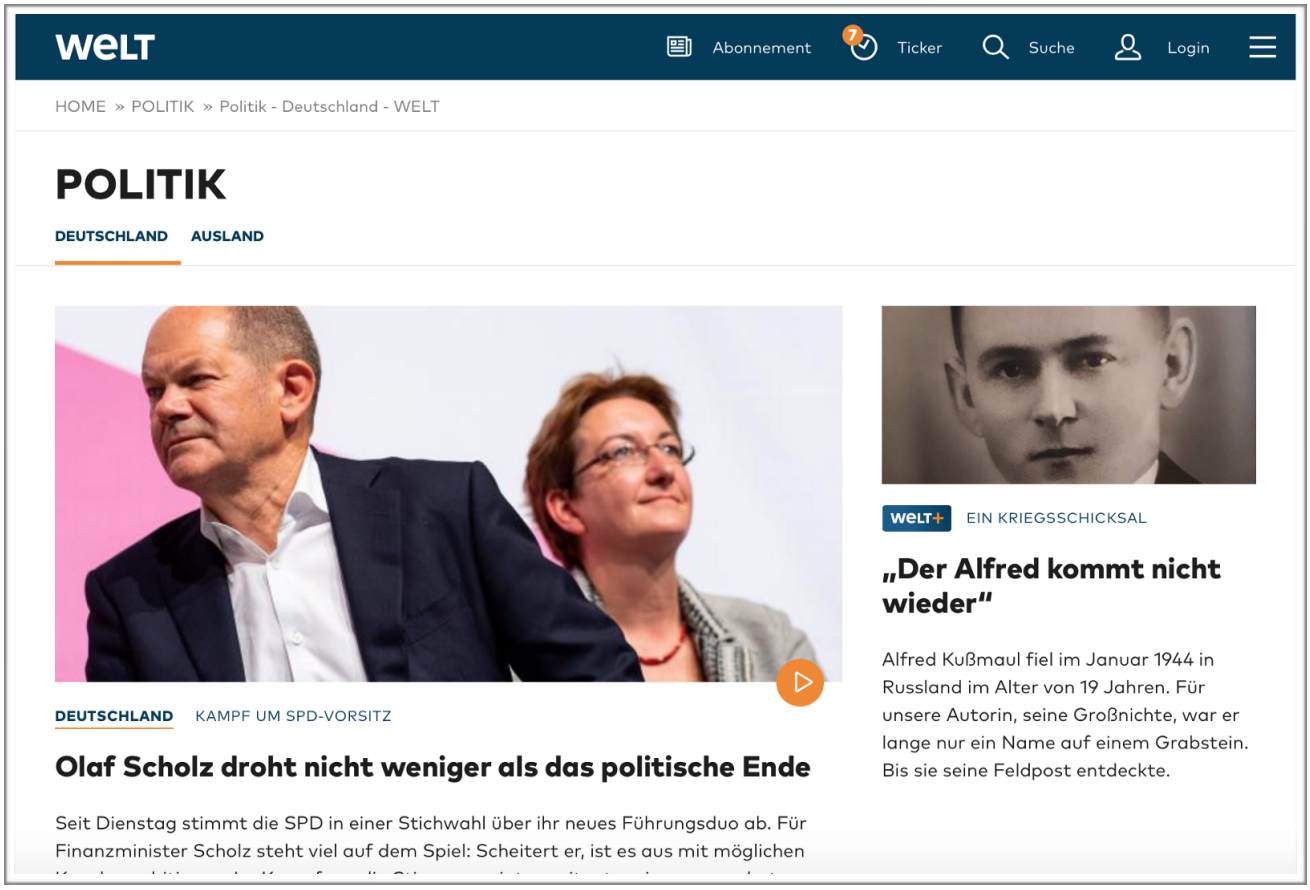
Web scraping individual assignment

Scraping a news site



https://www.welt.de/politik/deutschland/

Target: WELT.de

Aron Palkovics

2019. 12. 04.

Get up to date with German politics & people's reaction on it using R

I will combine AWS Machine Learning Services with the Scraping knowledge we learned

The idea:

Create a program what gets the author politically well informed about Germany and see the public's cumulated reaction on it, in a couple of minutes. The analysis based on the WELT.de news website using R-Studio.

