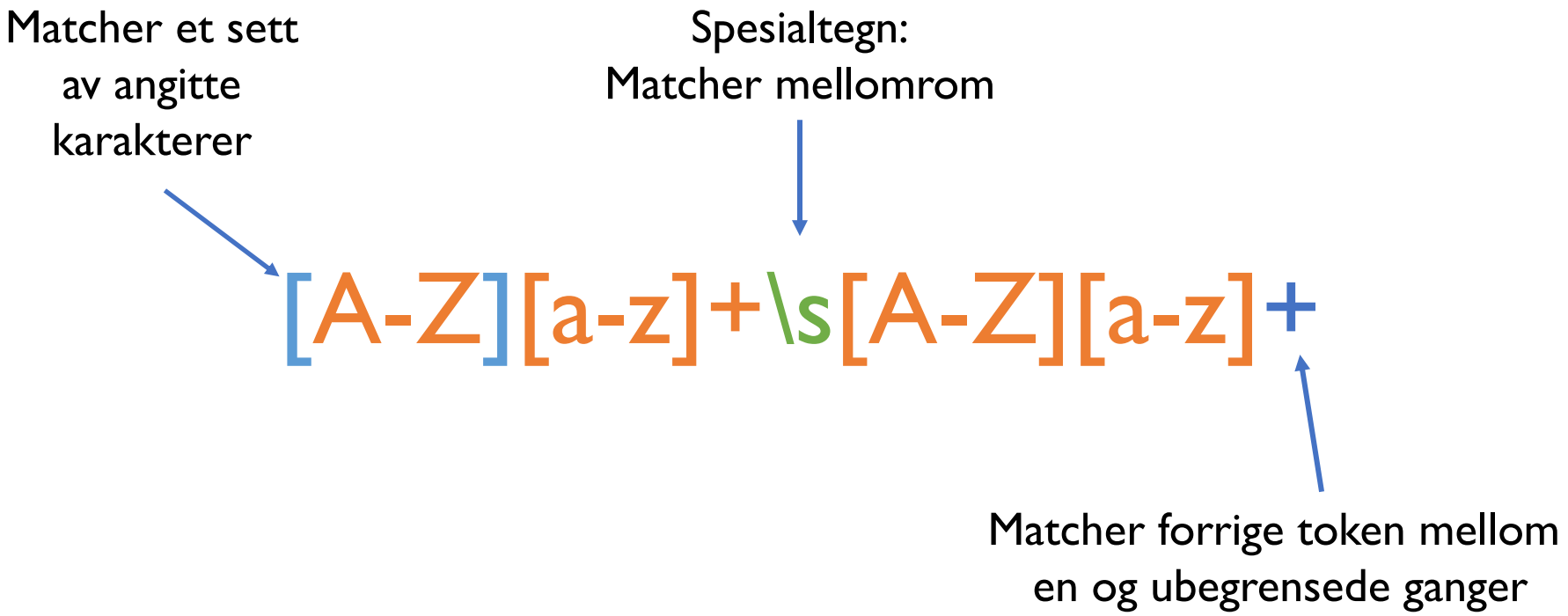
REGULÆRE UTTRYKK

En sekvens av karakterer som danner et søkbart mønster. Dette kan brukes til å ekstrahere ut det man ønsker i en tekst.


Matcher et sett av angitte karakterer
Spesialtegn: Matcher mellomrom
[A-Z][a-z]+\s[A-Z][a-z]+
Matcher forrige token mellom en og ubegrensede ganger

QUIZ!

[A-Z][a-z]+\s[A-Z][a-z]+

New York ✅
Ola Nordmann ✅
Jonas Gahr Støre ❌

new york ❌
Norge ❌
Frogner stadion ❌

SETT VS. GRUPPE

[hallo!]

«hallo! Hvordan går det?»

[«h» «a» «l» «l» «o» «!» «h» «o» «a»]

(hallo!)

«hallo! Hvordan går det?»

[«hallo!»]

SPESIALTEGN

[aA-zZ0-9_]+@[aA-zZ]+\.[aA-zZ]{2}\b

liljacs@uio.no ✅
liljacs@ifi.uio.no ❌

Eller ..

