TALL SOM FORTELLER

Text analysis in R

SOLVEIG BJØRKHOLT

Statistisk sentralbyrå
Statistics Norway

Agenda

Who am I?

Text gathering

Text prepping

Text usage

Example from using unstructured text

Statistisk sentralbyrå
Statistics Norway

Who am I?

- Solveig Bjørkholt

- Certified RStudio Trainer

- Working at Statistics Norway

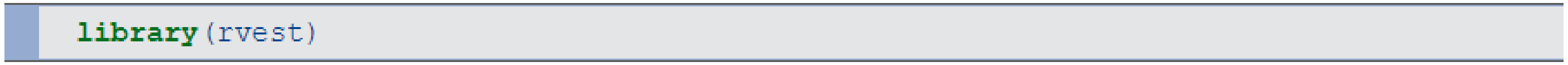
https://github.com/Zunny369

Fetch it

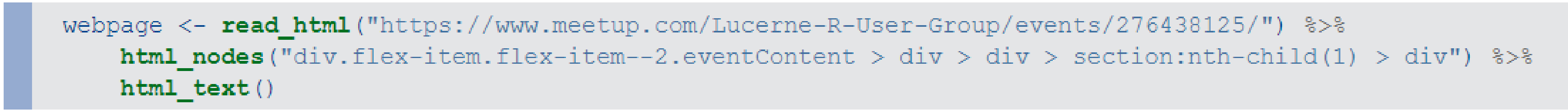
- You might have a .txt, .pdf, .docx, .json, .csv or any sort of file.
  - readtext

```r
list_of_files <- list.files("./path/to/dir", full.names = TRUE)
text <- readtext(list_of_files, encoding = "UTF-8")
```

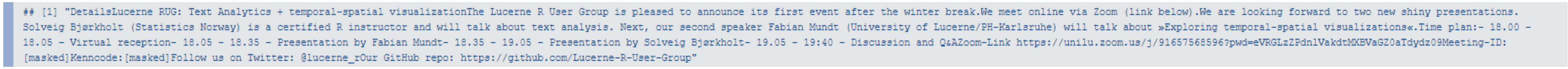
- For this example, I will use text from the Lucerne Group event webpage



```
library(rvest)
```



```
webpage <- read_html("https://www.meetup.com/Lucerne-R-User-Group/events/276438125/") %>%
    html_nodes("div.flex-item.flex-item--2.eventContent > div > div > section:nth-child(1) > div") %>%
    html_text()
```

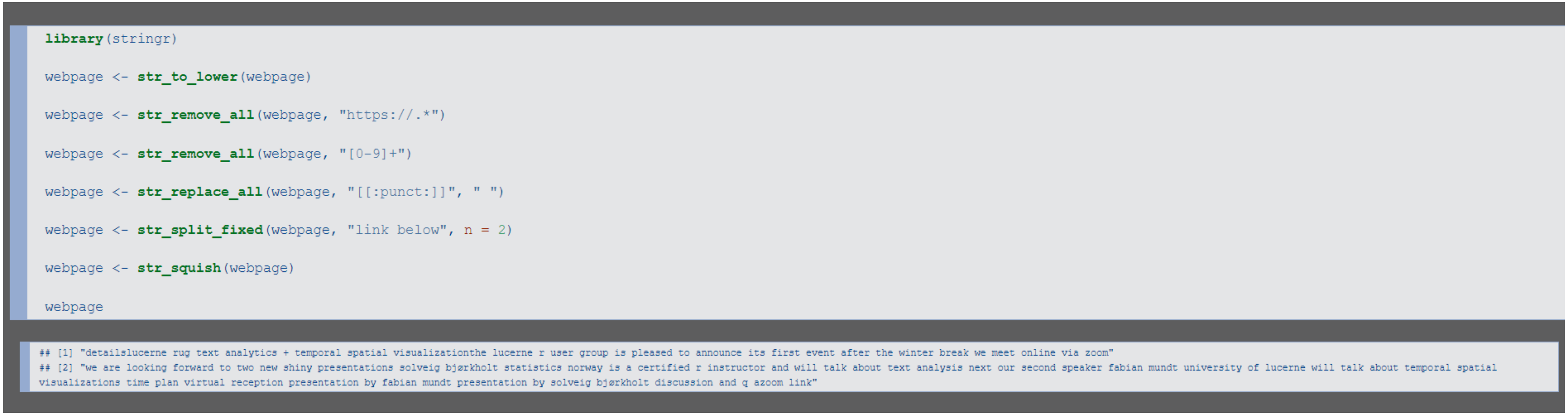


```
## [1] "DetailsLucerne RUG: Text Analytics + temporal-spatial visualizationThe Lucerne R User Group is pleased to announce its first event after the winter break.We meet online via Zoom (link below).We are looking forward to two new shiny presentations.
Solveig Bjørkholt (Statistics Norway) is a certified R instructor and will talk about text analysis. Next, our second speaker Fabian Mundt (University of Lucerne/PH-Karlsruhe) will talk about »Exploring temporal-spatial visualizations«.Time plan:- 18.00 -
18.05 - Virtual reception- 18.05 - 18.35 - Presentation by Fabian Mundt- 18.35 - 19.05 - Presentation by Solveig Bjørkholt- 19.05 - 19:40 - Discussion and Q&AZoom-Link https://unilu.zoom.us/j/91657568596?pwd=eVRGLzZPdnlVakdtMXBVaGZ0aTdydz09Meeting-ID:
[masked]Kenncode:[masked]Follow us on Twitter: @lucerne_rOur GitHub repo: https://github.com/Lucerne-R-User-Group"
```

Clean it

- What contains important meaning in your text?
  - Upper case? Lower case?
  - Stopwords?
  - Numbers?
  - Punctation?
  - Symbols?
  - URLs?
  - Email adresses?
  - Stem? Lemmatize?
  - Unigram, bigram, trigram?

