HathiTrust_analytics_v2

November 27, 2017

1 HathiTrust Usage Analytics and Metadata Analysis Notes

1.1 High-Level Steps

- Scrape analytics
- Process Analytics to extract volume IDs and usage counts
  - Fix the dollar sign barcode issue
- Ingest Hathifiles into postgres database
- Match IDs from analytics to current metadata in Hathifiles
- Create visualizations of interesting facets

1.2 Process Notes

1.2.1 Scrape analytics

Using Pyganalytics: https://github.com/chrpr/pyganalytics

Scrape the analytics using something like this (do this 4x, once for each quarter of the year, adjusting command as needed):

```
for i in {3..7}; do time python analytics.py -o ~/PATH/TO/DATA/OUTPUT/uniqueEvents_201$i\_01_01-
```

The yml config file is the following for the HathiTrust pageturner analytics account:

```
query:
    metrics: ga:uniqueEvents
    dimensions: ga:pagePath
    sort:
    filters:
profile: 26478114
```

Note: I adjusted the delimiters used in the Pyganalitcs `analytics.py` script because page paths in Google Analytics contain everything under the sun and I needed to amend to try and find something that would be unique enough to work as a field separator that Pandas can recognize. However, this also means that you have to use the Python parsing engine when reading the CSV into Pandas, but given that this is a one-time operation I think the tradeoff here of doing less surgery on the analytics CSVs is worth the potential slowdown here.

I also edited the `analytics.py` script to run daily instead of weekly in order to try and capture more granular results. See notes on how to do that in the Pyganalytics readme.

Notes: - these tended to run between 30-60 minutes for each quarter on my machine, via a wireless connection - Broken into annual quarters just to keep file sizes more manageable and prevent less data loss if API errored out mid-file - this uses Google Analytics API v3, not v4 which is current, so may break sooner or later Set up an API key as described in the readme - There is a 10000 API call per profile ID for the Google Analytics API; as a result, I had to run this over the course of 3 days. - Dask install via `pip install dask[dataframe]` won't work in zsh for some reason, case you use that

1.2.2 Other notes

Important note: there is sampling happening in the analytics A large number of events are getting grouped together under "(other)" I think the sampling is good, and should give a directional idea of trends, etc. But really, the analytics should be fixed before drawing 100% final conclusions

Also something to explore: there are a limited set of results with zero pageviews, even just limiting to event triggers

1.3 Next things to fix

- Data viz of monographs vs. serials

1.3.1 Setup

```
In [1]: %matplotlib inline

        import pandas as pd
        import dask.dataframe as dd #USE `pip install dask[dataframe]` (does not work in zsh for
        from dask.diagnostics import ProgressBar

        import matplotlib.pyplot as plt

        pbar = ProgressBar()
        pbar.register()
```

1.3.2 Process Analytics to extract volume IDs and usage counts

```
In [9]: def extract_ids():

            '''
            Uses Dask to extracts HathiTrust IDs from the raw analytics logs and writes them to
            '''

            df = dd.read_csv('./data/uniqueEvents_201*.csv', delimiter='\|\~', engine='python')

            #Mix of my regex experimentation and pattern supplied by Angelina Z at Hathi:
            pattern = '(?:id=|[a-z0-9]\/)([a-z][a-z0-9]{1,3}\.\$?[a-z0-9._:\/\-]+)'

            #Extract the ID matches into a new column
            df['id'] = df['ga:pagePath'].str.extract(pattern, expand=False)
            #Limit to just rows with ID matches
            df = df[df['id'].notnull()]

            #Remove rows that have 'skin=crms'
            df = df[df['ga:pagePath'].str.contains('skin=crms') == False]

            #Remove some junk punctuation from the end of some IDs
            df['id'] = df['id'].str.rstrip('._#')

            #Write the results to csvs
            df.to_csv('./data/all_ids_*.csv', index=False)

        #This takes roughly 26min to run

        %time extract_ids()
```

```
[########################################] | 100% Completed | 25min 24.7s
CPU times: user 23min 23s, sys: 5min 11s, total: 28min 34s
Wall time: 25min 25s
```

```
In [11]: def ids_count():

             '''
             Uses output csvs of extract_ids() function to create a tuple of volume IDs paired u
             (either unique or aggregate pageview counts depending on how analytics were scraped
             and writes them to a CSV
             '''

             #Read in all the files with the extracted IDs and their counts
             df = dd.read_csv('./data/all_ids_*.csv')

             #Group that data by volume identifier, and then record the sum total of all hit cou
             ids = df.groupby(by=['id'])['count'].sum()

             #Turn dask dataframe into pandas dataframe so we can use sort_values (not implement
             ids = ids.compute()

             #Do some column naming, and then export to a CSV
             ids.index.name = 'id'
             ids.columns = ['count']
             ids = ids.sort_values(ascending=False).to_csv('./csv/all_counts_sorted.csv', header

         #This takes roughly 12min to run
         %time ids_count()
```

```
[########################################] | 100% Completed | 11min 31.6s
CPU times: user 11min 3s, sys: 1min 24s, total: 12min 27s
```

```
Wall time: 11min 42s
```

```
In [ ]: def fix_dollar_sign_ids():

            '''
            Need to Fix UC $ issue, and roll up the total counts so they're only counted as sing
            There's some ugly pandas in here, could certainly be more effficient but it's quick
            that I'm calling it done for now
            '''

            #Need numpy briefly to do the conditional check later using .where() method
            import numpy as np

            #Import the extracted IDs and counts
            df = pd.read_csv('./csv/all_counts_sorted.csv')

            #New dataframe of just the potentially affected dollar sign IDs
            dollars = df[df['id'].str.contains('\$')]

            #New column with the id version minus dollar sign
            dollars.loc[:,'fixed_id'] = dollars['id'].str.replace('$','')

            #Merge with original data to get access to all counts we need to sum
            merged = df.merge(dollars, left_on='id', right_on='fixed_id', suffixes=['_df','_doll

            #Sum the hit counts for the dollar sign IDs and the non dollar sign IDs
            merged.loc[:,'total'] = merged['count_df'] + merged['count_dollars']

            #Merge the totals with the original dataset again, but in a new df just to be carefu
            df2 = df.merge(merged, how='outer', left_on='id', right_on='id_dollars')

            #Update the original 'count' column to the holistic total where needed
            df2.loc[:, 'count'] = np.where(df2['total'].notnull() == True, df2['total'], df['cou

            #Remove the non-dollar sign IDs from the df, since the dollar-sign-id totals now ref
            #Also removes the now extraneous extra columns
            df2 = df2[df2['id'].isin(df2['id_df']) == False][['id','count']]

            #Write it all out to a new csv
            df2.sort_values('count', ascending=False).to_csv('./csv/all_counts_sorted_dollar_fix

        #This takes like 35 seconds to run
        %time fix_dollar_sign_ids()
```