1. Scrape the website to get the leading article titles and summaries with the related links, and article IDs
2. With AWS services translate it to english
3. Create a dataframe on it and download as csv
4. Create a function which scrapes every comments for the realated article, while translating and analyzing the sentiments of it (file name is article ID)
5. Summarize and cumulate the findings with merging it to the original news dataset
6. Create a function which reads out the leading article titles
7. Upload all files to Amazon S3 bucket

Let's see how it works in R:

1. Scrape the website to get the leading article titles and summaries with the related links, and article IDs

```
keyTable <- read.csv("accessKeys.csv", header = T) # accessKeys.csv == the CSV downloaded from AWS containing your Acces & Secret keys
AWS_ACCESS_KEY_ID <- as.character(keyTable$Access.key.ID)
AWS_SECRET_ACCESS_KEY <- as.character(keyTable$Secret.access.key)

#activate
Sys.setenv("AWS_ACCESS_KEY_ID" = AWS_ACCESS_KEY_ID,
           "AWS_SECRET_ACCESS_KEY" = AWS_SECRET_ACCESS_KEY,
           "AWS_DEFAULT_REGION" = "eu-west-1")
```

connect to AWS

```r
library(rvest)
library(data.table)


my_url <- "https://www.welt.de/politik/deutschland/"

t <- read_html(my_url)

my_titles <- t %>%
  html_nodes('.c-grid__item .o-teaser__link--is-headline') %>%
  html_text()

my_summary <- t %>%
  html_nodes('.c-teaser-default__intro') %>%
  html_text()

my_links <-
  paste0('https://www.welt.de/politik/deutschland',
         t %>%
           html_nodes('.c-grid__item .o-teaser__link--is-headline') %>%
           html_attr('href')
  )


library(stringr)

my_link_id <- my_links %>% str_match_all("[0-9]{9}+") %>% unlist %>% as.numeric
```

The actual scraping

2. With AWS services translate it to english

```r
#  translate it to english and create a dataframe on it
library("aws.translate")

for (i in 1:34) {
  my_titles[i] <-  translate(as.character(my_titles[i]), from = "de", to = "en")
  my_summary[i] <-  translate(as.character(my_summary[i]), from = "de", to = "en")
}
```

Translation

3. Create a dataframe on it and download as csv

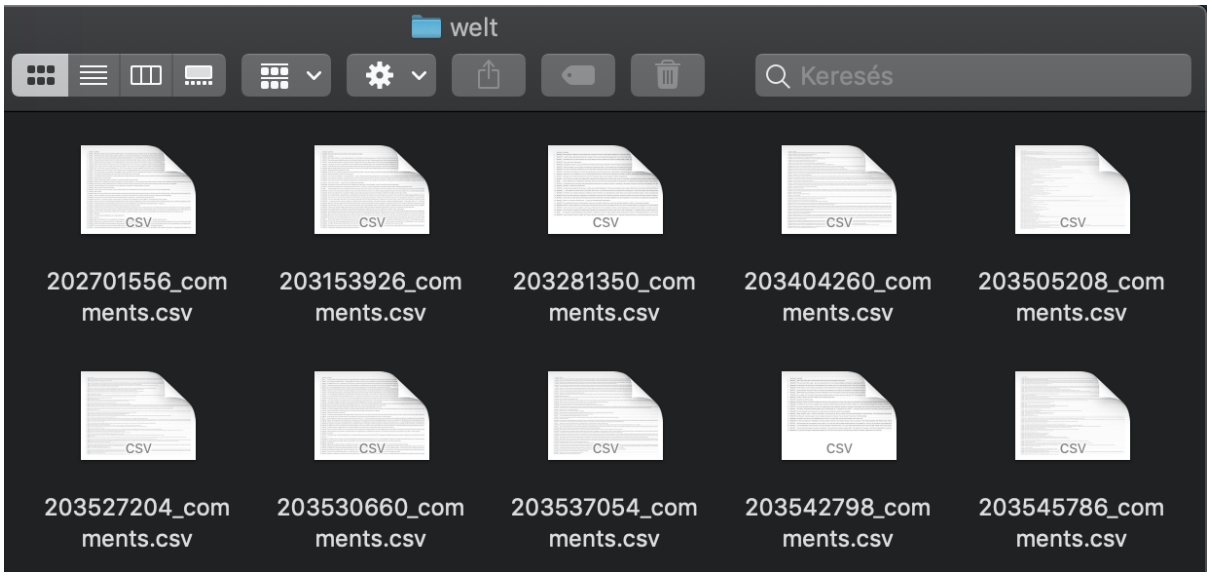
```r
welt_news <- data.frame('title'= my_titles,
                        'summary'=my_summary,
                        'links'=my_links,
                        'link_ID'=my_link_id)


View(welt_news)
```