- Packages:
  - stringr
  - stringi



```
library(stringr)

webpage <- str_to_lower(webpage)

webpage <- str_remove_all(webpage, "https://.*")

webpage <- str_remove_all(webpage, "[0-9]+")

webpage <- str_replace_all(webpage, "[[:punct:]]", " ")

webpage <- str_split_fixed(webpage, "link below", n = 2)

webpage <- str_squish(webpage)

webpage
```

```
## [1] "detailslucerne rug text analytics + temporal spatial visualizationthe lucerne r user group is pleased to announce its first event after the winter break we meet online via zoom"
## [2] "we are looking forward to two new shiny presentations solveig bjørkholt statistics norway is a certified r instructor and will talk about text analysis next our second speaker fabian mundt university of lucerne will talk about temporal spatial visualizations time plan virtual reception presentation by fabian mundt presentation by solveig bjørkholt discussion and q azoom link"
```

Vectorize it

```r
library(quanteda)
```

```r
dfm <- dfm(webpage)   # Make the string of text into a document feature matrix

dfm
```

- Bag of words

```r
dfm %>% convert(., to = "matrix") %>% knitr::kable()
```

	detailslucerne	rug	text	analytics	+	temporal	spatial	visualizationthe	lucerne	r	user	group	is	pleased	to	announce	its	first	event	after	the
text1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
text2	0	0	1	0	0	1	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0

- TF-IDF

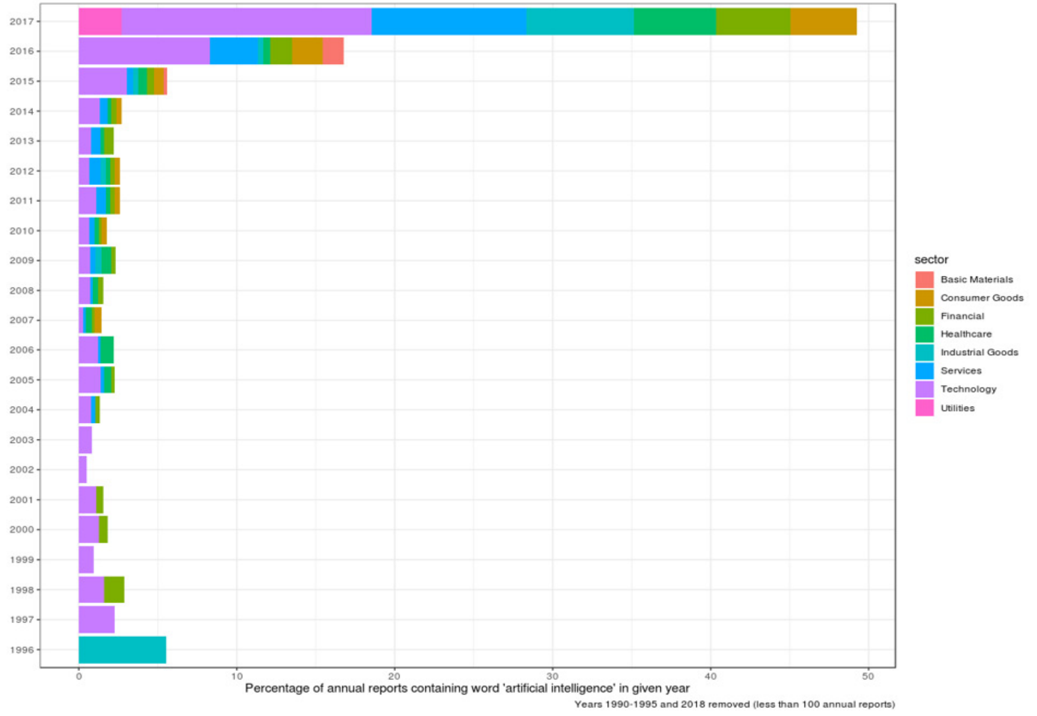
```r
dfm %>% dfm_tfidf() %>% convert(., to = "matrix") %>% knitr::kable()
```

	detailslucerne	rug	text	analytics	+	temporal	spatial	visualizationthe	lucerne	r	user	group	is	pleased	to	announce	its
text1	0.30103	0.30103	0	0.30103	0.30103	0	0	0.30103	0	0	0.30103	0.30103	0	0.30103	0	0.30103	0.30103
text2	0.00000	0.00000	0	0.00000	0.00000	0	0	0.00000	0	0	0.00000	0.00000	0	0.00000	0	0.00000	0.00000

