Problem One:

First, we need to get the data from Project Gutenbery.

This step took me some time to get the book information [https://github.com/c-w/Gutenberg/blob/master/gutenberg/acquire/text.py, http://stackoverflow.com/questions/19588952/python-reading-rdf-files-scraping-gutenberg-books]

I have known the RDF catalog, and then install the “rdflib” package and “urllib2” package. For RDF [http://rdflib.readthedocs.org/en/latest/intro_to_creating_rdf.html], there are three variables:
(1). URL reference — data resources
(2). Blank Nodes — data resources
(3). Literals — attribute values

The first step is to download the RDF from Project Gutenberg [http://www.gutenberg.org/wiki/Gutenberg%3aFeeds#Current_RDF_Format]. There are 50024 RDFs in this compression file.

The code to extract the 50024 files path in my local computer is:

```
# -*- coding: utf-8 -*-
"""
Created on Mon Sep 21 14:18:31 2015

@author: weizhi
"""
import glob, os
import re
# https://docs.python.org/2/library/os.html
def findFilePath(dataPath):
    """
    Input: dataPath: RDF folder path
          From: http://www.gutenberg.org/wiki/Gutenberg%3aFeeds#Current_RDF_Format
    Output:
            Each rdf file path

    """
    dataFileName = os.listdir(dataPath)
    filePaths = []
    keys = []
    for i in range(1,len(dataFileName)):
        item = dataFileName[i]
        keys.append(item)
        pathFolder = dataPath + '/' + item
      #  print pathFolder
        nameFile = os.listdir(pathFolder)
       # print nameFile
        filePaths.append(pathFolder  + '/'+nameFile[0])
    return (filePaths,keys)

rdfPath = '/Users/weizhi/Downloads/cache/epub'
dataFile,keys = findFilePath(rdfPath)

```

Second Step: Parse the RDF log files and extract the title, author name and release time, then we have data to do the data mining ideas.

For example:

Idea one: build a classifier that will take some text as input and detect which author it was written by

Idea two: discover a set of words that most distinguish a given author

I have wrote the function to extract the author name and title, release time from the RDF

Steps:

(1). Find the URL from the RDF — URL reference, and use the urllib2 to download the txt.

After downloading, we can find the author information from the data. However, after running the my code, they gave me forbidden to download the books. Please check the code and information below:

```
75      def readURL(self,subject,key):
76          """
77          https://gist.github.com/andreasvc/b3b4189120d84dec8857
78          """
79          endString =  '.txt'
80          URL = []
81          for item in subject:
82              if item.endswith(endString):
83                  URL.append(item)
84          if len(URL) !=0:
85              try:
86                  data = urllib2.urlopen(URL[0]).read()# read only 20 000 chars
87                  data = data.split("\n") # then split it into lines
88                 # return data
89              except:
90                  pass
91          if len(data) !=0:
92              return data
93          else:
94              return URL[0]
95      def metaData(self,data):
96        #  for line in data
97          Dict = {}
98          tag = ['author','title','release date']
99          count = 0
100         index = 0
101         for line in data:
102             for item in tag:
103                 if line.lower().startswith(item+':'):
104                     Dict[item] = line.split(':')[1].rstrip()
105                     count +=1
106             index +=1
107             if count ==3 or index>200:
108                 return Dict
109                 break
110
111         return Dict
```

Error 403

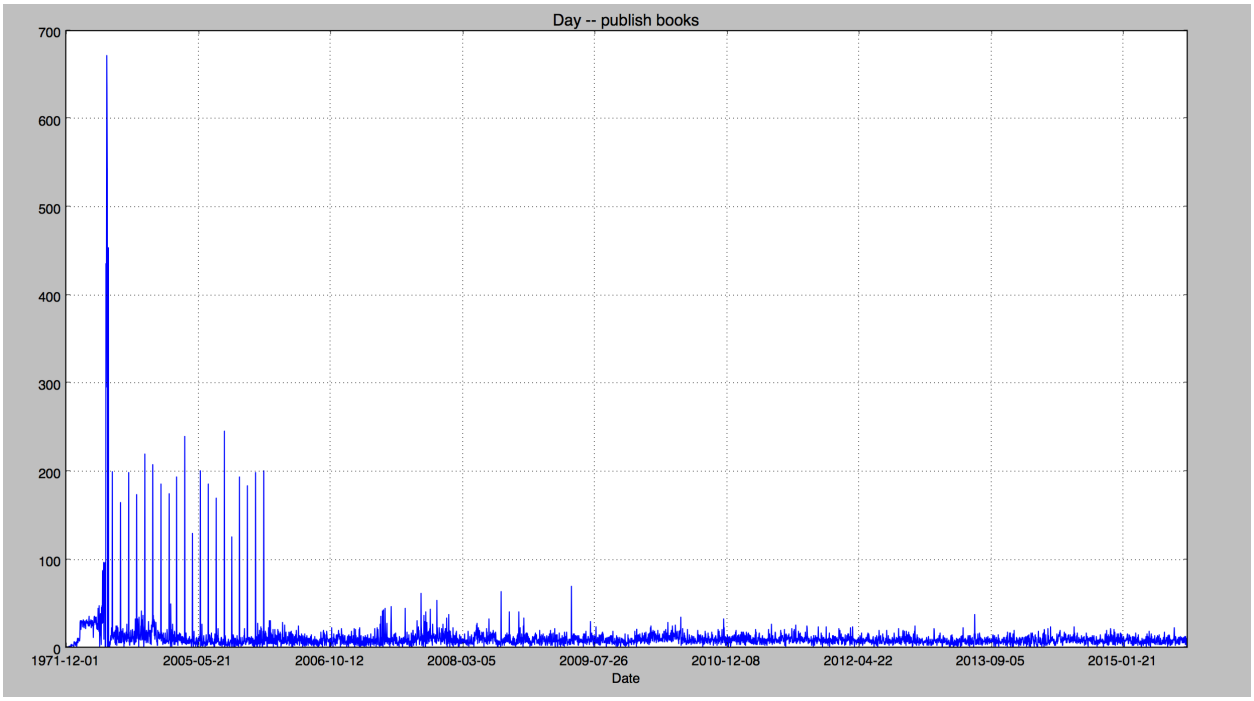
Maybe you have just a wrong url. Go to http://www.gutenberg.org/ebooks/ first to see if the error persists.

