Python Regular Expressions

re.compile – allows us to separate our patterns with a variable and enable us to perform multiple searches.

sentence = 'Start a sentence and then bring it to an end'

pattern = re.compile(r'start', re.I)

matches = pattern.search(sentence)

print(matches)

. - Any Character Except New Line

\d - Digit (0-9)

\D - Not a Digit

\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

Quantifiers:

\* - 0 or More

+ - 1 or More

? - 0 or One

{3} - Exact Number

{3, 4} - Range of Numbers (Minimum, Maximum)

#### Sample Regexs ####

[a-zA-Z0-9\_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+

import re

text\_to\_search = '''

abcdefghijklmnopqurtuvwxyz

ABCDEFGHIJKLMNOPQRSTUVWXYZ

1234567890

Ha HaHa

MetaCharacters (Need to be escaped):

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

coreyms.com

321-555-4321

123.555.1234

123\*555\*1234

800-555-1234

900-555-1234

Mr. Schafer

Mr Smith

Ms Davis

Mrs. Robinson

Mr. T

'''

sentence = 'Start a sentence and then bring it to an end'

pattern = re.compile(r'start', re.I)

matches = pattern.search(sentence)

print(matches)

import re

emails = '''

CoreyMSchafer@gmail.com

corey.schafer@university.edu

corey-321-schafer@my-work.net

'''

pattern = re.compile(r'[a-zA-Z0-9\_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+')

matches = pattern.finditer(emails)

for match in matches:

print(match)

# <re.Match object; span=(1, 24), match='CoreyMSchafer@gmail.com'>

# <re.Match object; span=(25, 53), match='corey.schafer@university.edu'>

# <re.Match object; span=(54, 83), match='corey-321-schafer@my-work.net'>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

import re

urls = '''

https://www.google.com

http://coreyms.com

https://youtube.com

https://www.nasa.gov

'''

pattern = re.compile(r'https?://(www\.)?(\w+)(\.\w+)')

subbed\_urls = pattern.sub(r'\2\3', urls)

print(subbed\_urls)

# google.com

# coreyms.com

# youtube.com

# nasa.gov

matches = pattern.finditer(urls)

for match in matches:

print(match.group(3))

# google

# coreyms

# youtube

# nasa

Replace the domain name and high level domain name with 2nd and 3rd groups.

