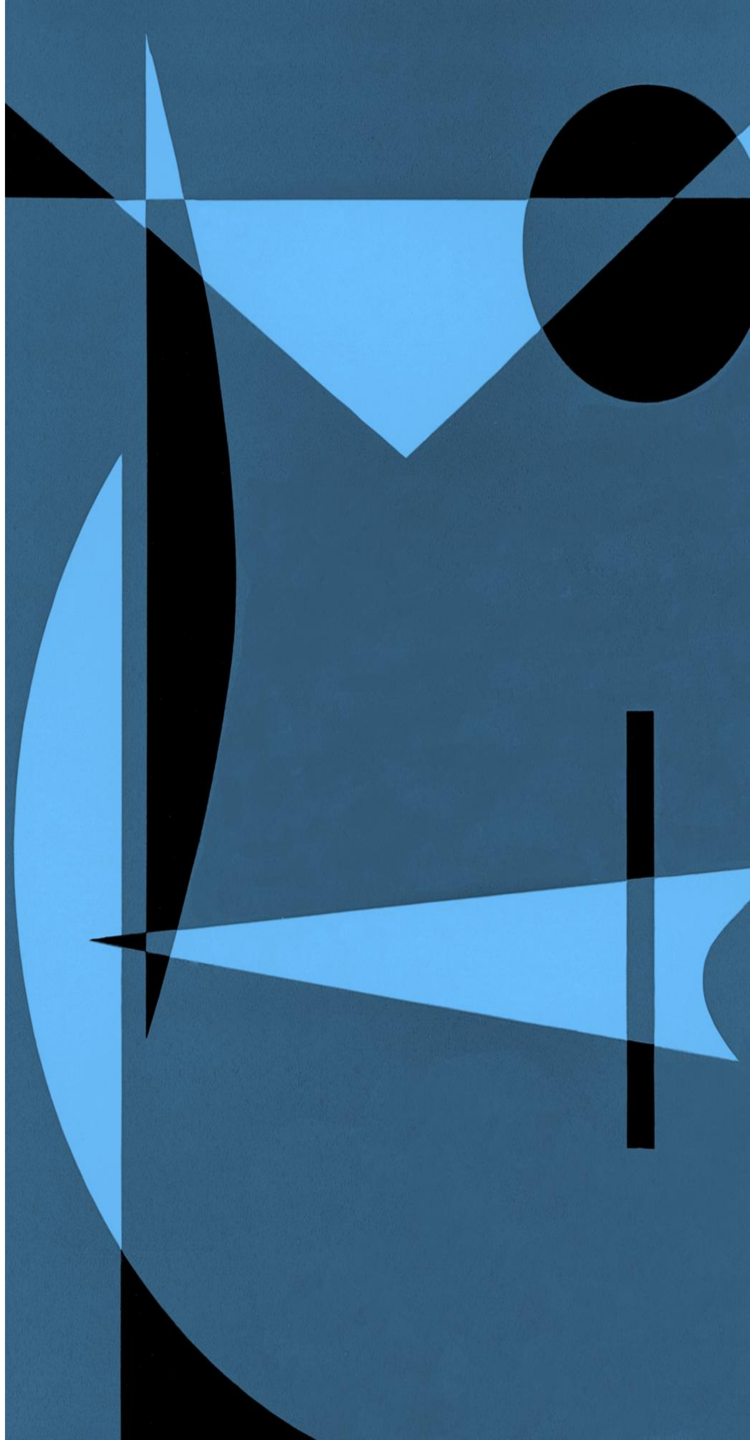
Regular expressions

Natural Language Processing ’24-’25
Tutorial 1

Yoanna Koleva



Contents

- What are regular expressions?
- Simplest form of regular expressions
- Metacharacters (special symbols)
- Sets
- Useful functions

Regular expressions

- Used to find common patterns in strings
- Useful to extract information in big bodies of text
- Useful for preprocessing

Example:

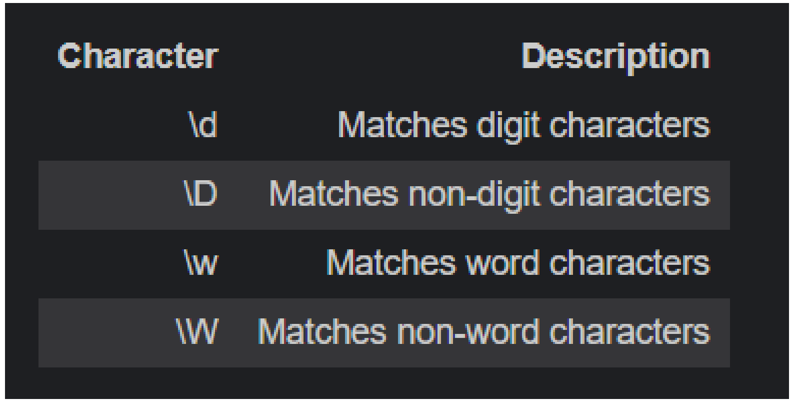
“ain” is a common pattern in: “The train travels through the mountain in the rain.”