1.4 Postgres import of Hathifiles

```
In [12]: #Set up the postgres connection

         import sqlalchemy as sa

         def connect(user, password, db, host='localhost', port=5432):
             '''Returns a connection and a metadata object'''
             # We connect with the help of the PostgreSQL URL
             url = 'postgresql://{}:{}@{}:{}/{}'
             url = url.format(user, password, host, port, db)

             # The return value of create_engine() is our connection object
             con = sa.create_engine(url, client_encoding='utf8')

             # We then bind the connection to MetaData()
             meta = sa.MetaData(bind=con).reflect()

             return con, meta

         con, meta = connect('postgres', '', 'hathifiles')
```

```
In [ ]: #Import the Hathifiles into a postgres database


        '''
        This cell does require ones step not included here which is adding the header row
        to the txt file with the column names, which I do in an elegant but fine way:
        - Copy row headers, including tab delimiters, to new file
        - Append text of massive hathi text file (~4gb) to that new file
            `cat hathi_full_20171101.txt >> headers_for_hathi_full_20171101.txt`
        - Delete old file, and rename new file same as the old (but now includes column headers)

        '''

        #Struggled for a long time with this, but turns out the delimiter needs to be r'\t', not
        hathi_data = pd.read_csv('./data/hathifiles/hathi_full_20171101.txt', engine='python', d


        def postgres_import():
            i = 0
            for chunk in hathi_data:
                chunk = chunk[['id','access','rights','hathitrust_record_number','enumeration_ch
                try:
                    chunk.to_sql('hathifiles', con, if_exists='append')
                    print i, chunk.index[0]
                    i += 1
                except:
                    print chunk.index[0]
        #This takes roughly 2 hours 10min to run on my macbook air
        #%time postgres_import()
```

1.4.1 Match IDs from analytics to current metadata in Hathifiles

```
In [15]: def get_access_and_date():

             gf = pd.read_csv('./csv/all_counts_sorted_dollar_fixed.csv')

             gf['access'] = ''
             gf['date'] = ''
             gf['title'] = ''
             gf['oclc'] = ''
         #     gf['format'] = ''
         #     gf['pub_place'] = ''
             print len(gf)

             header = 'id,title,access,date,oclc'
             text_file = open("./csv/all_id_title_access_date_oclc.csv", "w")
             text_file.write(header+'\n')
             text_file.close()

             for i in range(0,len(gf),250000):
                 xf = gf[i:(i+250000)]
                 ids = []
                 for index, row in xf.iterrows():
                     ids.append(row['id'])
                 x = pd.read_sql_query("select id, title, access, publication_date, oclc_numbers
                                         from hathifiles where id in"+str(tuple(ids)), con=con)
                 x.to_csv('./csv/all_id_title_access_date_oclc.csv', mode='a', encoding='utf-8',
                 if i % 10000 == 0:
                     print i

         #This takes about 33min
         %time get_access_and_date()
```

```
5570514
0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
2250000
```

```
2500000
2750000
3000000
3250000
3500000
3750000
4000000
4250000
4500000
4750000
5000000
5250000
5500000
CPU times: user 11min 31s, sys: 47.8 s, total: 12min 19s
Wall time: 33min 49s
```

