Homework1

October 11, 2019

In [7]:
```
import nltk
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
import spacy
import os
import tarfile
import re
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\tonyc\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\tonyc\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

In [2]:
```
## Open .tar.gz file if not already open
# file = tarfile.open('20news-19997.tar.gz',mode='r:gz')
# file.extractall()
# file.close()
os.listdir('20_newsgroups/')
```

Out[2]:
```
['alt.atheism',
 'comp.graphics',
 'comp.os.ms-windows.misc',
 'comp.sys.ibm.pc.hardware',
 'comp.sys.mac.hardware',
 'comp.windows.x',
 'misc.forsale',
 'rec.autos',
 'rec.motorcycles',
 'rec.sport.baseball',
 'rec.sport.hockey',
 'sci.crypt',
 'sci.electronics',
 'sci.med',
 'sci.space',
```

```
         'soc.religion.christian',
         'talk.politics.guns',
         'talk.politics.mideast',
         'talk.politics.misc',
         'talk.religion.misc']
```

```
In [3]: corpus_list = []
        for dir_ext in os.listdir('20_newsgroups/'):
            dir_name = '20_newsgroups/{}/'.format(dir_ext)
            for file_ext in os.listdir(dir_name):
                file_name = '{}{}'.format(dir_name,file_ext)
                file = open(file_name, 'r')
                text_list = file.readlines()
                text = ''.join(text_list)
                corpus_list.append(text)
        #         corpus = corpus + text
        #         print(text)
        #         if corpus != '':
        #             break
        corpus = ''.join(corpus_list)
```

```
In [4]: print(corpus[0:2000])
```

```
Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:49960 alt.atheism.moderated:713 news.answers:7054 a
Path: cantaloupe.srv.cs.cmu.edu!crabapple.srv.cs.cmu.edu!bb3.andrew.cmu.edu!news.sei.cmu.edu!c
From: mathew <mathew@mantis.co.uk>
Newsgroups: alt.atheism,alt.atheism.moderated,news.answers,alt.answers
Subject: Alt.Atheism FAQ: Atheist Resources
Summary: Books, addresses, music -- anything related to atheism
Keywords: FAQ, atheism, books, music, fiction, addresses, contacts
Message-ID: <19930329115719@mantis.co.uk>
Date: Mon, 29 Mar 1993 11:57:19 GMT
Expires: Thu, 29 Apr 1993 11:57:19 GMT
Followup-To: alt.atheism
Distribution: world
Organization: Mantis Consultants, Cambridge. UK.
Approved: news-answers-request@mit.edu
Supersedes: <19930301143317@mantis.co.uk>
Lines: 290

Archive-name: atheism/resources
Alt-atheism-archive-name: resources
Last-modified: 11 December 1992
Version: 1.0

                                Atheist Resources

                        Addresses of Atheist Organizations
```

```
                                        USA

FREEDOM FROM RELIGION FOUNDATION

Darwin fish bumper stickers and assorted other atheist paraphernalia are
available from the Freedom From Religion Foundation in the US.

Write to:  FFRF, P.O. Box 750, Madison, WI 53701.
Telephone: (608) 256-8900

EVOLUTION DESIGNS

Evolution Designs sell the "Darwin fish".  It's a fish symbol, like the ones
Christians stick on their cars, but with feet and the word "Darwin" written
inside.  The deluxe moulded 3D plastic fish is $4.95 postpaid in the US.

Write to:  Evolution Designs, 7119 Laurel Canyon #4, North Hollywood,
           CA 91605.

People in the San Francisco Bay area can get Darwin Fish from Lynn Gold --
try mailing <figmo@netcom.com>.  For net people who go to Lynn directly, the
price is $4.95 per fish.