	title	summary	links	link_ID
1	Laschet would be the candidate f...	When it comes to possible chancellor...	https://www.welt.de/politik/deutschland/politik/deu...	203621910
2	With the Greens, women decide ...	At the party congress in Bielefeld, th...	https://www.welt.de/politik/deutschland/politik/deu...	203624286
3	Federal Court of Auditors raises ...	According to a report by the Court of...	https://www.welt.de/politik/deutschland/politik/deu...	203623570
4	And then Merkel discovers the "n...	The Federal Government is committe...	https://www.welt.de/politik/deutschland/politik/deu...	203620526
5	After Green Vote, the way for Ke...	A clear majority of the Greens voted i...	https://www.welt.de/politik/deutschland/politik/deu...	203619744
6	The rental cover awakens fear of...	Contractors warn against Berlin's coll...	https://www.welt.de/politik/deutschland/politik/deu...	203537054
7	Why the state of Nato Macron wo...	The crisis belt from Libya to Nigeria i...	https://www.welt.de/politik/deutschland/politik/deu...	203572868
8	"People expect us to think bigger"	Before the CDU party congress, Phili...	https://www.welt.de/politik/deutschland/politik/deu...	203580470
9	One billion euros are to be stuffe...	More than one billion euros for mobi...	https://www.welt.de/politik/deutschland/politik/deu...	203598360
10	The demands of "Fridays For Fut...	Habeck and Baerbock manage to pos...	https://www.welt.de/politik/deutschland/politik/deu...	203593954
11	Even if Höcke-opponents are sus...	In Rhineland-Palatinate, AFD politicia...	https://www.welt.de/politik/deutschland/politik/deu...	203591396
12	Green adopt economic program ...	The Greens have strengthened their t...	https://www.welt.de/politik/deutschland/politik/deu...	203586398
13	Stegner makes a complaint again...	Former Schleswig-Holstein head of A...	https://www.welt.de/politik/deutschland/politik/deu...	203578204
14	Scholz demands legal entitlemen...	Citizens should "be able to do a new ...	https://www.welt.de/politik/deutschland/politik/deu...	203574854
15	"Macron wants to replace Nato. ...	CDU leader Annegret Kramp-Karrenb...	https://www.welt.de/politik/deutschland/politik/deu...	203561142

View data

4. Create a function which scrapes every comments for the related article, while translating and analyzing the sentiments of it

```
get_page_sentiment <- function(LINKID) {


  comments_data <- fromJSON(paste0('https://api-co.la.welt.de/api/comments?document-id=', LINKID, '&sort=NEWEST&limit=100'))

  number_of_comments <- length(comments_data$comments$id)
  my_comments <- head(as.character(comments_data$comments$contents), number_of_comments)

  detect_language(my_comments[1])

  sentiment <- c(1:number_of_comments)

  for (i in 1:number_of_comments) {
    my_comments[i] <-  translate(my_comments[i], from = "de", to = "en")
    sentiment[i] <- detect_sentiment(my_comments[i])[2]
  }

  sentiment <- unlist(sentiment)


  my_comments_data <- data.frame('sentiment'=sentiment,
                                 'comments'= my_comments)

  write.csv(my_comments_data, file = paste0("./welt/", LINKID, "_comments.csv"))
  comments_sentiment <- my_comments_data %>% group_by(sentiment) %>% summarise(n())

  negative <- comments_sentiment$`n()`[1]
  positive <- comments_sentiment$`n()`[2]
  neutral <- comments_sentiment$`n()`[3]
  mixed <- comments_sentiment$`n()`[4]

  link_sentiment <- data.frame('link_ID'= LINKID,
                               'num_of_comments'= number_of_comments,
             'negative'=negative,
             'positive'=positive,
             'neutral'=neutral,
             'mixed'=mixed)

   return(link_sentiment)


}
```



