Unstructured text as a source of information

SOLVEIG BJØRKHOLT



A world of data



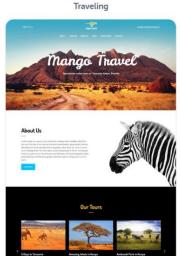


But what kind of data?



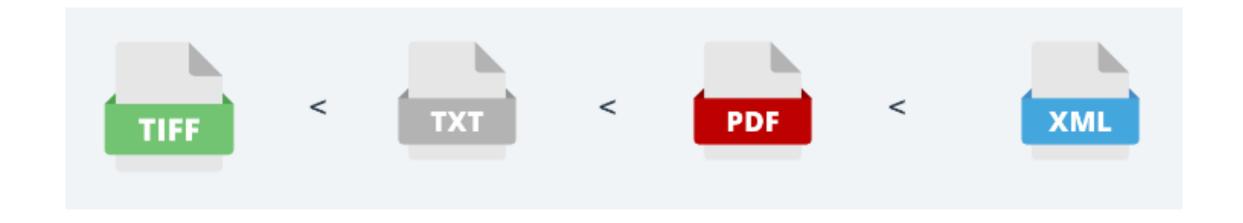






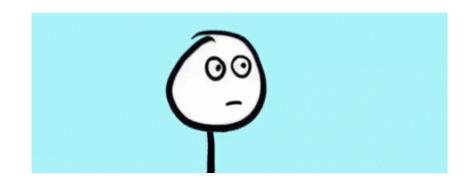


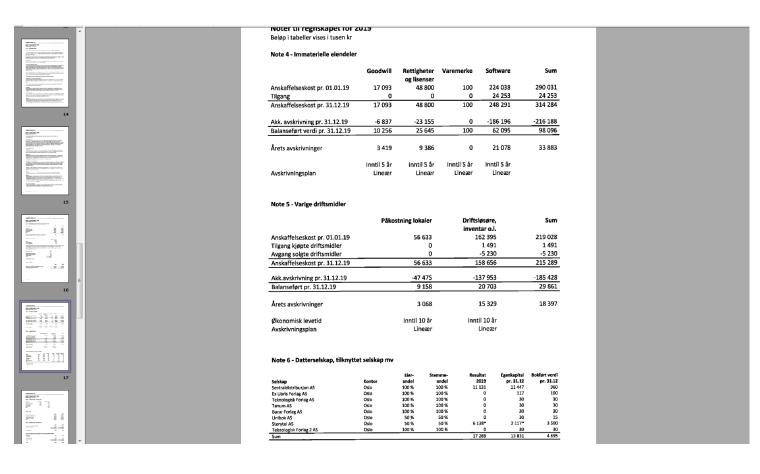






Archives with scanned pictures



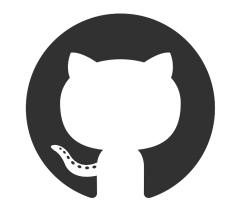


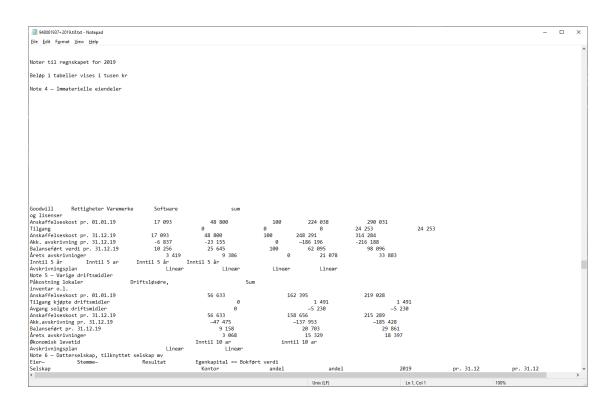


From picture to text











From text to table

- regex
- dplyr
- stringr
- stringi
- lubridate



```
14 130 000\nGoodwill 1 877 000\
920 000 14 130 000\nvarige driftomidler\nMaskiner og
 m immaterielle eiendeler
                                            20 920 000 14 130 000\nvarige driftomidler\nMaskiner og ar
1 896 000\nDriftsløsøre, inventar,\nverktøy, kontormaskiner
                                        18 538 000 6 821 000\nSum varige driftsmidler
 nog lignende
                         8 717 000\nFinansielle anleggsmidler\nInvestering i datterselskap 3
                 26 184 000\nInvesteringer i aksjer og\nandeler
               48 000\nAndre fordringer
                                                                                         1 106 000\nSum finans
                                                                       27 338 000\nSum anleggøsmidler
                  47 458 000
                                      50 185 000\nomløpsmidler\nvarer\nFordringer\nKundefordringer
                  248 643 000
                                      224 566 000\nAndre fordringer
     448 135 000\nSum fordringer
 ontanter\nog lignende\nBankinnskudd, kontanter\nog lignende
000\nSum bankinnskudd,\nkontanter og lignende
                                                                      37 728 000
                                                                                          33 223 000\nSum omle
\n\nSum kortsiktig gjeld
                      599 949 000\n\nSUM EGENKAPITAL OG GJELD
                                                                                                   756 109 000
```

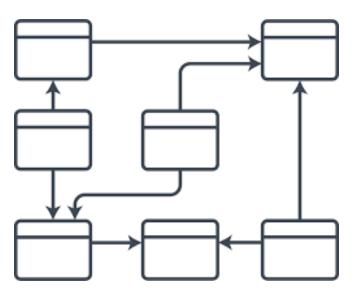
```
liness <- list()
[63] "Driftsløsøre, inventar,
    "verktøy, kontormaskiner'
                              1, 13
                                                18 538 000
    "og lignende
                                                                   6 821 000"
    "Sum varige driftsmidler
"Finansielle anleggsmidler"
                                                       26 319 000
                                                                             8 717 000"