AMERICAN ATHEIST PRESS

AAP publish various atheist books -- critiq
```

Out[4]:
```
['Xref:',
 'cantaloupe.srv.cs.cmu.edu',
 'alt.atheism:49960',
 'alt.atheism.moderated:713',
 'news.answers:7054',
 'alt.answers:126\nPath:',
 'cantaloupe.srv.cs.cmu.edu!crabapple.srv.cs.cmu.edu!bb3.andrew.cmu.edu!news.sei.cmu.ed
 'mathew',
 '<mathew@mantis.co.uk>\nNewsgroups:',
 'alt.atheism,alt.atheism.moderated,news.answers,alt.answers\nSubject:',
 'Alt.Atheism',
 'FAQ:',
 'Atheist',
 'Resources\nSummary:',
 'Books,',
 'addresses,',
 'music',
 '--',
 'anything',
```

```
'related',
'to',
'atheism\nKeywords:',
'FAQ,',
'atheism,',
'books,',
'music,',
'fiction,',
'addresses,',
'contacts\nMessage-ID:',
'<19930329115719@mantis.co.uk>\nDate:',
'Mon,',
'29',
'Mar',
'1993',
'11:57:19',
'GMT\nExpires:',
'Thu,',
'29',
'Apr',
'1993',
'11:57:19',
'GMT\nFollowup-To:',
'alt.atheism\nDistribution:',
'world\nOrganization:',
'Mantis',
'Consultants,',
'Cambridge.',
'UK.\nApproved:',
'news-answers-request@mit.edu\nSupersedes:',
'<19930301143317@mantis.co.uk>\nLines:',
'290\n\nArchive-name:',
'atheism/resources\nAlt-atheism-archive-name:',
'resources\nLast-modified:',
'11',
'December',
'1992\nVersion:',
'1.0\n\n',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
```

```
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'Atheist',
'Resources\n\n',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'Addresses',
'of',
'Atheist',
'Organizations\n\n',
'',
'',
```

'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'USA\n\nFREEDOM',
'FROM',
'RELIGION',
'FOUNDATION\n\nDarwin',
'fish',
'bumper',
'stickers',
'and',
'assorted',
'other',
'atheist',
'paraphernalia',
'are\navailable',
'from',

```
'the',
'Freedom',
'From',
'Religion',
'Foundation',
'in',
'the',
'US.\n\nWrite',
'to:',
'',
'FFRF,',
'P.O.',
'Box',
'750,',
'Madison,',
'WI',
'53701.\nTelephone:',
'(608)',
'256-8900\n\nEVOLUTION',
'DESIGNS\n\nEvolution',
'Designs',
'sell',
'the',
'"Darwin',
'fish".',
'',
"It's",
'a',
'fish',
'symbol,',
'like',
'the',
'ones\nChristians',
'stick',
'on',
'their',
'cars,',
'but',
'with',
'feet',
'and',
'the',
'word',
'"Darwin"',
'written\ninside.',
'',
'The',
'deluxe',
```

```
'moulded',
'3D',
'plastic',
'fish',
'is',
'$4.95',
'postpaid',
'in',
'the',
'US.\n\nWrite',
'to:',
'',
'Evolution',
'Designs,',
'7119',
'Laurel',
'Canyon',
'#4,',
'North',
'Hollywood,\n',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'CA',
'91605.\n\nPeople',
'in',
'the',
'San',
'Francisco',
'Bay',
'area',
'can',
'get',
'Darwin',
'Fish',
'from',
'Lynn',
'Gold',
'--\ntry',
'mailing',
'<figmo@netcom.com>.',
```

```
 '',
 'For',
 'net',
 'people',
 'who',
 'go',
 'to',
 'Lynn',
 'directly,',
 'the\nprice',
 'is',
 '$4.95',
 'per',
 'fish.\n\nAMERICAN',
 'ATHEIST',
 'PRESS\n\nAAP',
 'publish',
 'various',
 'atheist',
 'books',
 '--',
 'critiq']
```

In [108]:
```
tokens = nltk.word_tokenize(corpus[0:2000])
tokens
tagged = nltk.pos_tag(tokens)
tagged
```

Out[108]:
```
[('Xref', 'NN'),
 (':', ':'),
 ('cantaloupe.srv.cs.cmu.edu', 'NN'),
 ('alt.atheism:49960', 'NN'),
 ('alt.atheism.moderated:713', 'NN'),
 ('news.answers:7054', 'JJ'),
 ('alt.answers:126', 'JJ'),
 ('Path', 'NN'),
 (':', ':'),
 ('cantaloupe.srv.cs.cmu.edu', 'NN'),
 ('!', '.'),
 ('crabapple.srv.cs.cmu.edu', 'NN'),
 ('!', '.'),
 ('bb3.andrew.cmu.edu', 'NN'),
 ('!', '.'),
 ('news.sei.cmu.edu', 'NN'),
 ('!', '.'),
 ('cis.ohio-state.edu', 'NN'),
 ('!', '.'),
 ('magnus.acs.ohio-state.edu', 'NN'),
```