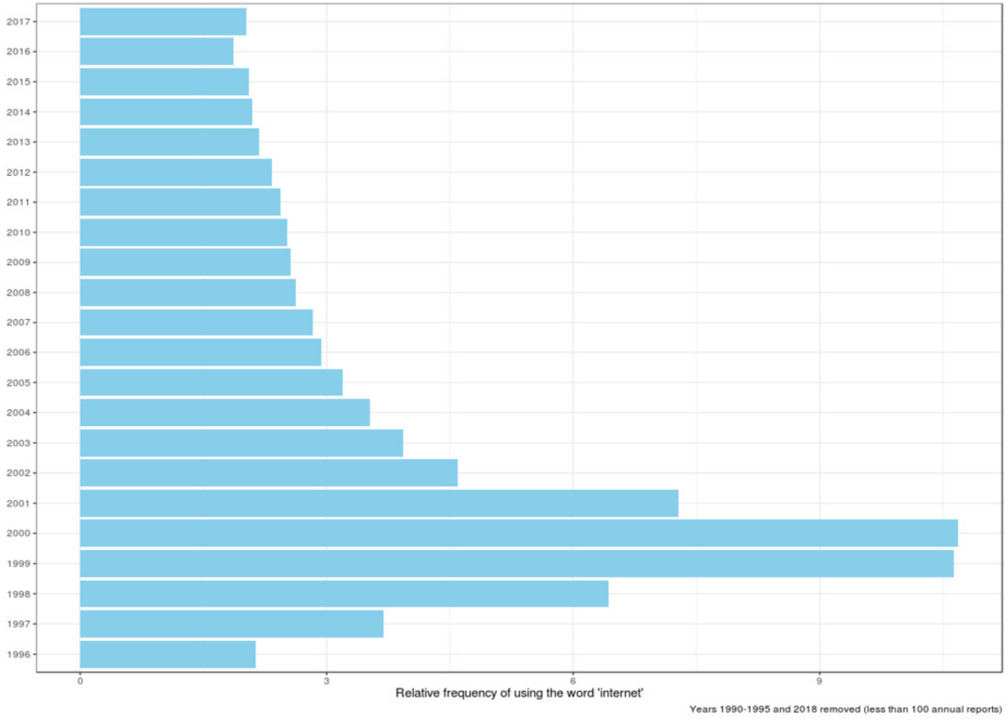
What can you do with this?