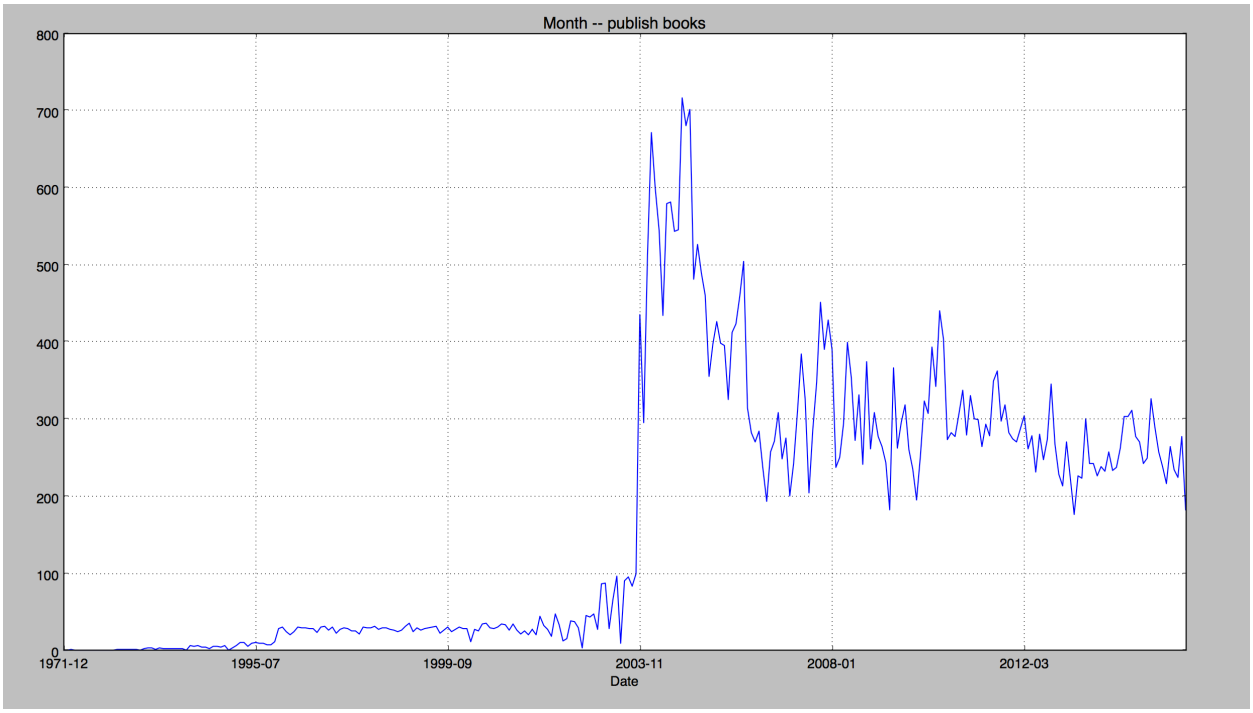
If you get the error again check that you:

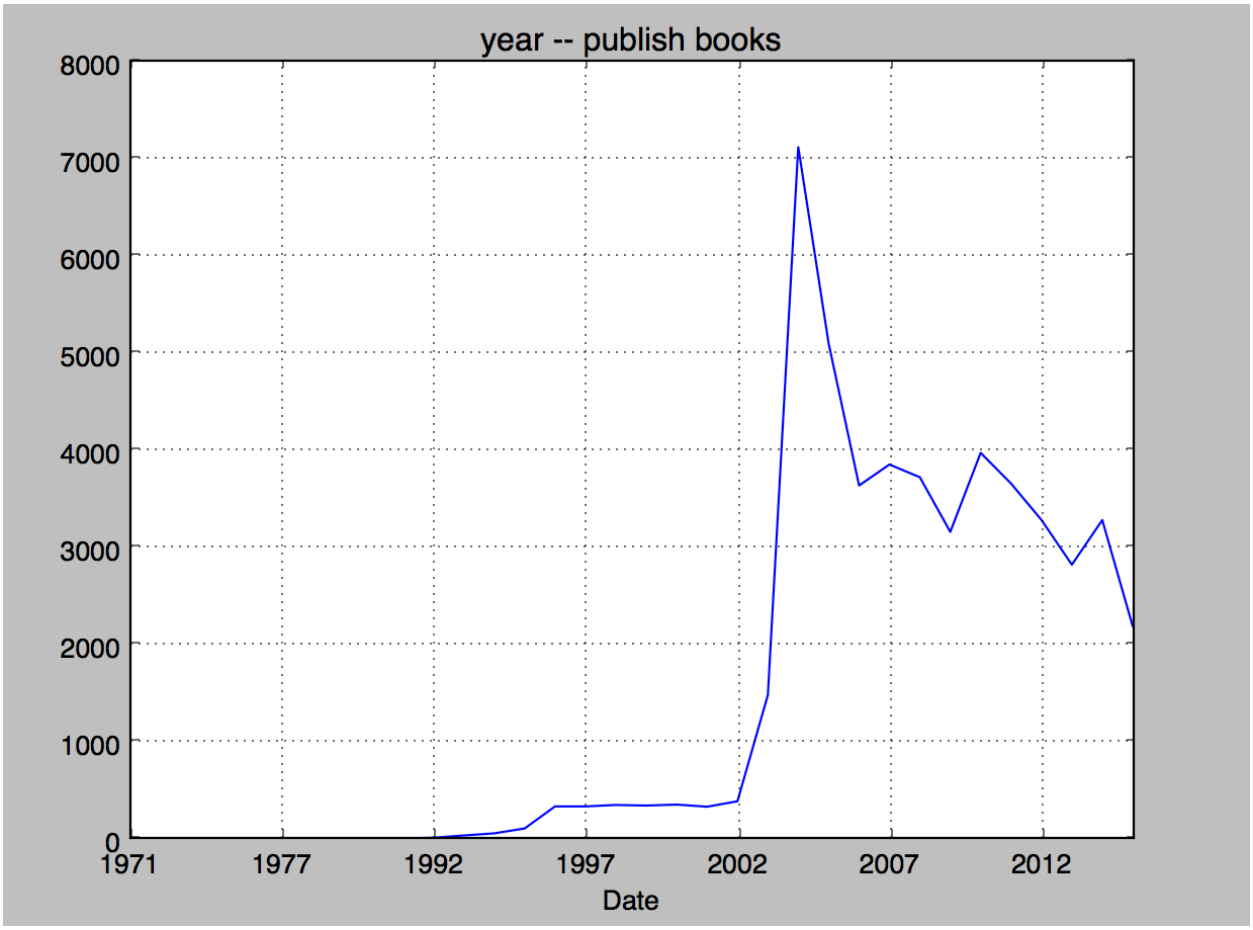
- Don't use anonymizers, open proxies, VPNs, or TOR to access Project Gutenberg. This includes the Google proxies that are used by Chrome.
- Don't access Project Gutenberg from hosted servers.
- Don't use automated software to download lots of books. We have a limit on how fast you can go while using this site. If you surpass this limit you get blocked for 24h.
- We have a daily limit on how many books you can download. If you exceeded this limit you get blocked for 24h.
- If you use the RSS feed, set your update interval to 24 hours.

If you are sure that none of the above applies to you, and wish us to investigate the problem, we need to know your IP address. Go to this site, don't sign up, just copy the IP address (it looks like: 12.34.56.78 but your numbers will be different) and mail it to us. If that page also shows a proxy address, we need that one too.

(2). After this step, I found that we can find the book release date from the Literals, after cacladting the 50024 books publish time, we can find the result below:




Month -- publish books
800
700
600
500
400
300
200
100
0
1971-12
1995-07
1999-09
2003-11
2008-01
2012-03
Date


year -- publish books
8000
7000
6000
5000
4000
3000
2000
1000
0
1971
1977
1992
1997
2002
2007
2012
Date

Third step:

After getting the forbidden information, I have search “wheel” (Open Source) to download the books. I found that Gutenberg 0.4.0 (https://pypi.python.org/pypi/Gutenberg). I have tried the function, I can download the text, however, the meta_data did not work.

In case of forbidden again, I decide to download 10000 books randomly to local computer. And using the function in step one to get the title and author information.

(Not finish yet, need more time. In this point, we have got the author, title, txt, etc, use NLTK, Scikit-learn to do the data mining projects)