```
('!', '.'),
('usenet.ins.cwru.edu', 'JJ'),
('!', '.'),
('agate', 'NN'),
('!', '.'),
('spool.mu.edu', 'NN'),
('!', '.'),
('uunet', 'NN'),
('!', '.'),
('pipex', 'NN'),
('!', '.'),
('ibmpcug', 'NN'),
('!', '.'),
('mantis', 'NN'),
('!', '.'),
('mathew', 'NN'),
('From', 'IN'),
(':', ':'),
('mathew', 'NN'),
('<', 'NNP'),
('mathew', 'NN'),
('@', 'NNP'),
('mantis.co.uk', 'NN'),
('>', 'NN'),
('Newsgroups', 'NNP'),
(':', ':'),
('alt.atheism', 'NN'),
(',', ','),
('alt.atheism.moderated', 'VBN'),
(',', ','),
('news.answers', 'NNS'),
(',', ','),
('alt.answers', 'NNS'),
('Subject', 'VBP'),
(':', ':'),
('Alt.Atheism', 'NN'),
('FAQ', 'NNP'),
(':', ':'),
('Atheist', 'JJ'),
('Resources', 'NNS'),
('Summary', 'JJ'),
(':', ':'),
('Books', 'NNP'),
(',', ','),
('addresses', 'VBZ'),
(',', ','),
('music', 'NN'),
('--', ':'),
```

```
('anything', 'NN'),
('related', 'JJ'),
('to', 'TO'),
('atheism', 'NN'),
('Keywords', 'NNS'),
(':', ':'),
('FAQ', 'NNP'),
(',', ','),
('atheism', 'NN'),
(',', ','),
('books', 'NNS'),
(',', ','),
('music', 'NN'),
(',', ','),
('fiction', 'NN'),
(',', ','),
('addresses', 'NNS'),
(',', ','),
('contacts', 'NNS'),
('Message-ID', 'NNP'),
(':', ':'),
('<', 'NN'),
('19930329115719', 'CD'),
('@', 'NN'),
('mantis.co.uk', 'NN'),
('>', 'NN'),
('Date', 'NNP'),
(':', ':'),
('Mon', 'NNP'),
(',', ','),
('29', 'CD'),
('Mar', 'NNP'),
('1993', 'CD'),
('11:57:19', 'CD'),
('GMT', 'NNP'),
('Expires', 'NNS'),
(':', ':'),
('Thu', 'NNP'),
(',', ','),
('29', 'CD'),
('Apr', 'NNP'),
('1993', 'CD'),
('11:57:19', 'CD'),
('GMT', 'NNP'),
('Followup-To', 'NN'),
(':', ':'),
('alt.atheism', 'NN'),
('Distribution', 'NN'),
```

```
(':', ':'),
('world', 'NN'),
('Organization', 'NN'),
(':', ':'),
('Mantis', 'NN'),
('Consultants', 'NNS'),
(',', ','),
('Cambridge', 'NNP'),
('.', '.'),
('UK', 'NNP'),
('.', '.'),
('Approved', 'VBN'),
(':', ':'),
('news-answers-request', 'JJ'),
('@', 'NNP'),
('mit.edu', 'NN'),
('Supersedes', 'NNP'),
(':', ':'),
('<', 'NN'),
('19930301143317', 'CD'),
('@', 'NN'),
('mantis.co.uk', 'NN'),
('>', 'JJ'),
('Lines', 'NNS'),
(':', ':'),
('290', 'CD'),
('Archive-name', 'NN'),
(':', ':'),
('atheism/resources', 'NNS'),
('Alt-atheism-archive-name', 'VBP'),
(':', ':'),
('resources', 'NNS'),
('Last-modified', 'JJ'),
(':', ':'),
('11', 'CD'),
('December', 'NNP'),
('1992', 'CD'),
('Version', 'NNP'),
(':', ':'),
('1.0', 'CD'),
('Atheist', 'NNP'),
('Resources', 'NNPS'),
('Addresses', 'NNP'),
('of', 'IN'),
('Atheist', 'NNP'),
('Organizations', 'NNP'),
('USA', 'NNP'),
('FREEDOM', 'NNP'),
```