Descriptives

Artificial intelligence

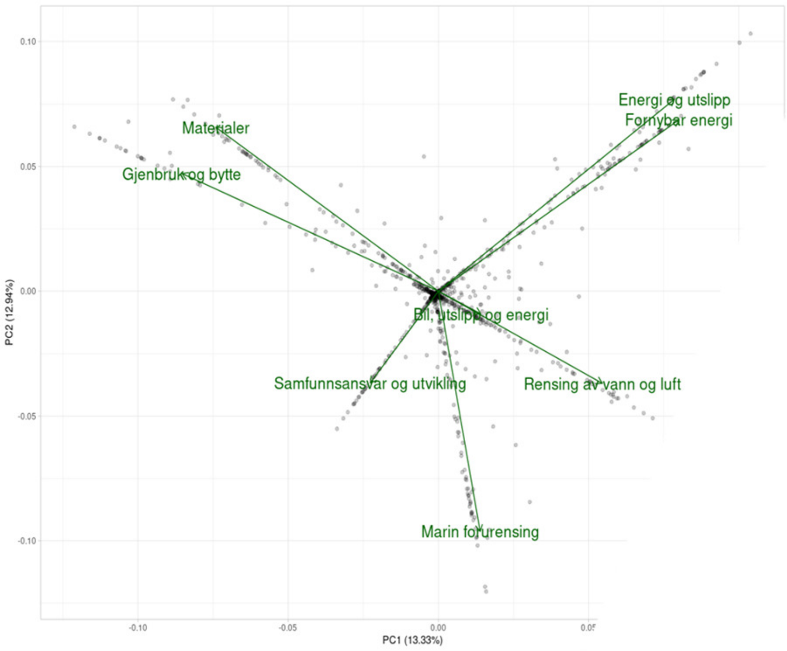


Internet

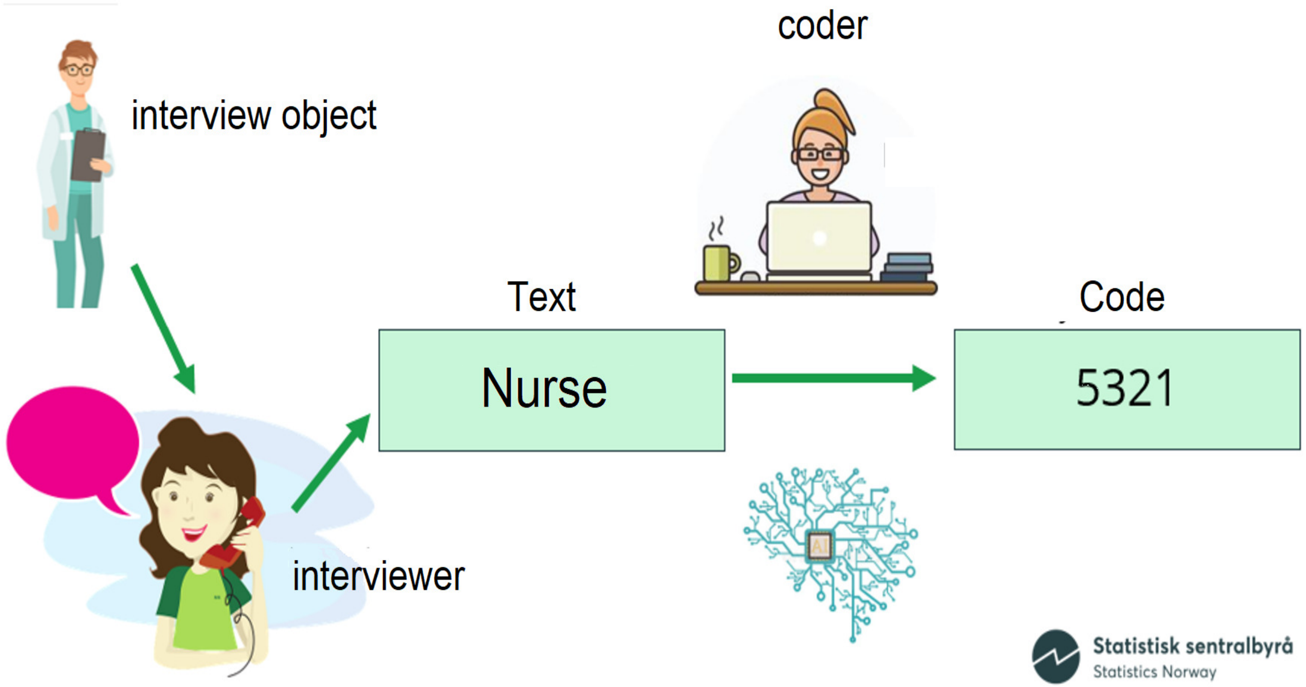


Classification

Unsupervised



Supervised

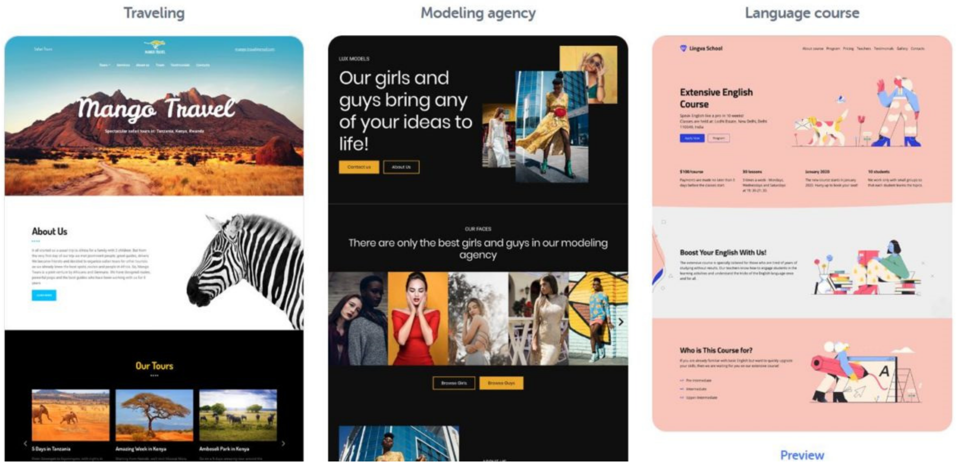


Example:
Unstructured data as a source of information

A world of data

Statistisk sentralbyrå
Statistics Norway

But what kind of data?



Traveling

Mango Travel

Spectacular safari tours in: Tanzania, Kenya, Rwanda

About Us

Our Tours

5 Days in Tanzania

Amazing Week in Kenya

Amboseli Park in Kenya

Modeling agency

LUX MODELS

Our girls and guys bring any of your ideas to life!

Contact us

About Us

OUR FACES

There are only the best girls and guys in our modeling agency

Browse Girls

Browse Guys

Language course

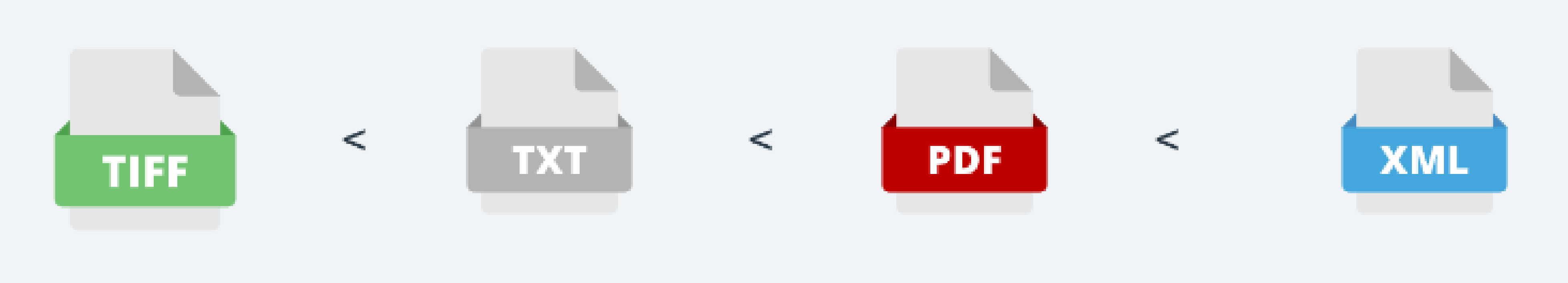
Lingva School

Extensive English Course

Boost Your English With Us!

