

Sathish_Kumar_Rajendiran_Week9.4_Finding_Patterns

August 27, 2020

Name: Sathish Kumar Rajendiran
Task: Week 9: 9.4 Patterns in Text
Date: 8/26/2020

```
[11]: #import libraries

# standard library
import os
import sys
from datetime import datetime
import time

# text tokenization
import nltk
import re

#MongoDB libraries
import pymongo
from pymongo import MongoClient

os.getcwd()
```

```
[11]: '/Users/sathishrajendiran/ist652-python'
```

```
[5]: # sample text
text= '''Celebrate #NationalPetDay with our puppy playlist: https://t.co/
↳eBHHPW0z7 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.'
```

```
[109]: #Emails
if re.search('[\w\.-]+@[\w\.-]+',text):
    text_url = re.findall(r'[\w\.-]+@[\w\.-]+',text)
    print('Found a match on emails:\n',text_url )
else:
    print('No match.')
```

```
Found a match on emails:
['ashishg@cs.stanford.edu', 'rozm@stanford.edu']
```

```
[174]: #Phone Numbers
if re.search('([]?[0-9]+[)]?[-]?[0-9]+[-]?[0-9]+',text):
    text_phone = re.findall(r'([]?[0-9]+[)]?[-]?[0-9]+[-]?[0-9]+',text)
    print('Found a match on Phone Numbers:\n',text_phone )
else:
    print('No match.')
```

```
Found a match on Phone Numbers:
['(650)814-9999', '(650)723-9999', '650-723-9999']
```

```
[175]: #Acronyms
if re.search('[A-Z]\.[A-Z]\.[A-Z]?\.?\s',text):
    text_phone = re.findall(r'[A-Z]\.[A-Z]\.[A-Z]?\.?\s',text)
    print('Found a match on Acronyms:\n',text_phone )
else:
    print('No match.')
```

```
Found a match on Acronyms:
['CA ', 'U.S.A. ']
```

```
[177]: #URLs
if re.search('https?:/(?:[-\w.]|(?:%[\da-fA-F]{2}))+',text):
    text_phone = re.findall(r'https?:/(?:[-\w.]|(?:%[\da-fA-F]{2}))+',text)
    print('Found a match on URLs:\n',text_phone )
else:
    print('No match.')
```

```
Found a match on URLs:
['https://t.co/eBHHPW0z7', 'https://t.co/uix5AY2FFQ',
'http://msande.stanford.edu']
```

```
[178]: # words with internal hiphens
if re.search('\w.\-',text):
    text_hiphens = re.findall(r'\w.\-',text)
    print('Found a match on internal hiphens:\n',text_hiphens )
```

```
else:  
    print('No match.')
```

Found a match on internal hiphens:

```
['14-', '23-', '50-', '23-', 'st-', 'up-', 'to-']
```