```
('FROM', 'NNP'),
('RELIGION', 'NNP'),
('FOUNDATION', 'NNP'),
('Darwin', 'NNP'),
('fish', 'JJ'),
('bumper', 'NN'),
('stickers', 'NNS'),
('and', 'CC'),
('assorted', 'VBD'),
('other', 'JJ'),
('atheist', 'JJ'),
('paraphernalia', 'NNS'),
('are', 'VBP'),
('available', 'JJ'),
('from', 'IN'),
('the', 'DT'),
('Freedom', 'NN'),
('From', 'NNP'),
('Religion', 'NNP'),
('Foundation', 'NNP'),
('in', 'IN'),
('the', 'DT'),
('US', 'NNP'),
('.', '.'),
('Write', 'NNP'),
('to', 'TO'),
(':', ':'),
('FFRF', 'NNP'),
(',', ','),
('P.O', 'NNP'),
('.', '.'),
('Box', 'NNP'),
('750', 'CD'),
(',', ','),
('Madison', 'NNP'),
(',', ','),
('WI', 'NNP'),
('53701', 'CD'),
('.', '.'),
('Telephone', 'NN'),
(':', ':'),
('(', '('),
('608', 'CD'),
(')', ')'),
('256-8900', 'CD'),
('EVOLUTION', 'NNP'),
('DESIGNS', 'NNP'),
('Evolution', 'NNP'),
```

```
('Designs', 'NNP'),
('sell', 'VB'),
('the', 'DT'),
('``', '``'),
('Darwin', 'NNP'),
('fish', 'NN'),
("''", "''"),
('.', '.'),
('It', 'PRP'),
("'s", 'VBZ'),
('a', 'DT'),
('fish', 'JJ'),
('symbol', 'NN'),
(',', ','),
('like', 'IN'),
('the', 'DT'),
('ones', 'NNS'),
('Christians', 'NNPS'),
('stick', 'VBP'),
('on', 'IN'),
('their', 'PRP$'),
('cars', 'NNS'),
(',', ','),
('but', 'CC'),
('with', 'IN'),
('feet', 'NNS'),
('and', 'CC'),
('the', 'DT'),
('word', 'NN'),
('``', '``'),
('Darwin', 'NNP'),
("''", "''"),
('written', 'VBN'),
('inside', 'RB'),
('.', '.'),
('The', 'DT'),
('deluxe', 'NN'),
('moulded', 'VBD'),
('3D', 'CD'),
('plastic', 'JJ'),
('fish', 'NN'),
('is', 'VBZ'),
('$', '$'),
('4.95', 'CD'),
('postpaid', 'NN'),
('in', 'IN'),
('the', 'DT'),
('US', 'NNP'),
```

```
('.', '.'),
('Write', 'NNP'),
('to', 'TO'),
(':', ':'),
('Evolution', 'NN'),
('Designs', 'NNP'),
(',', ','),
('7119', 'CD'),
('Laurel', 'NNP'),
('Canyon', 'NNP'),
('#', '#'),
('4', 'CD'),
(',', ','),
('North', 'NNP'),
('Hollywood', 'NNP'),
(',', ','),
('CA', 'NNP'),
('91605', 'CD'),
('.', '.'),
('People', 'NNS'),
('in', 'IN'),
('the', 'DT'),
('San', 'NNP'),
('Francisco', 'NNP'),
('Bay', 'NNP'),
('area', 'NN'),
('can', 'MD'),
('get', 'VB'),
('Darwin', 'NNP'),
('Fish', 'NNP'),
('from', 'IN'),
('Lynn', 'NNP'),
('Gold', 'NNP'),
('--', ':'),
('try', 'VB'),
('mailing', 'VBG'),
('<', 'JJ'),
('figmo', 'NN'),
('@', 'NNP'),
('netcom.com', 'NN'),
('>', 'NNP'),
('.', '.'),
('For', 'IN'),
('net', 'JJ'),
('people', 'NNS'),
('who', 'WP'),
('go', 'VBP'),
('to', 'TO'),
```