1.5 Analysis and Viz steps

Below cells are messy now, and some are just checks to make sure outputs look roughly correct
Cleanup TK

```
In [2]: #Read our various CSVs into dataframes so we can work with them
        counts = pd.read_csv('./csv/all_counts_sorted.csv')
        d_counts = pd.read_csv('./csv/all_counts_sorted_dollar_fixed.csv')
        full = pd.read_csv('./csv/all_id_title_access_date_oclc.csv')
```

```
In [3]: #Gut check on dollar sign fix and missing volumes
        cs = set(counts['id'])
        ds = set(d_counts['id'])
        fs = set(full['id'])

        #Total number of dollar sign ids fixed
        print("dollar sign ids fixed: %s" % (len(cs - ds)))


        '''
        To get a sense if we've missed things with the above data transformations
        This spits out the things that the parsing found as IDs, but couldn't find in the HathiF
        Generally, there are a few hundred here, things that have been added to Hathi via ingest
        Are available to analytics events, but haven't had their metadata added to the monthly H
        Gut checking, a delta of less than a thousand a month seems ok, and not something I'm su
        in terms of skewing data
        '''
        diff = ds - fs

        missing = []
        for d in diff:
            if "uc1" not in d:
                missing.append(d)
            else:
                pass
        print ("Total non-dollar sign IDs in data, but not found in hathifiles: %s" % (len(missi

        for item in missing:
            print item
```

```
dollar sign ids fixed: 24061
Total non-dollar sign IDs in data, but not found in hathifiles: 683
txa.ark:/81423/m3rh0q
gri.ark:/13960/t86j0w112
txa.ark:/81423/m3j63w
txa.ark:/81423/m3r90n
txa.ark:/81423/m3jk8k
uiuo.ark:/13960/t88h4v72m
mdp.39015013273209
txa.ark:/81423/m3fk8n
osu.32435083339473
txa.ark:/81423/m3f048
txa.ark:/81423/m3nk8h
txa.ark:/81423/m3x050
txa.ark:/81423/m36p8h
txa.ark:/81423/m3gs68
txa.ark:/81423/m33p76
txa.ark:/81423/m37633
txa.ark:/81423/m35p86
dul1.ark:/13960/t1jh9s44v
dul1.ark:/13960/t3wt4wz94
mdp.39015011399360
txa.ark:/81423/m30916
txa.ark:/81423/m3v040
mdp.39015008476239
mdp.39015020142413
osu.32435062842273
loc.ark:/13960/t4sj23nlj
txa.ark:/81423/m3zs70
osu.32435051449767
txa.ark:/81423/m3g62w
mdl.reflections.000461
mdp.39015015668075
txa.ark:/81423/m37h02
txa.ark:/81423/m3333t
emu.010000427626
mdp.39015008354709
uiug.30112116638161
```

```
emu.010002718580
txa.ark:/81423/m37p7f
txa.ark:/81423/m31637
txa.ark:/81423/m3wk8p
txa.ark:/81423/m3m056
mdp.39015021088474
loc.rbc/rbnawsa.n2748
emu.010002426988
txa.ark:/81423/m3nw5g
txa.ark:/81423/m3bp60
mdp.39015002955402
txa.ark:/81423/m3fw5m
txa.ark:/81423/m3m61r
emu.010002701790
emu.010000663182
emu.010000663181
emu.010001334218
txa.ark:/81423/m3233h
emu.010002701864
emu.010002701865
emu.010002701866
emu.010002701867
emu.010002701860
emu.010002701861
emu.010002701862
emu.010002701863
emu.000011035783
mdp.39015033941496
kero.htm
mdp.39015005293009
txa.ark:/81423/m3ps64
txa.ark:/81423/m3633r
txa.ark:/81423/m38s6d
mdp.39015014555836
uiug.30112116638146
mdp.39015040826888
txa.ark:/81423/m32625
mdp.39015041019350
mdp.39015015826400
mdp.39015001183980
txa.ark:/81423/m3p032
txa.ark:/81423/m3hs6k
txa.ark:/81423/m39p8f
txa.ark:/81423/m30d0v
txa.ark:/81423/m3r644
txa.ark:/81423/m3w34b
txa.ark:/81423/m3n33s
txa.ark:/81423/m3dp70
```

```
osu.32435069053379
mdp.39015007227047
mdp.35112104920675
txa.ark:/81423/m31336
uc2.ark:/13960
txa.ark:/81423/m3j32g
txa.ark:/81423/m3rp8g
ufl1.ark:/13960/t3b00qb6s
txa.ark:/81423/m3qw6g
txa.ark:/81423/m3tp7q
mdp.39015021021566
mdp.39015018601248
uiuo.ark:/13960/t7rn9g73w
txa.ark:/81423/m3c62z
mdp.39015041741011
uiuo.ark:/13960/t14n5gz3j
emu.010001341370
txa.ark:/81423/m3fc9z
txa.ark:/81423/m3np75
txa.ark:/81423/m3d33m
txa.ark:/81423/m35s7v
txa.ark:/81423/m3ch09
txa.ark:/81423/m3qd0c
txa.ark:/81423/m3062j
emu.010001295799
txa.ark:/81423/m3td09
mdp.39015014595527
txa.ark:/81423/m3wg8n
mdp.39015015746988
mdp.39015010881103
txa.ark:/81423/m3wp7b
mdp.39015002927039
txa.ark:/81423/m3462s
mdp.39015030236254
mdp.39015038884402
mdp.39015021963197
uiug.30112121934746
txa.ark:/81423/m3hc9k
txa.ark:/81423/m3h626
mdp.39015040826813
mdp.39015011020131
uiuo.ark:/13960/t25b6bb7m
emu.010002701731
emu.010002701732
txa.ark:/81423/m33w6w
mdp.39015039868727
uiug.30112116641256
txa.ark:/81423/m3pd2t
```

```
pt.id:mdp.39015051174947
emu.010001334535
emu.10002361311
oyp.33433066661673
txa.ark:/81423/m3fp6x
mdp.39015002906900
txa.ark:/81423/m3503d
uiuo.ark:/13960/t79s80910
txa.ark:/81423/m3rw5d
mdp.39015014534294
uiuo.ark:/13960/t7np89d52
uiuo.ark:/13960/t3908c06w
txa.ark:/81423/m3b919
mdp.39015010880477
dul1.ark:/13960/t45r1558k
txa.ark:/81423/m3x639
mdp.39015005860161l
txa.ark:/81423/m3xg9b
txa.ark:/81423/m3jh0v
uiug.30112121933672
txa.ark:/81423/m34w7k
emu.010002701778
ucl.b4164139:view
txa.ark:/81423/m3ms7w
emu.010001278509
mdp.39015033870489
txa.ark:/81423/m3k34j
mdp.39015038815679
txa.ark:/81423/m3cw4m
txa.ark:/81423/m3dc9n
mdp.39015025352868
emu.010002701463
txa.ark:/81423/m3dk9q
mdp.39015022480159
mdp.39015029148338
txa.ark:/81423/m36d13
emu.010002718789
emu.010002718788
txa.ark:/81423/m3bk8q
mdp.39015025999403
uiuo.ark:/13960/t6n07fp67
txa.ark:/81423/m3p61c
mdp.39015019954422
emu.010000427766
uiug.30112121929597