     "Investering i datterselskap 3
                                                                            26 184 000"
     "Investeringer i aksjer og'
                                                           48 000
                                                                                48 000"
     "andeler
     "Andre fordringer
"Sum finansielle"
                                                          170 000
                                                                             1 106 000"
                                                                            27 338 000"
     "anleggsmidler
                                                          219 000
     "Sum anleggøsmidler
                                                        47 458 000
                                                                              50 185 000"
     "Omløpsmidler"
     "varer"
     "Fordringer'
     "Kundefordringer
                                   2, 11, 12
                                                       248 643 000
                                                                            224 566 000"
     "Andre fordringer
                                   2, 11
                                                       508 860 000
                                                                            448 135 000"
     "Sum fordringer
                                                      757 503 000
                                                                           672 701 000"
     "Bankinnskudd, kontanter"
     "og lignende'
     "Bankinnskudd, kontanter"
     "og lignende
                                               37 728 000
                                                                  33 223 000"
     "Sum bankinnskudd,"
                                                                            33 223 000"
     "kontanter og lignende
                                                       37 728 000
     "Sum omløpsmidler
                                                      795 231 000
                                                                           705 924 000"
     "SUM EIENDELER
                                                                           756 109 000"
                                                      842 689 000
     "BALANSE - EGENKAPITAL OG"
```



From table to database

- dbplyr
- RSQLite







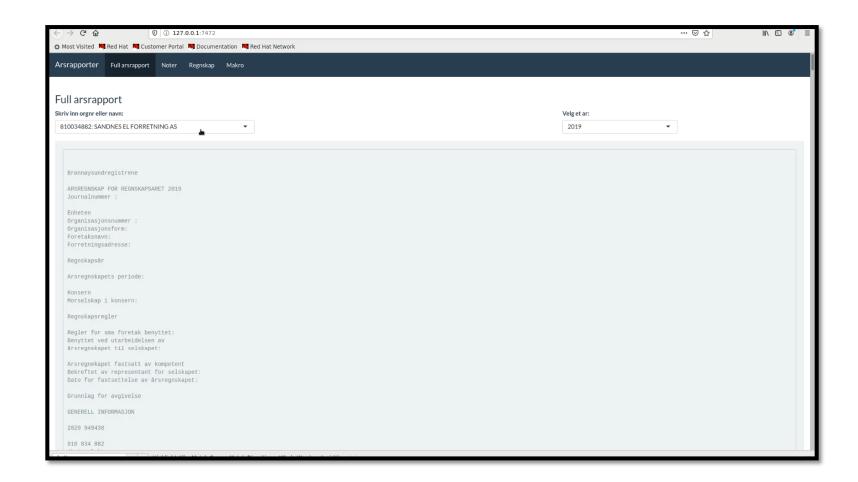
From database to app

- Shiny
- data.table
- Docker











Thanks!