```
 ('Lynn', 'NNP'),
 ('directly', 'RB'),
 (',', ','),
 ('the', 'DT'),
 ('price', 'NN'),
 ('is', 'VBZ'),
 ('$', '$'),
 ('4.95', 'CD'),
 ('per', 'IN'),
 ('fish', 'NN'),
 ('.', '.'),
 ('AMERICAN', 'NNP'),
 ('ATHEIST', 'NNP'),
 ('PRESS', 'NNP'),
 ('AAP', 'NNP'),
 ('publish', 'JJ'),
 ('various', 'JJ'),
 ('atheist', 'NN'),
 ('books', 'NNS'),
 ('--', ':'),
 ('critiq', 'NN')]
```

In [79]:
```
# Regex
## Find all emails
email_pattern = '[\w\._]+@\w+\.[\w\.]+'

## Test email cases
print(re.match(email_pattern,'tony'))
print(re.match(email_pattern,'tony@'))
print(re.match(email_pattern,'tony@gmail'))
print(re.match(email_pattern,'tony@gmail.'))
print(re.match(email_pattern,'tony@gmail.com'))
print(re.match(email_pattern,'@gmail.com'))
print(re.match(email_pattern,'gmail.com'))
print(re.match(email_pattern,'tony@u.northwestern.edu'))
print(re.match(email_pattern,'t_o_n_Y..@gmail.co.uk'))
print()
```

```
None
None
None
None
<re.Match object; span=(0, 14), match='tony@gmail.com'>
None
None
<re.Match object; span=(0, 23), match='tony@u.northwestern.edu'>
<re.Match object; span=(0, 21), match='t_o_n_Y..@gmail.co.uk'>
```

```
In [73]: all_emails = re.findall(email_pattern,corpus)
         print(len(all_emails))
         print(all_emails[0:2000])
```

```
Out[73]: ['93111.074840LIBRBA@BYUVM.BITNET',
          'C5vGyD.H7s@acsu.buffalo.edu',
          'C5vGyD.H7s@acsu.buffalo.edu',
          'psyrobtw@ubvmsd.cc.buffalo.edu',
          '93111.074840LIBRBA@BYUVM.BITNET',
          'LIBRBA@BYUVM.BITNET',
          'psyrobtw@ubvms.cc.buffalo.edu',
          'kltensme@kt8127.b23a.ingr.com',
          'ktikkane@phoenix.oulu.fi',
          '1993Apr27.151411.8912@ousrvr.oulu.fi',
          'news@ousrvr.oulu.fi',
          '1993Apr26.174041.25444@daffy.cs.wisc.edu',
          'mccullou@snake2.cs.wisc.edu',
          'ktikkane@phoenix.oulu.fi',
          'kltensme@infonode.ingr.com',
          '1993Apr27.153731.623@infonode.ingr.com',
          'C5wIA1.4Hr@apollo.hp.com',
          'hil@agate.berkeley.edu',
          'hil@agate.berkeley.edu',
          'isaackuo@spam.berkeley.edu',
          'C5wIA1.4Hr@apollo.hp.com',
          'goykhman@apollo.hp.com',
          'isaackuo@math.berkeley.edu',
          'kltensme@kt8127.b23a.ingr.com',
          'frank@D012S658.uucp',
          'nbs@horus.ap.mchp.sni.de',
          'jep@kyle.eitech.com',
          '42g@horus.ap.mchp.sni.de',
          '9be@squick.eitech.com',
          '9be@squick.eitech.com',
          'ekr@squick.eitech.com',
          'odwyer@sse.ie',
          'frank@D012S658.uucp',
          'npc@horus.ap.mchp.sni.de',
          'exuptr.1436.0@exu.ericsson.se',
          'C6224D.1EH@news.cso.uiuc.edu',
          '1993Apr26.163627.11364@csrd.uiuc.edu',
          '1993Apr26.163627.11364@csrd.uiuc.edu',
          'skinner@uiuc.edu',
          'odwyer@sse.ie',
          'C5nKwo.ILu@sunfish.usd.edu',
          'rfox@charlie.usd.edu',
          'rfox@charlie.usd.edu',
          'news@sunfish.usd.edu',
```