osu.32435063978308
dul1.ark:/13960/t41s30z6q
txa.ark:/81423/m3tk82
```

```
mdp.39015008679634
emu.010002701775
emu.010002701776
emu.010002701777
emu.010002701779
mdp.39015036922766
uiug.30112120239493
mdp.39015015667531
osu.32435029888989
mdp.39015087418763
ufl1.ark:/13960/t1sf3j445
txa.ark:/81423/m3f63z
txa.ark:/81423/m38w4p
txa.ark:/81423/m3791c
txa.ark:/81423/m3vp71
emu.000011713904
gri.ark:/13960/t2v471r7s
txa.ark:/81423/m3ns76
txa.ark:/81423/m3c91m
txa.ark:/81423/m3305j
uiuo.ark:/13960/t6zw7rm79
mdp.39015028059247
txa.ark:/81423/m3v63p
txa.ark:/81423/m31s7m
mdp.39015040826912
emu.10002335078
txa.ark:/81423/m3405v
mdp.39015079105444-1453734647
txa.ark:/81423/m3sp95
txa.ark:/81423/m38k7q
txa.ark:/81423/m3d929
txa.ark:/81423/m3h927
mdp.39013999346109
bc.ark:/13960/t9c59v05k
mdl.reflections.mhs03165
mdp.35128001457157
mdp.39015013031235
mdp.39013999315249
txa.ark:/81423/m3k04h
txa.ark:/81423/m3xp68
emu.010002701857
mdp.39015022480076
txa.ark:/81423/m38909
txa.ark:/81423/m3k62t
mdp.39015004530310
txa.ark:/81423/m3t921
mdp.39015026300627
emu.010002718809
```

```
txa.ark:/81423/m3s04c
mdp.39015012642164
mdp.39015087085455
mdp.39015009101885
txa.ark:/81423/m3g905
txa.ark:/81423/m3p35v
emu.010002701475
txa.ark:/81423/m3xk80
txa.ark:/81423/m3j05k
mdp.39015034573736
txa.ark:/81423/m3704d
coo.31924032664173:view1up:seq
txa.ark:/81423/m3pw5s
emu.010001341268
txa.ark:/81423/m3dg89
txa.ark:/81423/m3qs52
mdp.39015021307601
txa.ark:/81423/m3105x
emu.010002701696
txa.ark:/81423/m3b05q
uiuo.ark:/13960/t0cw0t79k
mdp.39015037277178
txa.ark:/81423/m3gg99
txa.ark:/81423/m3dw6p
mdp.39015015363842
txa.ark:/81423/m32s7x
mdp.39015020205269
txa.ark:/81423/m3mm0h
umn.31951002275060y
emu.010002701473
emu.010002701472
emu.010002701471
emu.010002701470
mdp.39015019354250
emu.010002701474
txa.ark:/81423/m3vh11
mdp.39015022779014
coo.31924105929362
txa.ark:/81423/m3m33g
txa.ark:/81423/m3192w
txa.ark:/81423/m3w92n
txa.ark:/81423/m3qp85
txa.ark:/81423/m3m925
txa.ark:/81423/m32w4t
uiuo.ark:/13960/t3b05fw90
txa.ark:/81423/m3bs7d
uiuo.ark:/13960/t3xt1xs3k
txa.ark:/81423/m3534t
txa.ark:/81423/m33s6v
emu.010000666051
mdp.39014003016989
mdp.39015009984819
mdp.39015031567889
mdp.39015021737690
emu.010002634732
txa.ark:/81423/m3sg93
emu.010002634735
emu.010002701451
emu.010002701450
emu.010002701453
emu.010002701452
emu.010002701454
emu.010002701457
emu.010002701456
emu.010002701459
emu.010002701458
mdp.39015014636669
txa.ark:/81423/m38p84
txa.ark:/81423/m3xw6b
txa.ark:/81423/m3q04r
emu.010002426117
txa.ark:/81423/m3x91k
emu.010000666530
txa.ark:/81423/m3cs6b
mdp.39015001793531
mdp.39015027876492
txa.ark:/81423/m35623
txa.ark:/81423/m3nd0r
txa.ark:/81423/m3005m
mdp.39015016862750
mdp.39015019365058
mdp.39015039331312.pdf
txa.ark:/81423/m31c97
txa.ark:/81423/m31k99
txa.ark:/81423/m3ck7n
mdp.39015026623861
txa.ark:/81423/m37k95
txa.ark:/81423/m34s7j
txa.ark:/81423/m3ks7k
uiuo.ark:/13960/t32296x8s
txa.ark:/81423/m3f92m
mdp.39015019906489
mdp.39015015345260
txa.ark:/81423/m3h33j
txa.ark:/81423/m31w5w
mdp.39015005257889
```

```
emu.000011066932
txa.ark:/81423/m3mp6g
uiuo.ark:/13960/t40s62x3g
txa.ark:/81423/m3n044
txa.ark:/81423/m37w5r
uiuo.ark:/13960/t2d85jq36
txa.ark:/81423/m34c95
mdp.39015019057242
emu.000011713920
dul1.ark:/13960/t4pk6mn9j
mdp.39015013022150
uiug.30112121934795
ucbk.ark:/28722/h2g13z
dul1.ark:/13960/t4sj7pq59
txa.ark:/81423/m36p9w
txa.ark:/81423/m3jw5j
emu.010000427432
uiug.30112120239246
wu.890662919985
mdp.39015026793987
txa.ark:/81423/m3n913
emu.10002350260
txa.ark:/81423/m3zg9n
txa.ark:/81423/m3gd1k
mdp.39015010742172
txa.ark:/81423/m3rw41
mdp.39015016639877
uiuo.ark:/13960/t8wb1gg7b
txa.ark:/81423/m3tc9c
txa.ark:/81423/m3bd0z
txa.ark:/81423/m3jc9w
txa.ark:/81423/m3j915
emu.010000427575
txa.ark:/81423/m3n62f
txa.ark:/81423/m3th0b
txa.ark:/81423/m3mw6j
dul1.ark:/13960/t1jh9s023
mdp.39015002258310
txa.ark:/81423/m3191h
txa.ark:/81423/m3rk8f
txa.ark:/81423/m3kk8w
emu.010002634320
emu.010002634321
txa.ark:/81423/m3803b
txa.ark:/81423/m3zh0k
mdp.39099999999999
txa.ark:/81423/m30s5j
txa.ark:/81423/m36344
```

```
mdp.39015011282608
mdp.39015033568489
txa.ark:/81423/m3qk9h
emu.010002701455
txa.ark:/81423/m3z63m
uiug.30112116638179
mdp.39015028555665
mdp.39015009573711
txa.ark:/81423/m3w617
mdp.39015039589869
uiug.30112121934720
txa.ark:/81423/m3nw43
txa.ark:/81423/m30g8k
txa.ark:/81423/m3290f
txa.ark:/81423/m3q622
mdp.39015000421480
mdp.39015008501689
txa.ark:/81423/m3g638
txa.ark:/81423/m30348
mdp.39015007227336
mdp.39015021119592
uiug.30112066813590
txa.ark:/81423/m3k91g
txa.ark:/81423/m3ps5r
emu.010000427431
emu.000011207069
emu.010002426856
txa.ark:/81423/m38s51
coo.31924012245027
uiug.30112121936592
uiug.30112116638153
txa.ark:/81423/m3b32m
emu.010002701774
prev.89.6.861.22100
txa.ark:/81423/m3ww5n
txa.ark:/81423/m39d0n
mdp.39015006749819
txa.ark:/81423/m3432r
coo.31924014779577
emu.010002701789
emu.010002701788
emu.010002701781
emu.010002701780
emu.010002701783
emu.010002701782
emu.010002701785
emu.010002701784
emu.010002701787
emu.010002701786
mdp.39015000640618
txa.ark:/81423/m30d17
uiuo.ark:/13960/t08w9nf5r
dul1.ark:/13960/t7fr63h31
mdp.39015002121419
mdp.39015024388079
mdp.39015010693912
emu.010002634673
mdp.39013999345819
txa.ark:/81423/m3qw7v
mdp.39015012554393
mdp.39015087085307
emu.010002718557
txa.ark:/81423/m34p64
txa.ark:/81423/m3fh0x