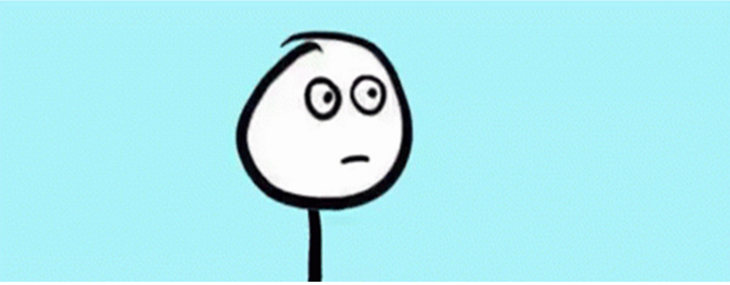
Who is This Course for?

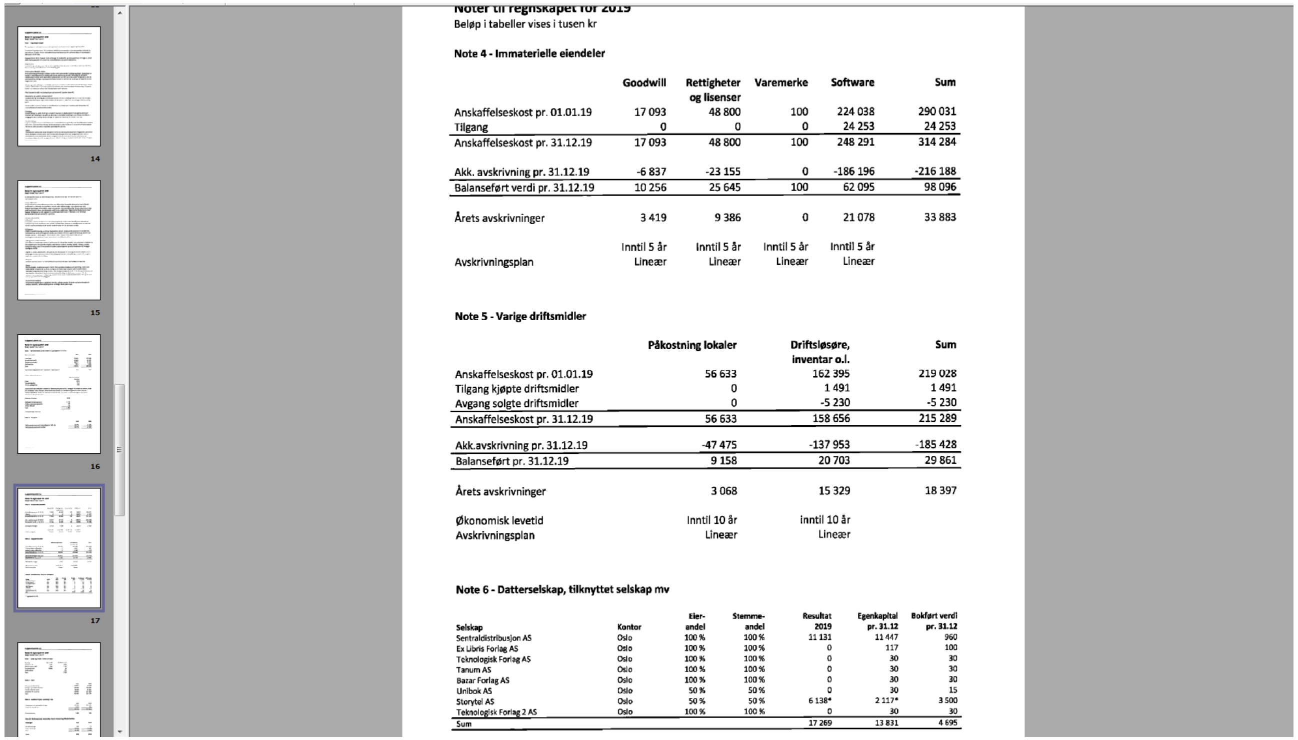
Preview



TIFF < TXT < PDF < XML

Archives with scanned pictures





Noter til regnskapet for 2019

Beløp i tabeller vises i tusen kr

Note 4 - Immaterielle eiendeler

	Goodwill	Rettigheter og lisenser	Varemerke	Software	Sum
Anskaffelseskost pr. 01.01.19	17 093	48 800	100	224 038	290 031
Tilgang	0	0	0	24 253	24 253
Anskaffelseskost pr. 31.12.19	17 093	48 800	100	248 291	314 284
Akk. avskrivning pr. 31.12.19	-6 837	-23 155	0	-186 196	-216 188
Balanseført verdi pr. 31.12.19	10 256	25 645	100	62 095	98 096
Årets avskrivninger	3 419	9 386	0	21 078	33 883
Avskrivningsplan	Inntil 5 år Lineær	Inntil 5 år Lineær	Inntil 5 år Lineær	Inntil 5 år Lineær	