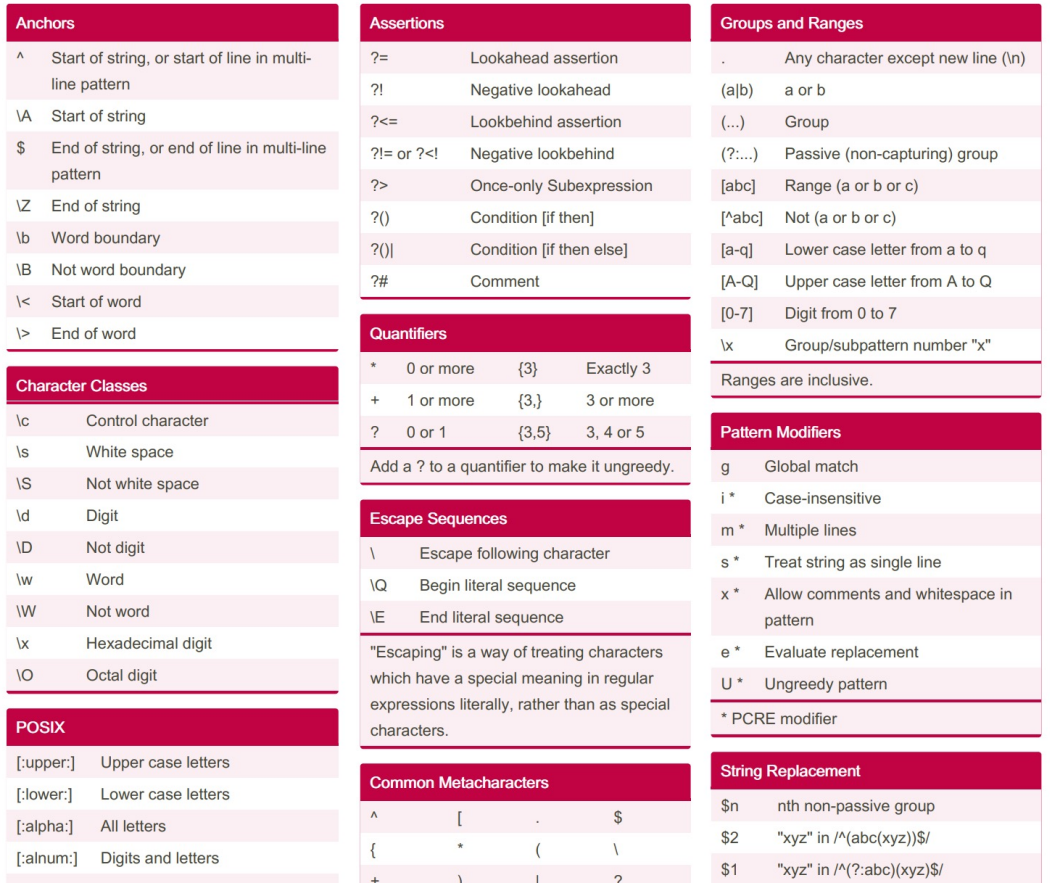
\w+@\w+\.\w{2}\b

\w matcher alle bokstaver inkludert æ, ø, å, é, ä, ñ, o.l. (ikke case sensitive), tall mellom 0-9, og '_'

MYE Å HUSKE?

https://www.uio.no/studier/emner/matnat/ifi/IN1140/h22/Undervisningsmateriale/gruppe-3/regulere-uttrykk.pdf

Eller google «regex cheat sheet» →



Anchors

^	Start of string, or start of line in multi-line pattern
\A	Start of string
$	End of string, or end of line in multi-line pattern
\Z	End of string
\b	Word boundary
\B	Not word boundary
\<	Start of word
\>	End of word

Character Classes

\c	Control character
\s	White space
\S	Not white space
\d	Digit
\D	Not digit
\w	Word
\W	Not word
\x	Hexadecimal digit
\O	Octal digit

POSIX

[:upper:]	Upper case letters
[:lower:]	Lower case letters
[:alpha:]	All letters
[:alnum:]	Digits and letters

Assertions

?=	Lookahead assertion
?!	Negative lookahead
?<=	Lookbehind assertion
?!= or ?<!	Negative lookbehind
?>	Once-only Subexpression
?()	Condition [if then]
?()|	Condition [if then else]
?#	Comment

Quantifiers

*	0 or more	{3}	Exactly 3
+	1 or more	{3,}	3 or more
?	0 or 1	{3,5}	3, 4 or 5

Add a ? to a quantifier to make it ungreedy.

