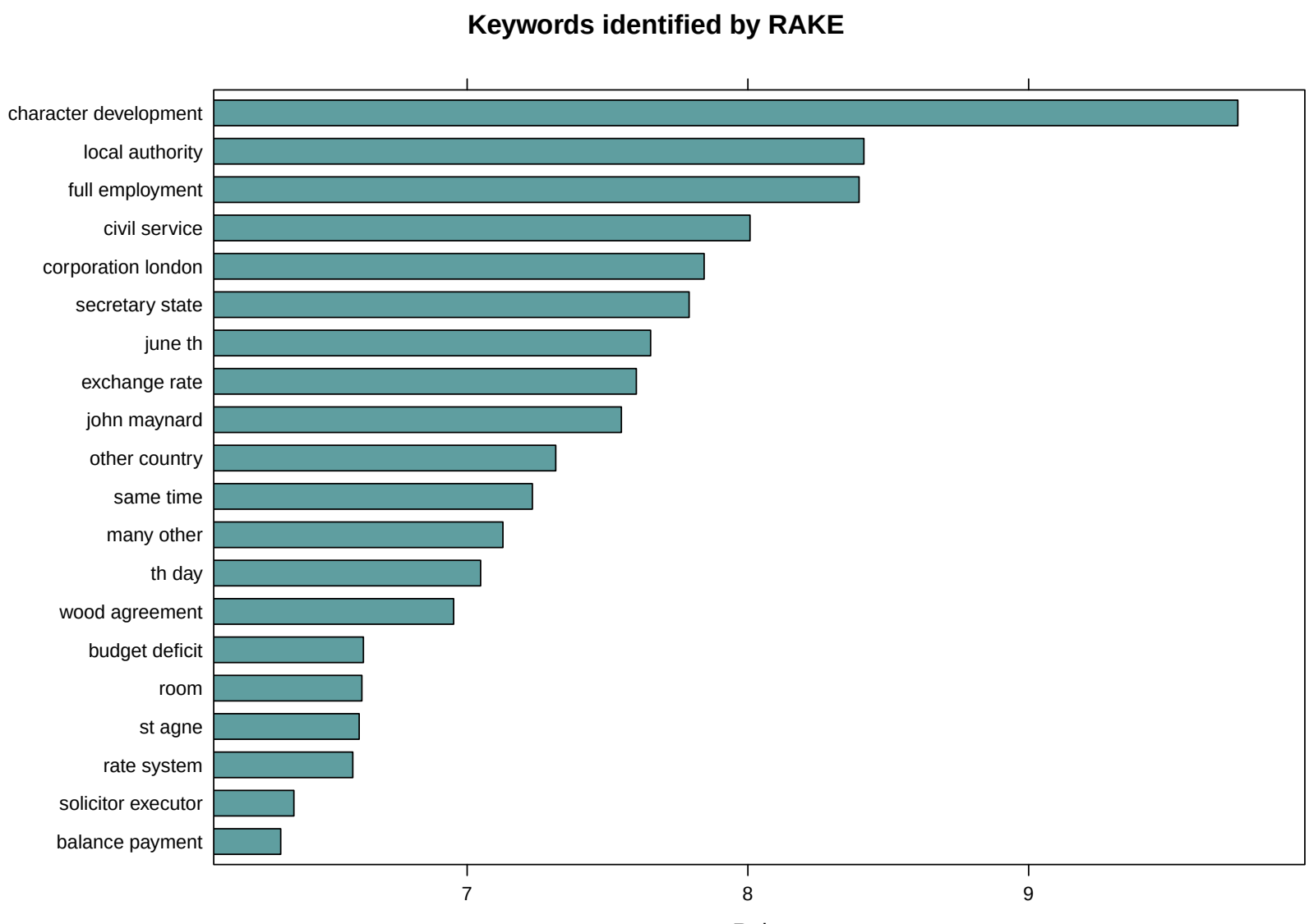
And adjectives:



Keyword combinations in text

Identifying keywords with RAKE

```
keyrake <- keywords_rake(x = cleanfieldPOS, term = "lemma",
                         relevant = cleanfieldPOS$upos %in% c
```



NER = Named Entity Recognition

Tags for NER

Correct NER tagging requires a model for the language used.
Wirtschafts universität in Vienna offers models on this website:
https://datacube.wu.ac.at/.

The Wikipedia text on the first election debate with Biden and Trump

```
President <- readLines('Debat.txt')
President <- as.String(President)
```

Wrapper around library

Open NLP needs the file to be tokenized, either in words or in sentences and the result will be stored as a list.

```
sent_ann <- Maxent_Sent_Token_Annotator()
```

Annotated text to replace the original

```
President_sent <- NLP::annotate(President, list(sent_ann))
President_nmbrd <- President[President_sent]
President_nmbrd
```

```
##  [1] "President <- \"Entering into the debate, Biden hac
##  [2] "Biden's lead was compounded by a funding shortage
##  [3] "Since Biden's successful nomination in the Democra
##  [4] "Trump called for Biden to be drug tested before th
##  [5] "Biden mocked the idea.[20]"
##  [6] "Trump also claimed that Biden would use a hidden e
##  [7] "Again, Biden declined."
##  [8] "Running up to the debate, Trump made repeated clai
##  [9] "When asked if he would commit to a peaceful transi
## [10] "In the weeks leading up to the debate, Trump becam
## [11] "Bob Woodward released his second book on the Trump
## [12] "In one recording made in February 2020, Trump indi
## [13] "Trump confirmed that he downplayed the severity of
## [14] "The New York Times published an investigation into
```

Questions?

Practical