Note 5 - Varige driftsmidler

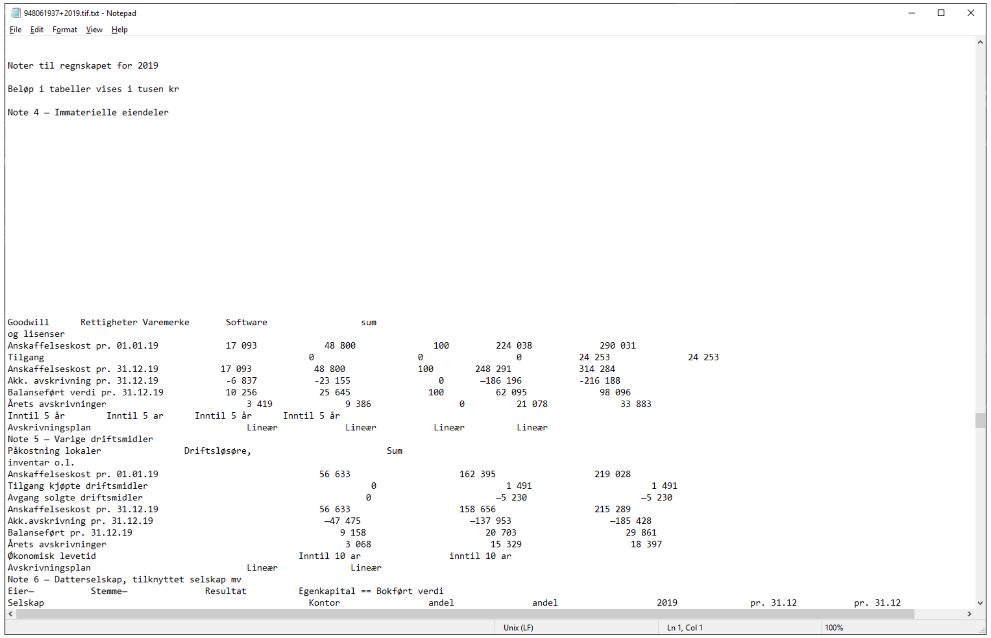
	Påkostning lokaler	Driftsløsøre, inventar o.l.	Sum
Anskaffelseskost pr. 01.01.19	56 633	162 395	219 028
Tilgang kjøpte driftsmidler	0	1 491	1 491
Avgang solgte driftsmidler	0	-5 230	-5 230
Anskaffelseskost pr. 31.12.19	56 633	158 656	215 289
Akk.avskrivning pr. 31.12.19	-47 475	-137 953	-185 428
Balanseført pr. 31.12.19	9 158	20 703	29 861
Årets avskrivninger	3 068	15 329	18 397
Økonomisk levetid Avskrivningsplan	Inntil 10 år Lineær	inntil 10 år Lineær	

Note 6 - Datterselskap, tilknyttet selskap mv

Selskap	Kontor	Eier-andel	Stemme-andel	Resultat 2019	Egenkapital pr. 31.12	Bokført verdi pr. 31.12
Sentraldistribusjon AS	Oslo	100 %	100 %	11 131	11 447	960
Ex Libris Forlag AS	Oslo	100 %	100 %	0	117	100
Teknologisk Forlag AS	Oslo	100 %	100 %	0	30	30
Tanum AS	Oslo	100 %	100 %	0	30	30
Bazar Forlag AS	Oslo	100 %	100 %	0	30	30
Unibok AS	Oslo	50 %	50 %	0	30	15
Storytel AS	Oslo	50 %	50 %	6 138*	2 117*	3 500
Teknologisk Forlag 2 AS	Oslo	100 %	100 %	0	30	30
Sum				17 269	13 831	4 695

