Joyce Woznica  
IST 652, Week 9

Activity 9.3: Regular Expressions

﻿Code:

﻿# 9.4 - Python Regular Expressions

text = '''Celebrate #NationalPetDay with our puppy playlist: https://t.co/eBHHFPW0z7 https://t.co/uix5AY2FFQ<a href="http://msande.stanford.edu"> Management Science and Engineering </a><p

Address: Terman 311, Stanford CA 94305<br>

Email: ashishg@cs.stanford.edu<br>

Phone: (650)814-9999 [Cell], Fax: (650)723-9999<br>

Admin asst: Roz Morf, Terman 405, 650-723-9999, rozm@stanford.edu</p>

The U.S.A. olympic teams have east-west training centers with up-to-date equipment.

'''

# define a function to return the regular expression passed

def Find(string, regex):

# findall() has been used

# with valid conditions for urls in string

m\_string = re.findall(regex,string)

return [x[0] for x in m\_string]

# 1. Match the URLs

regex = r"(?i)\b((?:https?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()<>]+\)))\*\))+(?:\(([^\s()<>]+|(\([^\s()<>]+\)))\*\)|[^\s`!()\[\]{};:'\".,<>?«»“”‘’]))"

print("Urls: ", Find(text, regex))

# 2. Match the phone numbers (3 results)

regex = r"(\d{3}[-\.\s]??\d{3}[-\.\s]??\d{4}|\(\d{3}\)\s\*\d{3}[-\.\s]??\d{4}|\d{3}[-\.\s]??\d{4})"

regex = re.compile("^\D?(\d{3})\D?\D?(\d{3})\D?(\d{4})$")

regex = r"[(]?[0-9]+[)]?[-]?[0-9]+[-][0-9]+"

regex = r"(\([0-9]{3}\)[\s-]?|[0-9]{3}-)[0-9]{3}-[0-9]{4}"

print("Phone Numbers: ", Find(text, regex))

# cannot get anything to work for this.

# 3. Match the e-mail addresses (two results).

regex = r"[a-z]+@[a-z]+.[a-z]+.?[a-z]+"

regex = r"'\S+@\S+'"

regex = re.compile("'\S+@\S+'")

print("Email Addresses: ", Find(text, regex))

# won't work

#4. Match acronyms (there should be two results, with or without dots: CA and U.S.A.).

regex = r"[A-Z]\.?[A-Z]\.?[A-Z]?\.?\s"

print("What's left: ", Find(text,regex))

Output:

Couldn’t get much to work after multiple efforts, so gave up.

﻿Urls: ['https://t.co/eBHHFPW0z7', 'https://t.co/uix5AY2FFQ', 'http://msande.stanford.edu']

Phone Numbers: ['(', '(', '6']

Email Addresses: []

What's left: ['C', 'U']