emu.010000666234
mdp.39015019202301
aja.1000680106
emu.010002701851
emu.010002701853
emu.010002701852
emu.010002701859
emu.010002701858
txa.ark:/81423/m3dp6m
txa.ark:/81423/m3zp7z
uiuo.ark:/13960/t6q01f796
mdp.39015042484504
txa.ark:/81423/m3g34m
txa.ark:/81423/m3132t
mdp.39015008208350
uiuo.ark:/13960/t6c316m95
dul1.ark:/13960/t6f255k6n
txa.ark:/81423/m3wg91
mdp.39015016748108
txa.ark:/81423/m3tp6b
txa.ark:/81423/m3t32z
txa.ark:/81423/m3sh01
txa.ark:/81423/m3992c
mdp.39015008725700
txa.ark:/81423/m3kg97
txa.ark:/81423/m32k7v
mdp.39013998980019
gri.ark:/13960/t4xh5x606
mdp.39015026796758
txa.ark:/81423/m3v33n
txa.ark:/81423/m33346
txa.ark:/81423/m35s6g
txa.ark:/81423/m3pp8v
mdp.39015021111060
txa.ark:/81423/m3c34p
emu.000011713905
emu.010002634845
emu.010002634847
mdp.39015031696886
txa.ark:/81423/m3q33d
uiuo.ark:/13960/t79s80q6s
mdp.39015040826805
mdp.39015006107596
mdp.39015035385338
txa.ark:/81423/m33d0s
txa.ark:/81423/m3kw5v
txa.ark:/81423/m34w66
mdp.39015003763060
txa.ark:/81423/m3h058
piee.1973.0244
mdp.39015019099178
uiuo.ark:/13960/t4fn7dc8w
mdp.39015020571595
inyp.33
mdp.39015022480142
txa.ark:/81423/m3d628
txa.ark:/81423/m3j90s
mdp.39015023756219
mdp.39015037363648
txa.ark:/81423/m3b92p
txa.ark:/81423/m3933p
txa.ark:/81423/m33k9x
txa.ark:/81423/m34k97
txa.ark:/81423/m39p72
mdp.39015028536665
txa.ark:/81423/m3qp7s
mdp.39015021307692
txa.ark:/81423/m37s63
txa.ark:/81423/m3gp80
txa.ark:/81423/m3sw5q
mdp.39015036716143
txa.ark:/81423/m3cd1n
cn.11/63
mdp.39015018046774
txa.ark:/81423/m36k9v
emu.000011713918
emu.000011713917
emu.000011024626
mdp.39015004603000
mdp.39015014125887
mdp.39015028738683
txa.ark:/81423/m3cp82
mdp.39015022449550
txa.ark:/81423/m37031
txa.ark:/81423/m3492t
txa.ark:/81423/m36043
dul1.ark:/13960/t9z09c80h
txa.ark:/81423/m39g9r
txa.ark:/81423/m3z90h
txa.ark:/81423/m36w6t
txa.ark:/81423/m3rk72
txa.ark:/81423/m38g9f
txa.ark:/81423/m3wd19
txa.ark:/81423/m39k8d
txa.ark:/81423/m3vk70
txa.ark:/81423/m3sp8s
mdp.39015013261154
uio.ark:/13960/t9v11xq3s
mdp.39015033591358
txa.ark:/81423/m3pk96
mdp.39015020233865
coo1.ark
penn.ark:/81431/p33t06
emu.010001341372
emu.010001341375
emu.010001341374
txa.ark:/81423/m39s5b
txa.ark:/81423/m38h0c
txa.ark:/81423/m3z34z
uiuo.ark:/13960/t73v5t15z
mdp.39015006758232
osu.32435064981384
txa.ark:/81423/m31g98
uiuo.ark:/13960/t2s52wr7z
mdp.39015039438232
mdp.39015012061035
txa.ark:/81423/m3z91w
penn.ark:/81431/p3k93h
uiuo.ark:/13960/t1hj2rm1n
loc.ark:/13960
mdp.39015009172050
txa.ark:/81423/m35g9h
mdp.39015021015519
txa.ark:/81423/m3k61f
txa.ark:/81423/m33613
txa.ark:/81423/m37w64
txa.ark:/81423/m3404g
txa.ark:/81423/m3392h
```

```
mdp.39015006729019
txa.ark:/81423/m3dd0k
txa.ark:/81423/m3p34g
mdp.39015007031704
txa.ark:/81423/m3vs72
mdp.39015038815687
emu.10002335129
txa.ark:/81423/m36611
txa.ark:/81423/m36s5d
mdp.39015014664539
emu.010000667184
uiuo.ark:/13960/t67430p8z
mdp.39015077310806
umn.31951002092739e
txa.ark:/81423/m3s91b
txa.ark:/81423/m3h34x
uiuo.ark:/13960/t3hx7jz69
uiuo.ark:/13960/t4rk0p40n
txa.ark:/81423/m3vw4z
txa.ark:/81423/m3ws60
gri.ark:/13960/t2n64vd8d
mdp.39015005797199
txa.ark:/81423/m36g8f
txa.ark:/81423/m3n03r
txa.ark:/81423/m35w4r
txa.ark:/81423/m3vd0m
mdp.39015021128528
gri.ark:/13960/t2f82p41f
txa.ark:/81423/m3204v
txa.ark:/81423/m3j046
emu.010002701460
emu.010002701461
emu.010002701462
emu.010002701464
emu.010002701466
txa.ark:/81423/m3md0f
txa.ark:/81423/m3xc99
txa.ark:/81423/m3xk9c
mdp.39015039589885
txa.ark:/81423/m3t052
txa.ark:/81423/m3x92z
mdp.39015002952482
mdp.39015004897958
mdp.39015037759001
txa.ark:/81423/m3033w
mdl.reflections.mhs7508-all
txa.ark:/81423/m3hg9m
txa.ark:/81423/m30p8n
```

```
gri.ark:/13960/t4pk6q454
mdp.39015016639885
emu.010002701693
emu.010002701697
emu.010002701695
emu.010002701694
emu.010002701698
txa.ark:/81423/m3pg8s
mdp.39015006367331
db.aspx
txa.ark:/81423/m32h0h
emu.010000666040
mdp.39015033265888
uiuo.ark:/13960/t05x8jz39
txa.ark:/81423/m3gg8x
dul1.ark:/13960/t1gj5qr9q
ucbk.ark:/28722/h2043n
gri.ark:/13960/t1fj8s658
txa.ark:/81423/m3zd0j
txa.ark:/81423/m3s62p
txa.ark:/81423/m3g04k
txa.ark:/81423/m3vh0n
mdp.39015000590060
txa.ark:/81423/m3cg9c
txa.ark:/81423/m3qg9g
mdp.39015008454665
txa.ark:/81423/m3m323
mdp.39013999315729
ufl1.ark:/13960/t9t16k94b
mdp.39015008524004
txa.ark:/81423/m35d1s
mdp.39015013523652
txa.ark:/81423/m3hd0h
mdp.39015021102796
txa.ark:/81423/m32w56
mdp.39015005174803
bc.ark:/13960/t52g2jf57
txa.ark:/81423/m3js8n
txa.ark:/81423/m3863d
mdp.39015023552600
txa.ark:/81423/m3905d
txa.ark:/81423/m3bk7b
txa.ark:/81423/m3z03j
txa.ark:/81423/m3f917
uiuo.ark:/13960/t25b6dp8d
txa.ark:/81423/m3bw5p
mdp.39015024038179
mdp.39015013028785
```

```
uiuo.ark:/13960/t0dv7rx5d
txa.ark:/81423/m3sk7c
inu.skin
txa.ark:/81423/m3r03p
emu.000011015865
txa.ark:/81423/m37k8s
txa.ark:/81423/m3ss5p
txa.ark:/81423/m3v91z
txa.ark:/81423/m3ts6c
txa.ark:/81423/m3jw45
txa.ark:/81423/m3ks66
txa.ark:/81423/m3rs6r
emu.010002702177
txa.ark:/81423/m3561q
osu.32435083668939
txa.ark:/81423/m3hp7x
mdp.39015027331308
txa.ark:/81423/m32d0g
txa.ark:/81423/m3bh00
mdp.39015028430646
mdp.39015021114791
txa.ark:/81423/m3263j
mdp.39015034750060
```

```
In [194]: #How many items had event triggers recorded in order to appear in the analytics, but h
          zeroes = counts_ids[counts_ids['count'] < 1]
          zeroes