Statistisk sentralbyrå
Statistics Norway

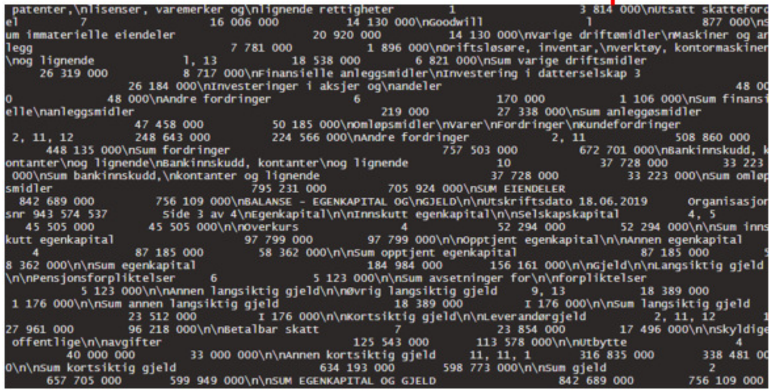
From picture to text



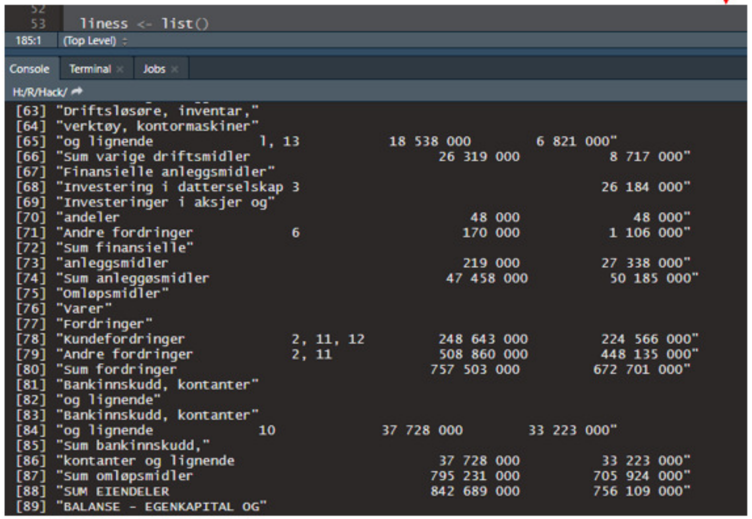
```
Noter til regnskapet for 2019

Beløp i tabeller vises i tusen kr

Note 4 – Immaterielle eiendeler









Goodwill      Rettigheter Varemerke       Software                  sum
og lisenser
Anskaffelseskost pr. 01.01.19             17 093             48 800               100         224 038             290 031
Tilgang                                                   0                    0                  0           24 253               24 253
Anskaffelseskost pr. 31.12.19            17 093            48 800              100        248 291             314 284
Akk. avskrivning pr. 31.12.19             -6 837            -23 155                0       –186 196            -216 188
Balanseført verdi pr. 31.12.19            10 256             25 645               100         62 095              98 096
Årets avskrivninger                           3 419              9 386                0          21 078              33 883
Inntil 5 år        Inntil 5 ar      Inntil 5 år      Inntil 5 år
Avskrivningsplan                              Lineær             Lineær            Lineær           Lineær
Note 5 – Varige driftsmidler
Påkostning lokaler                Driftsløsøre,                         Sum
inventar o.l.
Anskaffelseskost pr. 01.01.19                               56 633                     162 395                   219 028
Tilgang kjøpte driftsmidler                                           0                          1 491                        1 491
Avgang solgte driftsmidler                                           0                        –5 230                       –5 230
Anskaffelseskost pr. 31.12.19                               56 633                     158 656                   215 289
Akk.avskrivning pr. 31.12.19                                 –47 475                   –137 953                  –185 428
Balanseført pr. 31.12.19                                          9 158                       20 703                    29 861
Årets avskrivninger                                                3 068                      15 329                     18 397
Økonomisk levetid                                       Inntil 10 ar                 inntil 10 ar
Avskrivningsplan                              Lineær             Lineær
Note 6 – Datterselskap, tilknyttet selskap mv
Eier–           Stemme–              Resultat          Egenkapital == Bokført verdi
Selskap                                                  Kontor               andel               andel                    2019                pr. 31.12              pr. 31.12
```

From text to table

- *regex*
- dplyr
- stringr
- stringi
- lubridate



```
 patenter,\nlisenser, varemerker og\nlignende rettigheter        1         3 814 000\nUtsatt skatteford
el        7              16 006 000        14 130 000\nGoodwill             1              877 000\ns
um immaterielle eiendeler                  20 920 000        14 130 000\nvarige driftømidler\nMaskiner og an
legg                      7 781 000           1 896 000\nDriftsløsøre, inventar,\nverktøy, kontormaskiner
\nog lignende         1, 13           18 538 000       6 821 000\nSum varige driftsmidler
    26 319 000          8 717 000\nFinansielle anleggsmidler\nInvestering i datterselskap 3
                26 184 000\nInvesteringer i aksjer og\nandeler                                  48 00
0          48 000\nAndre fordringer            6              170 000          1 106 000\nSum finansi
elle\nanleggsmidler                           219 000         27 338 000\nSum anleggøsmidler
                47 458 000          50 185 000\nOmløpsmidler\nVarer\nFordringer\nKundefordringer
 2, 11, 12         248 643 000         224 566 000\nAndre fordringer        2, 11             508 860 000
      448 135 000\nSum fordringer                       757 503 000         672 701 000\nBankinnskudd, k
ontanter\nog lignende\nBankinnskudd, kontanter\nog lignende          10            37 728 000        33 223
 000\nSum bankinnskudd,\nkontanter og lignende              37 728 000         33 223 000\nSum omløp
smidler                        795 231 000         705 924 000\nSUM EIENDELER
  842 689 000        756 109 000\nBALANSE - EGENKAPITAL OG\nGJELD\n\nUtskriftsdato 18.06.2019      Organisasjon
snr 943 574 537       Side 3 av 4\nEgenkapital\n\nInnskutt egenkapital\n\nSelskapskapital        4, 5
  45 505 000        45 505 000\n\nOverkurs          4           52 294 000        52 294 000\n\nSum inns
kutt egenkapital                97 799 000         97 799 000\n\nOpptjent egenkapital\n\nAnnen egenkapital
   4              87 185 000        58 362 000\n\nSum opptjent egenkapital          87 185 000         5
8 362 000\n\nSum egenkapital                  184 984 000       156 161 000\n\nGjeld\n\nLangsiktig gjeld
\n\nPensjonsforpliktelser     6          5 123 000\n\nSum avsetninger for\n\nforpliktelser
          5 123 000\n\nAnnen langsiktig gjeld\n\nØvrig langsiktig gjeld     9, 13          18 389 000
 1 176 000\n\nSum annen langsiktig gjeld            18 389 000          I 176 000\n\nSum langsiktig gjeld
               23 512 000          I 176 000\n\nKortsiktig gjeld\n\nLeverandørgjeld          2, 11, 12         1
27 961 000        96 218 000\n\nBetalbar skatt          7            23 854 000        17 496 000\n\nSkyldige
 offentlige\n\navgifter                       125 543 000        113 578 000\n\nUtbytte              4
        40 000 000         33 000 000\n\nAnnen kortsiktig gjeld       11, 11, 1          316 835 000         338 481 00
0\n\nSum kortsiktig gjeld                    634 193 000        598 773 000\n\nSum gjeld              2
      657 705 000       599 949 000\n\nSUM EGENKAPITAL OG GJELD                  842 689 000         756 109 000
```



```
liness <- list()
```