```
'1993Mar16.200648.8005@rambo.atlanta.dg.com',
'C5FtJt.885@sunfish.usd.edu',
'1993Apr15.012537.26867@nntpd2.cxo.dec.com',
'1993Apr15.012537.26867@nntpd2.cxo.dec.com',
'sharpe@nmesis.enet.dec.com',
'C5FtJt.885@sunfish.usd.edu',
'rfox@charlie.usd.edu',
'1993Apr10.213547.17644@rambo.atlanta.dg.com',
'wpr@atlanta.dg.com',
'8frJnwm00iUxA2cXNE@andrew.cmu.edu',
'QfpBOqu00WBKIAO81C@andrew.cmu.edu',
'1993Apr21.182127.23528@advtech.uswest.com',
'1993Apr24.002509.4017@midway.uchicago.edu',
'1993Apr26.150845.28537@advtech.uswest.com',
'1993Apr26.150845.28537@advtech.uswest.com',
'Novak@advtech.uswe',
'eeb1@midway.uchicago.edu',
'neese@cerritos.edu',
'1993Apr27.083523.8145@cerritos.edu',
'sbuckley@fraser.sfu.ca',
'sbuckley.735506328@sfu.ca',
'news@sfu.ca',
'C5sqtK.F6r@noose.ecn.purdue.edu',
'sbuckley.735337212@sfu.ca',
'C5swMr.H3w@noose.ecn.purdue.edu',
'muttiah@thistle.ecn.purdue.edu',
'sbuckley.735337212@sfu.ca',
'sbuckley@fraser.sfu.ca',
'sbuckley@fraser.sfu.ca',
'sbuckley.735506617@sfu.ca',
'news@sfu.ca',
'1993Apr22.133142.23772@ifi.uio.no',
'joakimr@ifi.uio.no',
'sbuckley@fraser.sfu.ca',
'sbuckley.735895208@sfu.ca',
'news@sfu.ca',
'1993Apr27.015537.9149@magnus.acs.ohio',
'pboxrud@magnus.acs.ohio',
'bakerj@gtephx.UUCP',
'1993Apr26.204814.29342@gtephx.UUCP',
'19APR199313180801@utarlg.uta.edu',
'bskendigC5qyJ2.GEw@netcom.com',
'1993Apr23.111105.7703@ifi.uio.no',
'1993Apr23.111105.7703@ifi.uio.no',
'joakimr@ifi.uio.no',
'C5u5nv.JGs@ncratl.AtlantaGA.NCR.COM',
'mwilson@ncratl.AtlantaGA.NCR.COM',
'C5sqyA.F7v@noose.ecn.purdue.edu',
```