Escape Sequences

\	Escape following character
\Q	Begin literal sequence
\E	End literal sequence

"Escaping" is a way of treating characters which have a special meaning in regular expressions literally, rather than as special characters.

Common Metacharacters

^	[	.	$
{	*	(	\
+	)	|	?

Groups and Ranges

.	Any character except new line (\n)
(a|b)	a or b
(...)	Group
(?:...)	Passive (non-capturing) group
[abc]	Range (a or b or c)
[^abc]	Not (a or b or c)
[a-q]	Lower case letter from a to q
[A-Q]	Upper case letter from A to Q
[0-7]	Digit from 0 to 7
\x	Group/subpattern number "x"

Ranges are inclusive.

Pattern Modifiers

g	Global match
i *	Case-insensitive
m *	Multiple lines
s *	Treat string as single line
x *	Allow comments and whitespace in pattern
e *	Evaluate replacement
U *	Ungreedy pattern

* PCRE modifier

String Replacement

$n	nth non-passive group
$2	"xyz" in /^(abc(xyz))$/
$1	"xyz" in /^(?:abc)(xyz)$/

REGULÆRE UTTRYKK I PYTHON

```
import re

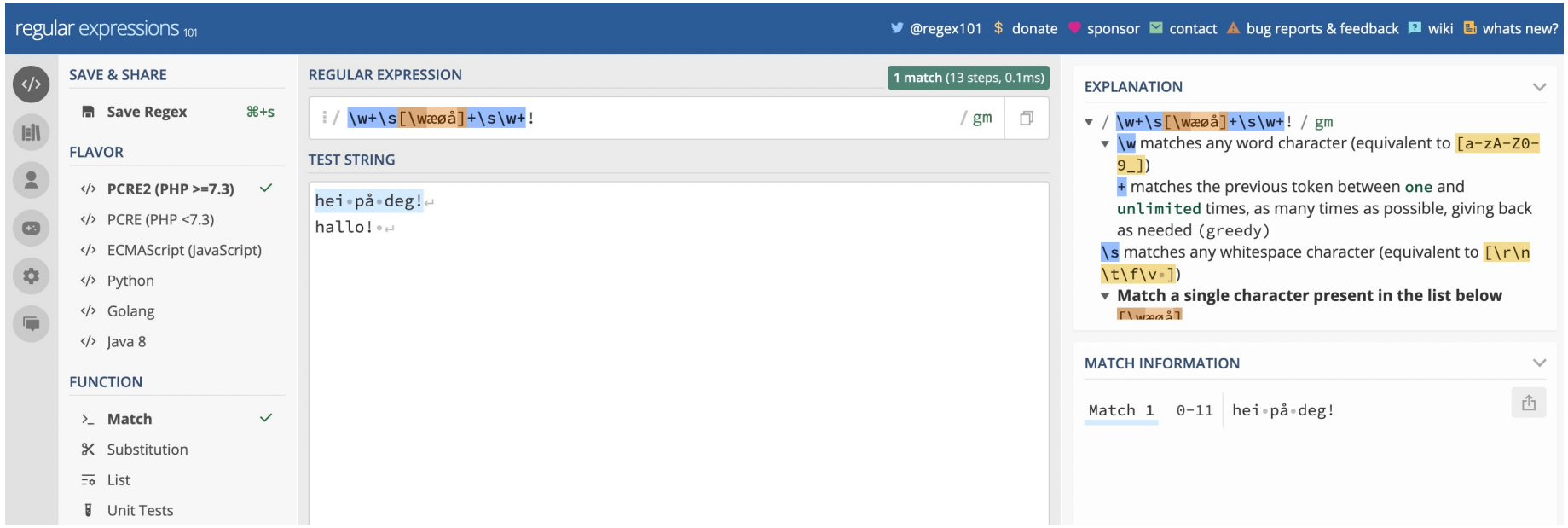
tekst = "Jeg er på tur i New York"
uttrykk = r"[A-Z][a-z]+\s[A-Z][a-z]+"
tokens = re.findall(uttrykk, tekst)

print(tokens)
```

>>> [«New York»]

REGEX101

Et genialt verktøy!🤩



regex101.com