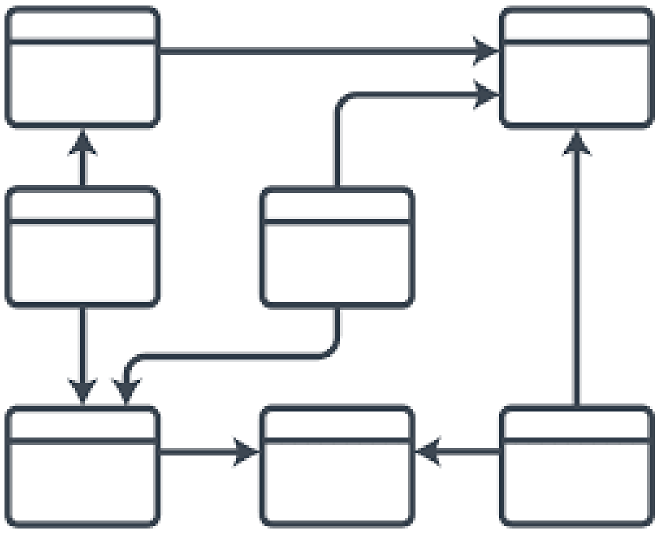
```
[63] "Driftsløsøre, inventar,"
[64] "verktøy, kontormaskiner"
[65] "og lignende          1, 13          18 538 000        6 821 000"
[66] "Sum varige driftsmidler                    26 319 000           8 717 000"
[67] "Finansielle anleggsmidler"
[68] "Investering i datterselskap 3                                   26 184 000"
[69] "Investeringer i aksjer og"
[70] "andeler                                        48 000              48 000"
[71] "Andre fordringer           6                  170 000           1 106 000"
[72] "Sum finansielle"
[73] "anleggsmidler                                 219 000          27 338 000"
[74] "Sum anleggøsmidler                         47 458 000           50 185 000"
[75] "Omløpsmidler"
[76] "Varer"
[77] "Fordringer"
[78] "Kundefordringer            2, 11, 12        248 643 000         224 566 000"
[79] "Andre fordringer           2, 11            508 860 000         448 135 000"
[80] "Sum fordringer                             757 503 000         672 701 000"
[81] "Bankinnskudd, kontanter"
[82] "og lignende"
[83] "Bankinnskudd, kontanter"
[84] "og lignende          10             37 728 000        33 223 000"
[85] "Sum bankinnskudd,"
[86] "kontanter og lignende                       37 728 000          33 223 000"
[87] "Sum omløpsmidler                           795 231 000         705 924 000"
[88] "SUM EIENDELER                              842 689 000         756 109 000"
[89] "BALANSE - EGENKAPITAL OG"
```

From table to database

- dbplyr
- RSQLite
- elastic
- *Elasticsearch*

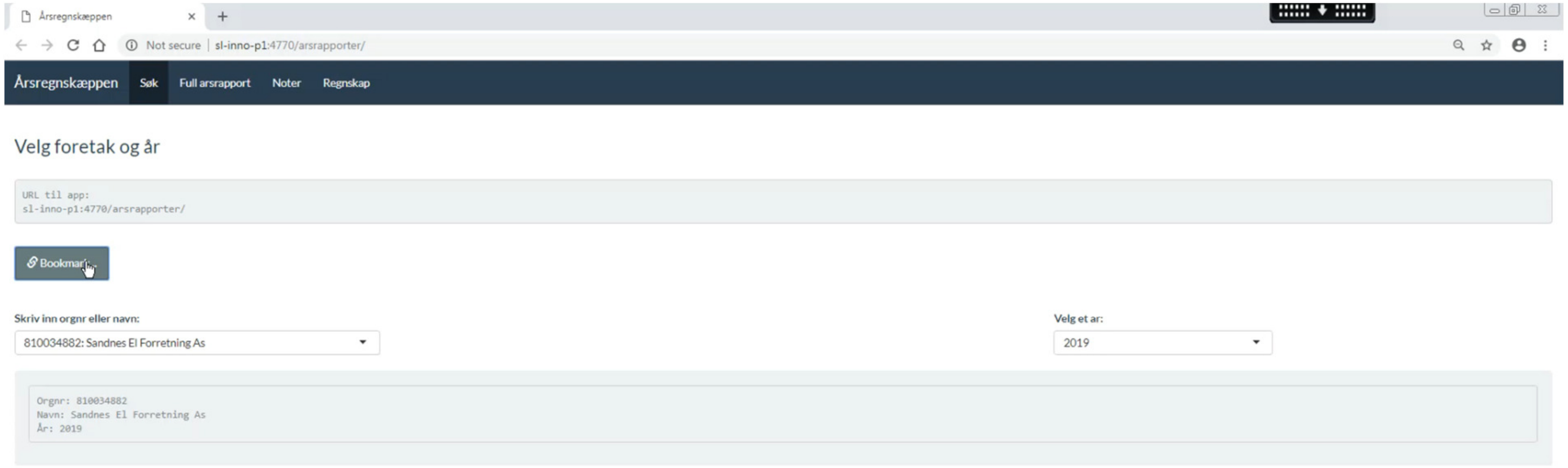
elastic

SQLite



From database to app

- Shiny
- data.table
- *Docker*



Velg foretak og år

```
URL til app:
sl-inno-p1:4770/arsrapporter/
```

Skriv inn orgnr eller navn: 810034882: Sandnes El Forretning As

Velg et ar: 2019

```
Orgnr: 810034882
Navn: Sandnes El Forretning As
År: 2019
```

TALL SOM FORTELLER

Thanks!

Statistisk sentralbyrå
Statistics Norway