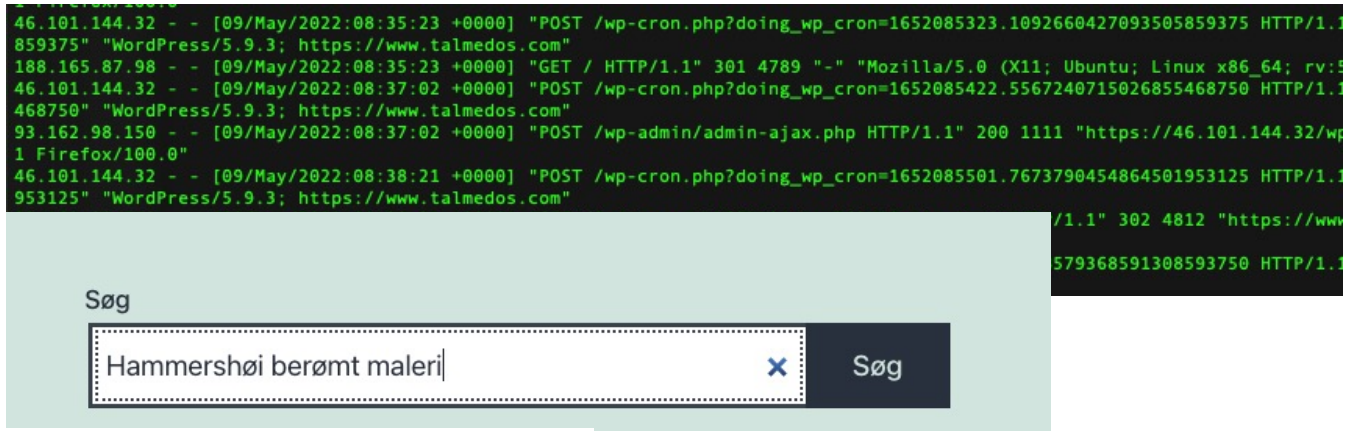
# Textmining with R – regular expressions

- Parse apache-logfil




- Parse html

**Parkvej 23,** 6510 Gram
E Villa fra EDC

m²
129

Dejlig villa beliggende på stille og rolig villavej

Beliggende på en stille og rolig villavej finder du denne fine gulstensvilla med eternittag fra omkring 2009 og trætermo vinduer, som er delvis skiftet omkring 2009. Villaen er opført i 1965 og er meget velholdt og indbydende. Der findes egen trappe med nedgang til Slotsvej, og dermed er der kort afstand til byens skole, indkøb og idrætsfaciliteter. Her er kort afstand til naturen omkring Gram Slot.

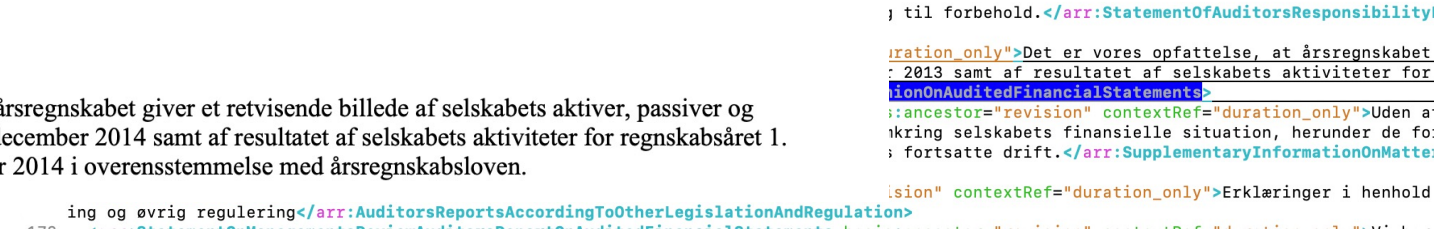
Tilhørende er en carport på 31 m² med plads til 2 biler. Derudover en god grusbelagt indkørsel med plads til familiens køretøjer.



- Parse Regnskab

**Konklusion**

Det er vores opfattelse, at årsregnskabet giver et retvisende billede af selskabets aktiver, passiver og finansielle stilling pr. 31. december 2014 samt af resultatet af selskabets aktiviteter for regnskabsåret 1. januar 2014 - 31. december 2014 i overensstemmelse med årsregnskabsloven.



# Dates and timestamps

```
print(realdate+12)
"2022-06-09"
```

| Symbol | Meaning | Example |
|---|---|---|
| %a | Abbreviated weekday name | Tue |
| %A | Full weekday name | Tuesday |
| %b | Abbreviated month name | Apr |
| %B | Full month name | April |
| %C | Century: the integer part of the year divided by 100 | 20 |
| %d | Day of the month | 09 |
| %H | Hours as decimal number (00–23) | 13 |
| %I | Hours as decimal number (01–12) | 1 |
| %m | Month as number (01–12) | 04 |
| %M | Minute as number (00–59) | 12 |
| %p | AM/PM indicator for 12-hour time (%I) | PM |
| %S | Second as integer (00–61) | 12 |
| %u | Weekday as a decimal number (1–7, Monday is 1) | 2 |
| %w | Weekday as decimal number (0–6, Sunday is 0) | 2 |
| %y | 2-Digit Year (00–99) | 19 |
| %Y | 4-Digit Year | 2019 |

```
realdate <- as.Date("09/May/2022:00:09:24",format='%d/%m%Y:%H%M%S')
class(realdate)
Sys.setlocale(category = "LC_ALL",locale ="en_GB.UTF-8")
```