```
'tbrent@bank.ecn.purdue.edu',
'GMILLS@CHEMICAL.watstar.uwaterloo.ca',
'GMILLS.45.735930174@CHEMICAL.watstar.uwaterloo.ca',
'news@watserv2.uwaterloo.ca',
'o1v@horus.ap.mchp.sni.de',
'930422.113807.7Q9.rusnews.w165w@mantis.co.uk',
'3lv@horus.ap.mchp.sni.de',
'930426.140835.4f1.rusnews.w165w@mantis.co.uk',
'930426.140835.4f1.rusnews.w165w@mantis.co.uk',
'mathew@mantis.co.uk',
'1993Apr27.073723.18577@csis.dit.csiro.au',
'1993Apr27.073723.18577@csis.dit.csiro.au',
'prl@csis.dit.csiro.au',
'king@ctron.com',
'dk@imager.llnl.gov',
'pharvey@quack.kfu.com',
'f2ujHTW@quack.kfu.com',
'1rc1f3INN7rl@emx.cc.utexas.edu',
'1rc1f3INN7rl@emx.cc.utexas.edu',
'bill@emx.cc.utexas.edu',
'KEVXU@CUNYVM.BITNET',
'93117.080750KEVXU@CUNYVM.BITNET',
'a8a@geraldo.cc.utexas.edu',
'1dx802lO40Rq01@JUTS.ccc.amdahl.com',
'1993Apr26.231845.13843@digi.lonestar.org',
'1993Apr26.231845.13843@digi.lonestar.org',
'qpalo@digi.lonestar.org',
'pharvey@quack.kfu.com',
'f2ui5RH@quack.kfu.com',
'kmr4.1697.735654694@po.CWRU.edu',
'C64H4w.BFH@darkside.osrhe.uoknor.edu',
'C64H4w.BFH@darkside.osrhe.uoknor.edu',
'bil@okcforum.osrhe.edu',
'kmr4@po.CWRU.edu']
```

In [110]:
```
## Find all addresses
# date_pattern_1 = '[01]\d-[0123]\d-[12]\d{3}'
# date_pattern_2 = '[1-9]-[0123]\d-[12]\d{3}'

date_patterns = re.compile(r"""
 (

     (1[012]|0?[1-9]|jan|january|feb|february|mar|march|apr|april|may|jun|june|jul|ju
   | ([12]\d|0?[1-9]|3[01])(st|nd|rd|th)*[ ]*([/-]|[ ]+)[ ]*(1[012]|0?[1-9]|jan|janua
   | ([12]\d{3}|\d{2})[ ]*([/-]|[ ]+)[ ]*(1[012]|0?[1-9]|jan|january|feb|february|mar

)
""", re.VERBOSE|re.IGNORECASE)
```

```
        ## Does not account for different numbers of days in months
        ## Does not include date formats with year between month and day (uncommon)
        ## Does not include date formats in format YYYY-DD-MM (uncommon)


        ## Test date cases
        print(re.match(date_patterns, '1-1-2000'))
        print(re.match(date_patterns, '12-35-2000'))
        print(re.match(date_patterns, 'January 1st 2000'))
        print(re.match(date_patterns, 'January 1 1212'))
        print(re.match(date_patterns, 'January 1sd 2000'))
        print(re.match(date_patterns, '1/1/00'))
        print(re.match(date_patterns, '20/3/2000'))
        print(re.match(date_patterns, '20/15/2000'))
        print(re.match(date_patterns, 'january\n20\n2000'))
        print(re.match(date_patterns, '2000 JAN 20'))
        print(re.match(date_patterns, '2010 February 18'))
        print(re.match(date_patterns, '2010-29-2'))
        print(re.match(date_patterns, '2010-2-28'))
        print(re.match(date_patterns, 'Jul 4 1776'))
        print(re.match(date_patterns, 'Octopus 23 99'))
```

```
<re.Match object; span=(0, 8), match='1-1-2000'>
None
<re.Match object; span=(0, 16), match='January 1st 2000'>
<re.Match object; span=(0, 14), match='January 1 1212'>
None
<re.Match object; span=(0, 6), match='1/1/00'>
<re.Match object; span=(0, 9), match='20/3/2000'>
None
None
<re.Match object; span=(0, 11), match='2000 JAN 20'>
<re.Match object; span=(0, 16), match='2010 February 18'>
None
<re.Match object; span=(0, 9), match='2010-2-28'>
<re.Match object; span=(0, 10), match='Jul 4 1776'>
None
```

```
In [103]: all_dates = re.findall(date_patterns, corpus)
          all_dates = [x[0] for x in all_dates]
          print(len(all_dates))
          print(all_dates[0:2000])
```

```
25525
['29 Mar 1993', '29 Apr 1993', '11 December 1992', '5 Apr 1993', '6 May 1993', '5 April 1993',
```