REGULÆRE UTTRYKK I PYTHON



## Hvordan få regulære uttrykk inn i python

# Hent den koden vi trenger fra

For å gjøre et søk etter noe i en tekst, så må vi bruke de følgende to linjene.

```
# python sin egen kodebase.
import re

# Et funksjonskall for
# sok etter det vi vil ha i en tekst,
# og leg resultatet i en variabel.
resultat = re.search(det_vi_vil_ha, tekst)
```



Du kan søke etter hvilken som helst tekststreng. import re

```
# en tekst variabel som forteller
# oss hva vi leter etter.
det_vi_leter_etter = "hund"

# En tekst vi vil lete i.
tekst = "det_var_en_gang_en_hund_som_het_Arne."

# Let etter svar
resultat = re.search(det_vi_leter_etter, tekst)
```

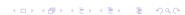




```
Og vi kan finne ut om vi har funnet noe ved å si:
```

```
# Hvis vi har funnet noe
if resultat:
    # Sier vi at vi har funnet noe
    print det_vi_leter_etter," _finnes_i_teksten."

# Hvis vi ikke har funnet noe
else:
    # Da sier vi at vi ikke har funnet noe
    print det_vi_leter_etter," _finnes_ikke_i_teksten."
```



- ▶ Bortsett fra å kunne søke etter noe i en tekst, hva er spessiellt med regulære uttrykk?
- ▶ Spessielle symboler i teksten vil gjøre søket mer avansert.



```
Symbolet . ( \setminus \setminus . )
```

. kan være hva som helst. Hvis du ville ha en dot (for eksempel søke etter filer som "hei.txt"), må du bruke \ \ . import re det\_vi\_leter\_etter = "h..d" # Her er det vi leter etter tekst = "det\_var\_en\_gang\_en\_hund\_som\_het\_Arne." resultat = re.search(det\_vi\_leter\_etter, tekst) if resultat: print det\_vi\_leter\_etter ,"\_finnes\_i\_teksten." else: print det\_vi\_leter\_etter ," \_finnes \_ikke \_i \_teksten ."

## UiO **Campus TG**



## Symbolet \*

\* vil finne alle symbol som er rett foran den 0 eller flere ganger. Eksempel: "hu\*nd" vil matche

- ► hnd
- hund
- huund
- huuund
- etc.



## Paranteser ()

Med () kan du lage større grupper av symboler og få ut hva du har funnet. Eksempel: "h(..)d" vil matche hund og i resultatet fra søket kan du finne ut at .. var "un"



Med dette kan vi begynne å crawle Webben for info. For eksempel, hvordan vil du prøve å finne tittelen på dette dokument?



Vi vet at tittelen må starte med "< title >" og slutte med "< /title >". Hva som er inni der er uvisst/ Så om vi ikke vet hva som er der, så kan vi bruke ".", som representerer hvilken som helst bokstav. Men vi trenger den mer enn 1 gang. Hvordan løser vi det problemet? "\*" vil ta et symbol mer enn 1 gang. ".\*" vil da ta hva som helst. Så da prøver vi "< title >.\*< /title >" Da vil du kunne finne ut om et dokument har en tittel, men for å hente det ut, må vi bruke (). "< title >(.\*) < /title >"