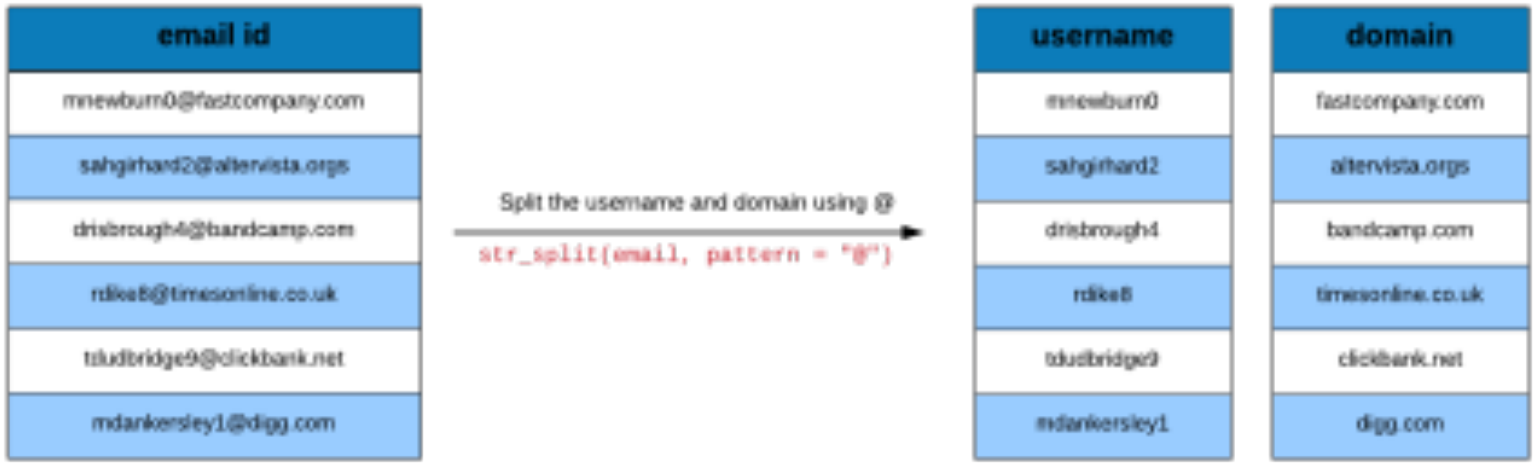
# String match

# Øvelse

| | IP | timestamp | words |
|---|---|---|---|
| 1 | *NA* | *NA* | *NA* |
| 2 | 93.162.98.150 | 09/May/2022:08:09:19 | Kurt+bor+her |
| 3 | 93.162.98.150 | 09/May/2022:08:13:47 | Burger+baren+er+%C… |
| 4 | 93.162.98.150 | 09/May/2022:08:13:59 | Burger+baren+er+%C… |
| 5 | 93.162.98.150 | 09/May/2022:08:15:03 | Burger+baren+er+%C… |
| 6 | 93.162.98.150 | 09/May/2022:08:16:15 | Pippi+langstr%C3%B8… |
| 7 | 93.162.98.150 | 09/May/2022:08:16:24 | Pippi+langstr%C3%B8… |
| 8 | 93.162.98.150 | 09/May/2022:08:18:27 | Pippi+langstr%C3%B8… |
| 9 | 93.162.98.150 | 09/May/2022:08:18:45 | Pippi+langstr%C3%B8… |
| 10 | 93.162.98.150 | 09/May/2022:08:19:04 | Pippi+langstr%C3%B8… |
| 11 | 93.162.98.150 | 09/May/2022:08:20:46 | Vlggo |
| 12 | 93.162.98.150 | 09/May/2022:08:21:22 | Vlggo |
| 13 | 93.162.98.150 | 09/May/2022:08:21:33 | S%C3%B8gren |
| 14 | 93.162.98.150 | 09/May/2022:08:25:03 | sdf+asdf+asdf+sesf |
| 15 | 93.162.98.150 | 09/May/2022:08:28:54 | Snurre+Snup |
| 16 | 93.162.98.150 | 09/May/2022:08:32:02 | Viggo+Kampnam |
| 17 | 93.162.98.150 | 09/May/2022:08:38:21 | Hammersh%C3%B8i+b… |

1) Lav denne dataframe ud fra access-loggen
2) Tilføj OS og Browser

# Split string



# String concatenate

# String subtract

| amount |
| --- |
| $67.37 |
| ¥34.37 |
| €33,85 |
| £22.80 |
| $77.29 |
| £51.10 |

Extract the currency from the amount

```
str_sub(amount, start = 1, end = 1)
```

| currency |
| --- |
| $ |
| ¥ |
| € |
| £ |
| $ |
| £ |

# String search and replace

| address |
| --- |
| 1 Elgar Road |
| 7814 Pennsylvania Street |
| 8 Manley Drive |
| 659 7th Avenue |
| 69 Gina Street |
| 9 Hoepker Street |

Replace Street with ST

```
str_replace(address, "Street", "ST)
```

| address |
| --- |
| 1 Elgar Road |
| 7814 Pennsylvania ST |
| 8 Manley Drive |
| 659 7th Avenue |
| 69 Gina ST |
| 9 Hoepker ST |

# String detect type

| email id |
| --- |
| mnewburn0@fastcompany.com |
| sahgirhard2@altervista.orgs |
| drisbrough4#bandcamp.com |
| rdike8@timesonline.co.uk |
| tdudbridge9$clickbank.net |
| mdankersley1@digg.com |

Detect the pattern @ in the email id

```
str_detect(pattern = "@")
```

| Detect @ |
| --- |
| TRUE |
| TRUE |
| FALSE |
| TRUE |
| FALSE |
| TRUE |

```
22 #if GET then find queries
23 if ( str_detect(ntst, "GET")) {
24   print("GOT IT")
25 }
26
```