TextAnalytics_Homework1_Final

October 11, 2019

0.1 Testing regex to match emails and dates in the Newsgroup text corpus

In [1]:
```
import re
```

In [2]:
```
#### Read in the data to test regular expressions on
with open("20-newsgroups/talk.religion.misc.txt", encoding="utf8", errors='ignore') as
    test_text = file.read().replace('\n', '')
```

0.2 Emails

In [6]:
```
#### Regular expressions that match the email
matched_emails = re.findall(r'\b[A-Za-z0-9_.%+-]+@[A-Za-z0-9.-]+\.[A-Za-z]{1,}\b',test_
matched_emails[0:25]
```

Out[6]:
```
['dsoconne@quads.uchicago.edu',
 'magarret@magnus.acs.ohio-state.edu',
 'dsoconne@uchicago.edu',
 'pharvey@quack.kfu.com',
 'C5OKDr.Duz@acsu.buffalo.edu',
 'psyrobtw@ubvmsb.cc.buffalo.edu',
 'dsav+@andrew.cmu.edu',
 'joslin@pogo.isp.pitt.edu',
 'Tyreaf664@yfn.ysu.edu',
 'boylan@sltg04.ljo.dec.com',
 '1993Apr1.024850.20111@sradzy.uucp',
 'radzy@sradzy.uucp',
 'swaim@owlnet.rice.edu',
 'radzy@sradzy.uucp',
 'boylan@sltg04.ljo.dec.com',
 'radzy%sradzy@jack.sns.com',
 'pharvey@quack.kfu.com',
 'martini@ccwf.cc.utexas.edu',
 'mayne@pipe.cs.fsu.edu',
 '7912@blue.cis.pitt.edu',
 'joslin@pogo.isp.pitt.edu',
 'af664@yfn.ysu.edu',
 'Thyagi@cup.portal.com',
 'Thanks.Thyagi@HouseofKaos.Abyss.comNewsgroup',
 'lovall@bohr.physics.purdue.edu']
```

In [7]:
```
#### Number of emails matched in the given text
len(matched_emails)
```

Out[7]: 4442

In [8]:
```
#### Number of UNIQUE emails matched in the given text
len(set(matched_emails))
```

Out[8]: 841

In [10]:
```
#### An alternative regular expression that was tried but finally given up since it m
matched_emails_alt = re.findall(r'[\w\.-]+@[\w\.-]+',test_text)
matched_emails_alt[0:25]
```

Out[10]:
```
['dsoconne@quads.uchicago.edu',
 'magarret@magnus.acs.ohio-state.edu',
 'dsoconne@uchicago.edu',
 'pharvey@quack.kfu.com',
 'C50KDr.Duz@acsu.buffalo.edu',
 'psyrobtw@ubvmsb.cc.buffalo.edu',
 'joslin@pogo.isp.pitt.edu',
 'Tyreaf664@yfn.ysu.edu',
 'boylan@sltg04.ljo.dec.com',
 '1993Apr1.024850.20111@sradzy.uucp',
 'radzy@sradzy.uucp',
 'swaim@owlnet.rice.edu',
 'radzy@sradzy.uucp',
 'boylan@sltg04.ljo.dec.com',
 'sradzy@jack.sns.com',
 'pharvey@quack.kfu.com',
 'martini@ccwf.cc.utexas.edu',
 'mayne@pipe.cs.fsu.edu',
 '7912@blue.cis.pitt.edu',
 'joslin@pogo.isp.pitt.edu',
 'af664@yfn.ysu.edu',
 'Thyagi@cup.portal.com',
 'Thanks.Thyagi@HouseofKaos.Abyss.comNewsgroup',
 'lovall@bohr.physics.purdue.edu',
 'zxmkr08.733955549@studserv']
```

In [11]:
```
#### Number of emails matched in the given text
len(matched_emails_alt)
```

Out[11]: 4456

In [12]:
```
#### Number of emails matched in the given text
len(set(matched_emails_alt))
```