```

```
Out[194]:                               id  \
          3997326  aeu.ark:/13960/t0001qn0c
          3997327  aeu.ark:/13960/t00z7fz4z
          3997329  aeu.ark:/13960/t00z7qh5x
          3997333  aeu.ark:/13960/t01z4v16z
          3997337  aeu.ark:/13960/t02z2864r
          3997338  aeu.ark:/13960/t02z29g19
          3997344  aeu.ark:/13960/t03x91k5w
          3997363  aeu.ark:/13960/t08w4f73q
          3997369  aeu.ark:/13960/t09w1fq2p
          3997371  aeu.ark:/13960/t09w1mw9c
          3997380  aeu.ark:/13960/t0dv1p82b
          3997384  aeu.ark:/13960/t0dv2mn8q
          3997391  aeu.ark:/13960/t0gt6g78g
          3997393  aeu.ark:/13960/t0ht30s5k
          3997397  aeu.ark:/13960/t0jt0pk1d
          3997398  aeu.ark:/13960/t0jt0sf0c
          3997399  aeu.ark:/13960/t0ks8dd68
          3997403  aeu.ark:/13960/t0ms4j92f
```

```
3997404  aeu.ark:/13960/t0ns16n2z
3997405  aeu.ark:/13960/t0ns1hz74
3997410  aeu.ark:/13960/t0pr89d35
3997412  aeu.ark:/13960/t0pr8m23d
3997441  aeu.ark:/13960/t0wq17j9w
3997442  aeu.ark:/13960/t0xp7ph4n
3997443  aeu.ark:/13960/t0xp7st00
3997452  aeu.ark:/13960/t0zp5bd94
3997457  aeu.ark:/13960/t10p1xj72
3997458  aeu.ark:/13960/t10p22s2t
3997495  aeu.ark:/13960/t1cj9j18b
3997496  aeu.ark:/13960/t1cj9jv3p
...                            ...
5561818       yale.39002088371829
5561819       yale.39002088374187
5561820       yale.39002088374484
5561821       yale.39002088375341
5561822       yale.39002088441077
5561823       yale.39002088441689
5561824       yale.39002088442000
5561825       yale.39002088442679
5561826       yale.39002088442687
5561827       yale.39002088445391
5561828       yale.39002088450003
5561829       yale.39002088545463
5561830       yale.39002088548251
5561831       yale.39002088549127
5561832       yale.39002088670220
5561833       yale.39002088672432
5561834       yale.39002088678033
5561835       yale.39002088678892
5561836       yale.39002088679015
5561837       yale.39002089373949
5561838       yale.39002089541990
5561839       yale.39002089549894
5561840       yul.11365223_000_00
5561841       yul.11729383_000_00
5561842       yul.11816619_000_00
5561843       yul.12221406_000_00
5561844       yul.12225202_000_00
5561845       yul.12240836_009_00
5561846       yul.12266189_004_00
5561847       yul.12557260_000_00

                                                     title access    date  \
3997326                       The bridge by Mark Somers.   deny  1929.0
3997327                    Sonnet to E. W. [N.F. Davin].  allow  1881.0
3997329       A lady's life on a ranche [Moira O'Neill].  allow  1898.0
```

```
3997333  Elective franchise, or, Why Reformed Presbyter...  allow  1878.0
3997337  Le voyageur françois, ou, La connoissance de l...  allow  1795.0
3997338  Géographie moderne précédée d'un petit traité ...  allow  1772.0
3997344  Letters from North America written during a to...  allow  1824.0
3997363  Causes of ministerial sadness a sermon preache...  allow  1866.0
3997369  Out on the Pampas, or, The young settlers by G...  allow  1899.0
3997371   Nene Karighyoston tsinihorighhoten ne Saint John  allow  1818.0
3997380  The equality of Greek with French and German (...  allow  1899.0
3997384  An Act respecting pilotage, assented to 23rd M...  allow  1877.0
3997391  Annual address, delivered by the Rev. John M. ...  allow  1851.0
3997393  The Laurentian and Huronian systems in the reg...  allow  1892.0
3997397  The author, or, Sketches from life by W.F. Dea...  allow  1866.0
3997398  Voters' list of the municipality of the townsh...  allow  1881.0
3997399  The effect of ferric salts on the rate of oxid...  allow  1908.0
3997403  Les Ursulines de Québec, depuis leur établisse...  allow  1863.0
3997404  The constable's guide a sketch of the office o...  allow  1861.0
3997405  Memoir upon the estates which the Jesuits poss...  allow  1845.0
3997410  Mexico, Texas, Canada message from the preside...  allow  1838.0
3997412  Starke's pocket almanac and general register f...  allow  1866.0
3997441  The community survey, a basis for social actio...  allow  1919.0
3997442  The history of the Church of England in the co...  allow  1845.0
3997443  Annual register of officers and members of the...  allow  1896.0
3997452  In the van, or, The builders by Price-Brown (E...  allow  1906.0
3997457  The Political progress of Britain, or, An impa...  allow  1794.0
3997458  A full history of the wonderful career of Mood...  allow  1876.0
3997495  Thèses de mathématiques et de physique, qui se...  allow  1792.0
3997496  Mémoire sur la question des corvées dans la se...  allow  1873.0
...                                                   ...    ...     ...
5561818                           The geography of Europe.  allow  1918.0
5561819  Psalms, in metre, selected from the Psalms of ...  allow  1843.0
5561820  Notes on the Psalms, chiefly explanatory of th...  allow  1869.0
5561821  A harmony of the gospels for historical study ...  allow  1902.0
5561822  The Book of books and its wonderful story : a ...  allow  1922.0
5561823  The people's Bible; discourses upon Holy Scrip...  allow  1895.0
5561824  The people's Bible; discourses upon Holy Scrip...  allow  1895.0
5561825  Ad fidem; or, Parish evidences of the Bible / ...  allow  1871.0