Simplest form of regular expressions

- Made out of ordinary characters => 'a', 'B', 'w', etc.
- Can be concatenated
- “string” matches “string”; “ing” matches “ing” in “string”
- But this is not very informative or useful

Metacharacters

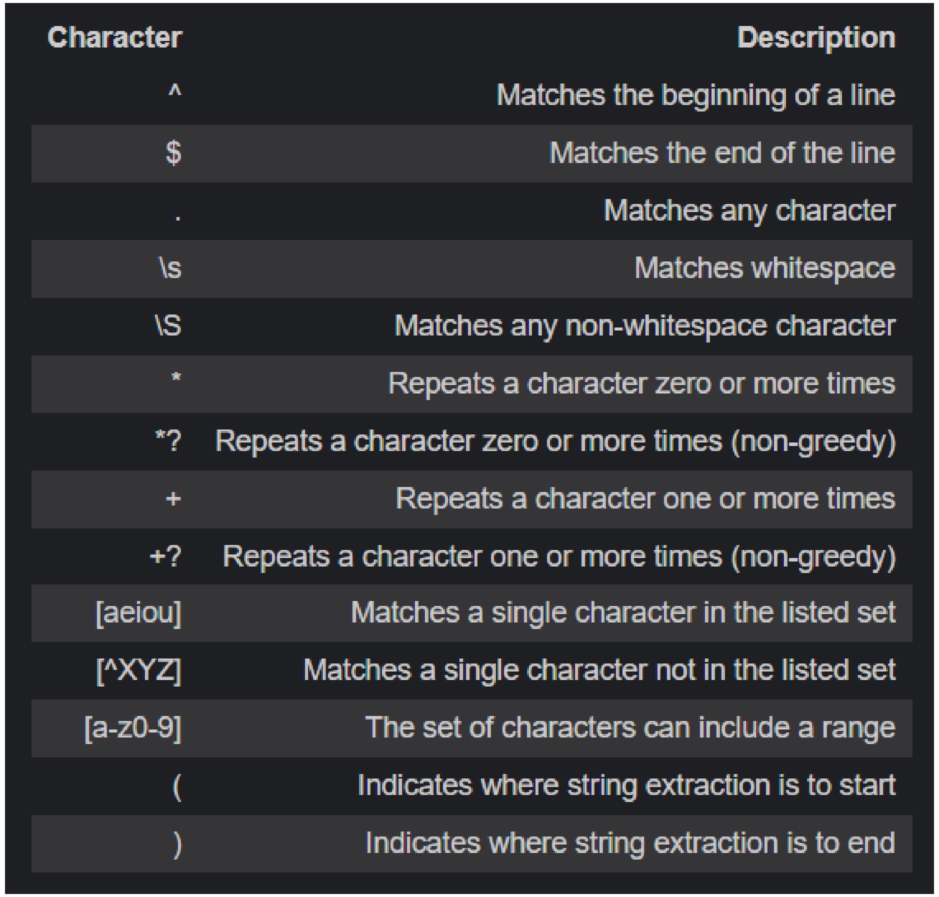
- Have special meaning
- Example: "\", "()", "|", "*", ".“, etc
- “\” used to signify special sequences



Character	Description
\d	Matches digit characters
\D	Matches non-digit characters
\w	Matches word characters
\W	Matches non-word characters

- Problem: Python uses “\” for escape characters: “\n”, “\t”, etc.
- Solution: define strings like this r”string”

Metacharacters



Character	Description
^	Matches the beginning of a line
$	Matches the end of the line
.	Matches any character
\s	Matches whitespace
\S	Matches any non-whitespace character
*	Repeats a character zero or more times
*?	Repeats a character zero or more times (non-greedy)
+	Repeats a character one or more times
+?	Repeats a character one or more times (non-greedy)
[aeiou]	Matches a single character in the listed set
[^XYZ]	Matches a single character not in the listed set
[a-z0-9]	The set of characters can include a range
(	Indicates where string extraction is to start
)	Indicates where string extraction is to end

Sets

- Defined with []
- Can be used with greedy operators : i.e: [abc]+
- Can also be used for ranges of characters: [a-m]
- NOTE: Inside sets special characters (except “\”) lose their meaning!
- Can also be used for digits: i.e: [0-5][0-9]
- A “^” in the beginning of a set-> indicates negation

Useful functions and properties

- **re.search()**
  - As soon as it finds a match it stops searching
  - Returns an object
  - Can be used with **match.start()** and **match.group()**
- **re.findall()**
  - Returns a list of all matches
- **re.sub()**
- **string.group()**
  - Used to extract previously defined In the RE “groups”

Resources

https://www.w3schools.com/python/python_regex.asp#split

https://docs.python.org/3/library/re.html

https://developers.google.com/edu/python/regular-expressions