INFO-664-01
Programming For Cultural Heritage

Regular Expressions

Regular Expressions

- To find things in Text in python we can do something like:

```
my_text = "Hello there!"

#start now equals 6
start = my_text.find("there")

#tell it to return the text that starts at start and ends at the length of the text
sub_text = my_text[start:len(my_text)]
```

Regular Expressions

- Matching text is difficult. Regular Expressions are used for pattern matching text.

- You'll find them in all languages and many tools (text editors, open refine, etc)

Regular Expressions

- You can build a pattern from characters such as: “[0-9]”

- And there are reserved tokens that do something specific such as: “\s”

Regular Expressions

- You build up this pattern to return parts of the text.
- Do ranges with brackets []
- Make “groups” with parentheses ()
- There are reserved tokens (for example “*” so if you want to match them escape them with a back slash “*”
- Demo:

William Wegman

COTTO
1970

William Wegman, *Cotto*, 1970. Gelatin silver print, 7 7/8 × 7 3/4 in. (20 × 19.7 cm). Whitney Museum of American Art, New York; purchase with funds from the Mrs. Percy Uris Purchase Fund and the Photography Committee 92.14

William Wegman, *Cotto*, 1970. Gelatin silver print, 7 7/8 × 7 3/4 in. (20 × 19.7 cm). Whitney Museum of American Art, New York; purchase with funds from the Mrs. Percy Uris Purchase Fund and the Photography Committee 92.14

Regular Expressions

- Let's figure out what the expression first will be and then implement it in python:

  - http://pythex.org/

William Wegman, Cotto, 1970. Gelatin silver print, 7 7/8 × 7 3/4 in. (20 × 19.7 cm). Whitney Museum of American Art, New York; purchase with funds from the Mrs. Percy Uris Purchase Fund and the Photography Committee 92.14
© William Wegman

regex.py

```python
import re

our_text = "William Wegman, Cotto, 1970. Gelatin silver print, 1999 7

#compile our pattern
p = re.compile('([0-9]{4})')

#Findall for it
m = p.findall(our_text)

print(m)
```

Line 14, Column 1 | Tab Size: 4 | Python

We build the regular expression and then use its .findall method to return a list of matches

regex.py

```python
import re

our_text = "William Wegman, Cotto, 1970. Gelatin silver print, 1999 7

#compile our pattern
p = re.compile('([0-9]{4})')

#search method
m = p.search(our_text)

#print out the first group
print(m.group(0))

#print the positions of it
print(m.span())

```

Line 17, Column 1 | Tab Size: 4 | Python

To find a exact bit of the text and its position we can use the .search method.

Regular Expressions

- Resources:

- https://docs.python.org/3.4/howto/regex.html

- Cheat sheet: http://www.cheatography.com/davechild/cheat-sheets/regular-expressions/

- http://www.diveintopython3.net/regular-expressions.html

Challenge

- Try to use a regular expression in you API results. Or try a regular expression using possible project data.