5561826  Ad fidem; or, Parish evidences of the Bible / ...  allow  1871.0
5561827  Mosaics of Bible history; the Bible record wit...  allow  1883.0
5561828  An argument to prove the truth of the Christia...  allow  1834.0
5561829  The Victoria history of the county of Leiceste...   deny  9999.0
5561830                                La piedad del agua.  allow  1922.0
5561831  Los tres primeros historiadores de la isla de ...  allow  1877.0
5561832     ... The Gallery of portraits: with memoirs ...  allow  1837.0
5561833  The golden age of engraving; a specialist's st...  allow  1910.0
5561834     Man's place in the kosmos ... By S.A. Merrill.  allow  1906.0
5561835  The life of Jesus the Christ by Henry Ward Bee...  allow  1891.0
5561836  American Presbyterianism : a sermon, delivered...  allow  1854.0
5561837  Collections of the Worcester Society of Antiqu...  allow  1899.0
```

```
5561838  Vital record of Rhode Island : 1636-1850 : fir...   deny  9999.0
5561839  The duty of a canonical adherence to the ritua...  allow  1818.0
5561840                                  Winsted directory   deny  1927.0
5561841    The green bay tree a novel, by Louis Bromfield.   deny  1926.0
5561842  Dona Marina por el dr. Gustavo A. Rodriguez ...   deny  1935.0
5561843                   Soviet science by J.G. Crowther.   deny  1936.0
5561844  Statistika evreiskago naseleniia raspredi...  allow  1909.0
5561845  Sussex archaeological collections relating to ...   deny  9999.0
5561846  Prace Towarzystwa naukowego warszawskiego III...  allow  1913.0
5561847  Report of the chief engineer, October 1870 [Ja...  allow  1871.0

                       oclc  Unnamed: 0  count
3997326           861778360     5129046    0.0
3997327           716107670     5130961    0.0
3997329           719178160     5130942    0.0
3997333           716961216     5133445    0.0
3997337           875529021     5130941    0.0
3997338           862023747     5133444    0.0
3997344           719993021     5143277    0.0
3997363           867972702     5129045    0.0
3997369           867969327     5129044    0.0
3997371           861562623     5130940    0.0
3997380           716114277     5133443    0.0
3997384           768326489     5130939    0.0
3997391           768321462     5130938    0.0
3997393           716130996     5130937    0.0
3997397           867973723     5130936    0.0
3997398           861481202     5130935    0.0
3997399           861779706     5130934    0.0
3997403           726101156     5143276    0.0
3997404           716911770     5130933    0.0
3997405           719993647     5133442    0.0
3997410           716922024     5130932    0.0
3997412           717071502     5130931    0.0
3997441           861574437     5130930    0.0
3997442           719955011     5130929    0.0
3997443           719998329     5130928    0.0
3997452           679948599     5133441    0.0
3997457           768321183     5133440    0.0
3997458           867971236     5143275    0.0
3997495           862032604     5130943    0.0
3997496           862035382     5143274    0.0
...                     ...         ...    ...
5561818              682772     4682496    0.0
5561819  38735720,684319935     4742338    0.0
5561820   683671756,7471610     4674509    0.0
5561821   3391145,684260780     4682502    0.0
5561822             2281977     4701040    0.0
```

```
5561823  47646483,684886439     4726012    0.0
5561824  47646483,684886439     4711490    0.0
5561825   5867852,684167729     4682501    0.0
5561826   5984270,684168223     4701039    0.0
5561827   3154352,684517262     4733231    0.0
5561828  15086237,684731231     4674508    0.0
5561829   2098674,686236342     4711483    0.0
5561830  54251735,687623441     4682500    0.0
5561831   1857715,688056354     4674507    0.0
5561832   1930519,687696176     4686299    0.0
5561833   2994055,687217248     4726027    0.0
5561834  14121287,684886705     4726026    0.0
5561835             3308355     4674506    0.0
5561836  11487210,684347128     4701038    0.0
5561837  10840331,686690536     4682499    0.0
5561838   1358069,686968339     4743595    0.0
5561839  44450650,684487372     4742337    0.0
5561840                 NaN     4733233    0.0
5561841           890513898     4682498    0.0
5561842           890514180     4701037    0.0
5561843           907971991     4682497    0.0
5561844           915042914     4682495    0.0
5561845           923597726     4711489    0.0
5561846           915042971     4711488    0.0
5561847                 NaN     4711487    0.0

[1350284 rows x 7 columns]
```

```
In [4]: full = pd.read_csv('./csv/all_id_title_access_date_oclc.csv')

        allow = full[full.access == 'allow']
        deny = full[full.access == 'deny']

        print ("There are %s total volumes that have triggered analytics events in the collected
        print ("There are %s total open volumes that have triggered analytics events in the coll
        print ("There are %s total limited view volumes that have triggered analytics events in
```

```
There are 5561848 total volumes that have triggered analytics events in the collected data
There are 3812502 total open volumes that have triggered analytics events in the collected data
There are 1749346 total limited view volumes that have triggered analytics events in the collect
```