Comment Files with article IDs

202701556_comments

	sentiment	comments
1	NEUTRAL	World author Tim Röhn wrote down why he feels with the refugees — and why the borders have to be closed. An “If not, then things have to change quickly, then things have to be called by name: those who have the necessary In addition, asylum centres at the EU's external borders should be set up promptly, where asylum applications are You can read the entire article here: https://www.welt.de/debatte/plus180303378/Migrationskrise-Ich-fuehle-mit-den-Fluechtlingen-deswegen-schlies
2	NEUTRAL	There have always been economic refugees, migration consists only to a small extent of direct fleeing war, in esse What do we learn from this? If we really want to reduce migration, then we must help countries to control their ec The next effect is that the growth of the population is diminishing and entails a huge step towards reducing the gr After all, if these people want to achieve our prosperity and energy consumption, then we have to limit our comfo
3	NEUTRAL	Today's message: https://www.welt.de/regionales/nrw/article203330038/Schwarzfahrer-mit-fuenf-Identitaeten-ve Wanted to write “We can do this”, but again there was no comment field.
4	POSITIVE	This excellent series documents the failure of the Federal Government and the Federal Chancellor. In fact, even m
5	POSITIVE	Excellent, ideology-free documentation by an outstanding journalist. Information, careful research rather than “att
6	POSITIVE	Dear Mr. Aust, dear Mr. Büchel, Thank you very much for this informative series. There's a lot of work in it. This brings me directly to the subject: our politicians are making it too easy as they are putting more and more on

Comments and sentiments

5. Summarize and cumulate the findings with merging it to the original news dataset (Lapply in use)

```
my_page_sentiment <- lapply(my_link_id, get_page_sentiment)

final_df <- rbindlist(my_page_sentiment)
getwd()

welt_news_df <- left_join(final_df, welt_news, by = "link_ID")
welt_news_df <- welt_news_df %>% select(title, everything())
View(welt_news_df)

write.csv(welt_news_df, file = "./welt/welt_news_df.csv")
```

welt_news_df

	title	link_ID	num_of_comments	negative	positive	neutral	mixed	summary	links
1	Laschet would be the candidate for Jamaica	203621910	97	51	27	16	3	When it comes to possible ch	https://www.welt.de/politik
2	With the Greens, women decide whether to debate	203624286	144	47	66	29	2	At the party congress in Biele	https://www.welt.de/politik
3	Federal Court of Auditors raises serious accusations against Andreas Scheuer	203623570	62	31	25	1	5	According to a report by the (	https://www.welt.de/politik
4	And then Merkel discovers the “new oil”	203620526	131	59	51	20	1	The Federal Government is c	https://www.welt.de/politik
5	After Green Vote, the way for Kenya coalition in Brandenburg	203619744	143	36	83	23	1	A clear majority of the Greens	https://www.welt.de/politik
6	The rental cover awakens fear of Berlin's decay	203537054	48	3	17	25	3	Contractors warn against Ber	https://www.welt.de/politik
7	Why the state of Nato Macron worries	203572868	2	1	1	NA	NA	The crisis belt from Libya to N	https://www.welt.de/politik
8	“People expect us to think bigger”	203580470	115	1	31	30	53	Before the CDU party congre	https://www.welt.de/politik
9	One billion euros are to be stuffed with the funk-holes	203598360	134	61	56	13	4	More than one billion euros fo	https://www.welt.de/politik
10	The demands of “Fridays For Future” are applause, but no majority	203593954	106	50	38	16	2	Habeck and Baerbock manag	https://www.welt.de/politik
11	Even if Höcke-opponents are suspected of NPD	203591396	26	12	8	6	NA	In Rhineland-Palatinate, AFD	https://www.welt.de/politik
12	Green adopt economic program with eco-rules for the market	203586398	142	57	22	60	3	The Greens have strengthene	https://www.welt.de/politik
13	Stegner makes a complaint against ex-AFD politician Sayn-Wittgenstein	203578204	105	2	56	34	13	Former Schleswig-Holstein he	https://www.welt.de/politik
14	Scholz demands legal entitlement to a second training course	203574854	131	29	29	71	2	Citizens should “be able to do	https://www.welt.de/politik
15	“Macron wants to replace Nato. We want to strengthen them”	203561142	95	15	56	24	NA	CDU leader Annegret Kramp-	https://www.welt.de/politik
16	Seehofer's strength	203566948	129	6	63	44	16	The Ministry of the Interior ha	https://www.welt.de/politik
17	Unauthorized onward travel should lead to exclusion of social benefits	203574324	118	49	53	11	5	The Ministry of the Interior of	https://www.welt.de/politik
18	Berlin rental cover should violate the Basic Law	203564614	141	4	47	76	14	The Berlin CDU published an	https://www.welt.de/politik