Out[12]: 858

```
In [13]: #### Find out emails that are matched by one regular expression and not the other

In [14]: (set([x for x in matched_emails_alt if x not in matched_emails]))

Out[14]: {'------------------------------------------------------------------------cutter@gloster
          '------------------popec@brewich.hou.tx.us',
          '...@compuserve.comI',
          '...@compuserve.comTonyNewsgroup',
          'A54SI@CUNYVM',
          'ISSCCK@BYUVM',
          'Kent---sandvik@newton.apple.com.',
          'Lanphierlanphi872@snake.cs.uidaho.edulanph872',
          'Lippard@CCIT.ARIZONA.EDUDept.',
          'MARGOLI@YKTVMV',
          'Utidjian-utidjian@remarque.berkeley.edu-Newsgroup',
          'anyway.darinwilkins@scubed.com_____________________________',
          'bd@fluent',
          'boo@PacBell.COM',
          'dk@imager',
          'dlphknob.734986640@camelot',
          'got.Jim--jmd@handheld.com-----------------------------------------------------------
          'humanist...Kent---sandvik@newton.apple.com.',
          'ismarkp@avignon',
          'keng.735334134@tunfaire',
          'like.Kent---sandvik@newton.apple.com.',
          'lynn@cs.cmu.edu',
          'lynn@cs.cmu.eduNewsgroup',
          'mathew@mant',
          'medkeffjs@hirambPO',
          'merlyn.735422443@digibd',
          'mst4298@zeus._____',
          'psyrobtw@ubvms.cc.buffalo.edu--',
          'psyrobtw@ubvms.cc.buffalo.edu--Rick',
          'radian@natinst.com',
          'sail.LABS.TEK.COM@RELAY.CS.NET',
          'sandvik@newton.apple.com.',
          'scharle@irishmvsRoom',
          'sradzy@jack.sns.com',
          'them.Kent---sandvik@newton.apple.com.',
          'tph@drake_mallard.sbc.com',
          'wcscps.735321331@cunews',
          'zxmkr08.733955549@studserv'}

In [15]: set([x for x in matched_emails if x not in matched_emails_alt])

Out[15]: {'.edulanph872@uidaho.eduAnd',
          'Kent---sandvik@newton.apple.com',
          'Lanphierlanphi872@snake.cs.uidaho',
          'Lippard@CCIT.ARIZONA.EDUDept',
```

```
'Utidjian-utidjian@remarque.berkeley.edu',
'alizard%tweekco%boo@PacBell.COM',
'dba+lynn@cs.cmu.edu',
'dba+lynn@cs.cmu.eduNewsgroup',
'drake+@cs.cmu.edu',
'dsav+@andrew.cmu.edu',
'fluent@dartmouth.EDU',
'got.Jim--jmd@handheld.com',
'humanist...Kent---sandvik@newton.apple.com',
'like.Kent---sandvik@newton.apple.com',
'markbr%radian@natinst.com',
'mikec%sail.LABS.TEK.COM@RELAY.CS.NET',
'nm0w+@andrew.cmu.edu',
'psyrobtw@ubvms.cc.buffalo.edu',
'radzy%sradzy@jack.sns.com',
'rjl+@pitt.edu',
'them.Kent---sandvik@newton.apple.com'}
```

0.3 Dates

```
In [23]: #### String of dates to test regular expression
         dates = '25.04.2017 , 02.04.2017 , 2.4.2017 , 25/04/2017 , 5/12/2017 , 15/2/2017 , 25-

In [26]: print(re.findall(r'([0-3][0-9][-\/.\\\s][0-3][0-9][-\/.\\\s](?:[0-9]{4}|[0-9]{2}))|((

[('25.04.2017', '', '', ''), ('02.04.2017', '', '', ''), ('25/04/2017', '', '', ''), ('25-04-20


In [ ]: #### Testing the regex on the 'test_text' data read in earlier

In [27]: print(re.findall(r'([0-3][0-9][-\/.\\\s][0-3][0-9][-\/.\\\s](?:[0-9]{4}|[0-9]{2}))|((

[('', '', '5 Apr 93', ''), ('', '', '36 Jan 01', ''), ('', '', '48 Mar 31', ''), ('', '', '12 S


In [ ]:
```