Agenda

- Regex
- Simple pattern Matching
- Metacharacters
- Anchors
- Quantifiers and Groups
- Python implementation of Regex with re library

Regex

Regex can help in identying complex patterns in Documents

- Data Cleaning
- Data Validation
- Masking Information

8586836627 ⇒ **** ** 6627

→ Case-sensitive 'abc' != 'ABC'

→ Order matters 'abc' != 'cba'

'●'

→ Special character that matches Everything Except newline.

\ → Escape character / backslash

'\.' → metachar

Meta Characters

\d : Matches all digits

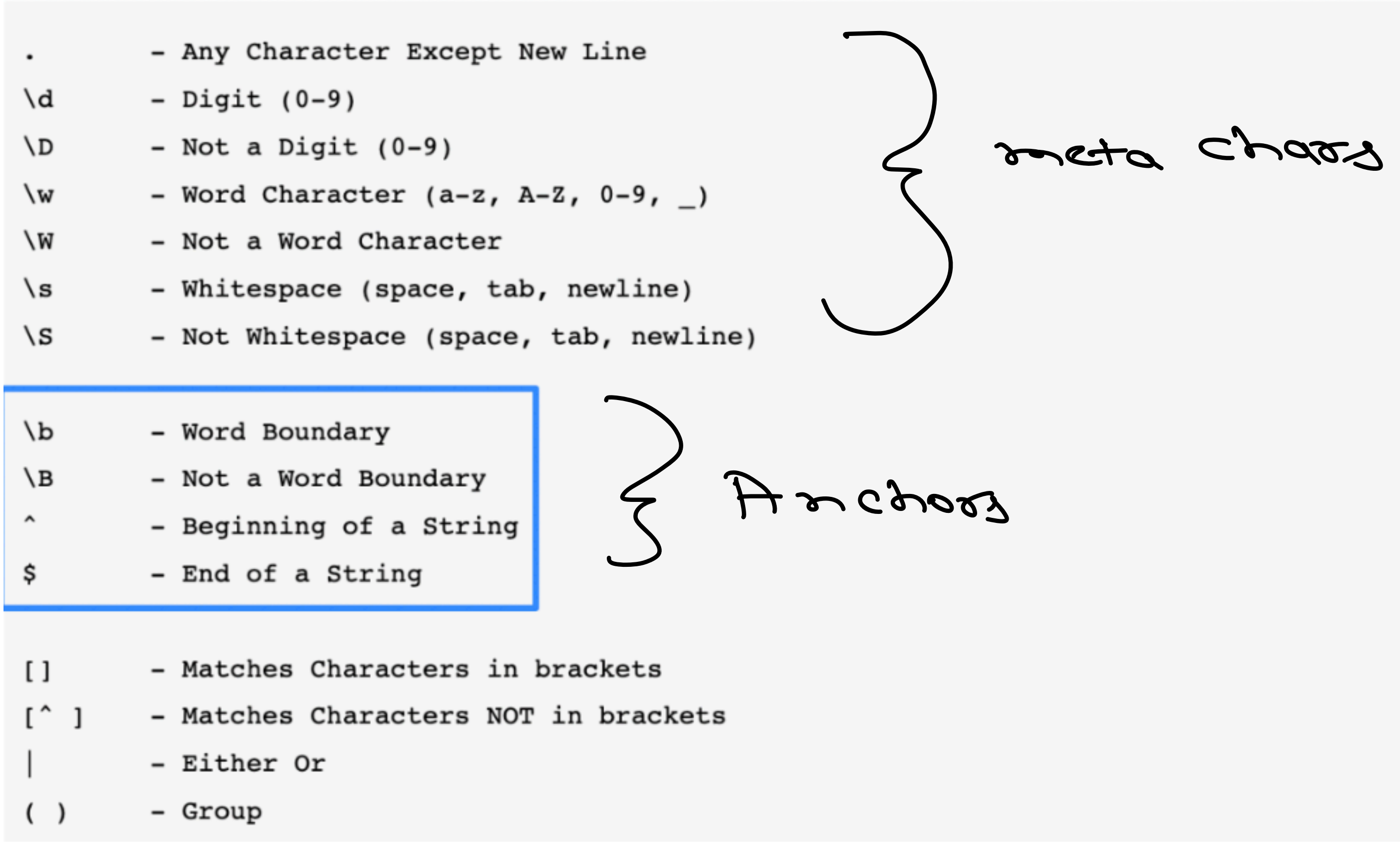
\D : ! Matches all digits

\w : Matches Alphanumeric and _

\W : ! \w

\s : matches whitespace chars

\S : ! \s



```
.       - Any Character Except New Line
\d      - Digit (0-9)
\D      - Not a Digit (0-9)
\w      - Word Character (a-z, A-Z, 0-9, _)
\W      - Not a Word Character
\s      - Whitespace (space, tab, newline)
\S      - Not Whitespace (space, tab, newline)

\b      - Word Boundary
\B      - Not a Word Boundary
^       - Beginning of a String
$       - End of a String

[]      - Matches Characters in brackets
[^ ]    - Matches Characters NOT in brackets
|       - Either Or
( )     - Group
```

} meta chars

} Anchors

Anchors

- Anchors Don't match any characters
- They match invisible positions before or after characters

1. \b : boundary of a word
   - Prefix

2. \B : suffix

3. ^ → matches Beginning of String

4. $ → Matches end of String

Pattern$

Character Set

[.-] => matches either . or -

[19] ✗ [1 2 3 4 5 6 7 8 9] ✓

↓

[1-9]

Ex

[ab]

[.-]

* numeric character set

=> [1 2 3 4 5 6]

=> [1-6] => [start - end]

=> [1-9]

* Alphabet Charset

[a-z] [A-Z] [a-zA-Z]

* Negation in character-set

[^a-z] ⇒ match everything but lowercase char

* Note: inside a character Set ^ works as negation of character Set

Quantifiers

⊙ match specific repetion of characters

① * ⇒ 0 or more

② + ⇒ 1 or more

③ ? ⇒ 0 or One

④ {3} ⇒ Exactly 3

{3, 4} ⇒ {Min 3, max-4}

{3,} → 3 or more

\d\d\d → \d{3}

Look for group of 3 digits

a*
→ a
→ aa
→ aaa
→ ——

a?
→
→ a

Mr\.? → Mr
Mr.

Group

(   )

M(r|s|rs)\.?[a-zA-Z]*

r ✓

or

s ✓

or

rs ✓

(Mr|Ms|Mrs)