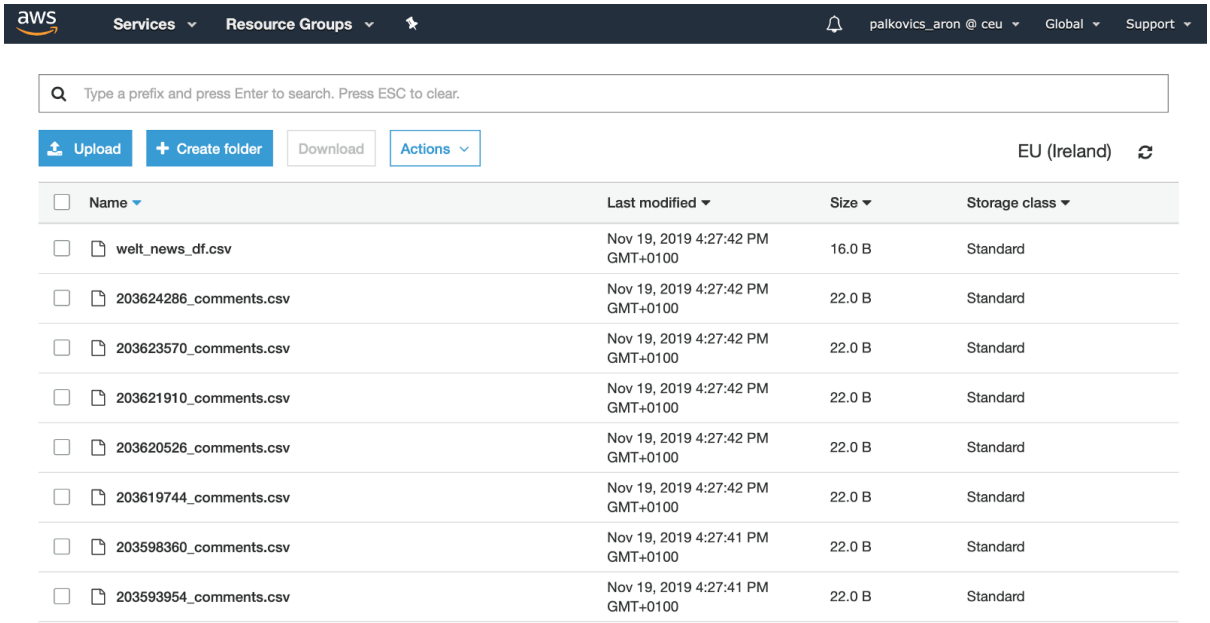
The final dataset

6. Create a function which reads out the leading article titles

```
#### READ TITLES

library("aws.polly")
library("tuneR")
setWavPlayer("afplay")
list_voices()


read <- function(text){
  read_in <- synthesize(text, voice = "Joanna")
  play(read_in)

}

my_titles
lapply(my_titles, read)
```

7. Upload all files to Amazon S3 bucket

```
##### CREATE BUCKET UPLOAD FILES

library(aws.s3)
bucketlist()

bucket_name <- "welt-news"


# Now, create the bucket on S3
put_bucket(bucket_name)


# Send the text file to your AWS S3 bucket

put_on_s3 <- function(filename) {

  put_object(filename, bucket = bucket_name)
}

welt_files <- list.files("./welt")


upload <- lapply(welt_files, put_on_s3)
```



END

The code is useful to refresh knowledge about German Politics every day, and get peoples reaction on it without knowing a single german word

(It's of course rather useful for learning scraping and R than get the actual news .. actually Welt.de site has an english version as well)