1.6 Top title analysis

```
In [187]: '''This merges the metadata extracted from the Hathifiles with the top counts from the
          and spits out the list of the most viewed items in HathiTrust
          But this could easily be tweaked to show top NYPL items, top items that were denied ac
          top items published in a given country, etc. '''
          counts_ids = full.merge(d_counts, on='id', suffixes=['_full','_counts'])
```

```
In [195]: #Top 25 titles in Hathi
          counts_ids.sort_values('count', ascending=False).head(25)
```

```
Out[195]:                          id  \
          105427   mdp.39015054061430
          75199    mdp.39015011274175
          65550    mdp.39015004111095
          172780     pst.000057937434
          61310    mdp.39015000804453
          111279   mdp.39015064340733
          61058    mdp.39015000566789
          71062    mdp.39015008158415
          189012   uc1.32106007458745
          99427    mdp.39015038069475
          174170  pur1.32754077064610
          182455          uc1.$b99721
          78781    mdp.39015014103017
          221532  uiug.30112101024682
          62432    mdp.39015002033903
          69360    mdp.39015006749868
          71061    mdp.39015008158407
          103043   mdp.39015048226941
          118999   mdp.39015071886035
          1497          chi.087013173
          238087       wu.89059402255
          60758    mdp.39015000379902
          200351         uc1.b4164139
          238089       wu.89059402289
          45970    inu.30000007109121

                                                            title access    date  \
          105427                        Quicksand, by Nella Larsen.  allow  1928.0
          75199   The surnames of Scotland, their origin meaning...  allow  1962.0
          65550                                   Godey's magazine.  allow  1850.0
          172780          The human figure / by John H. Vanderpoel.  allow  1907.0
          61310     Perfume and flavor materials of natural origin.  allow  1960.0
          111279  Solid mensuration, by Willis F. Kern and James...  allow  1934.0
          61058   America is in the heart, a personal history, b...  allow  1946.0
          71062   Quintus Curtius [History of Alexander] with an...  allow  1946.0
          189012  History of wages in the United States from Col...  allow  1934.0
          99427   Return to life through contrology, by Joseph H...  allow  1960.0
          174170  Investigation of Korean-American relations : R...  allow  1978.0
          182455  The five laws of library science, by S. R. Ran...  allow  1931.0
          78781                        The book of a hundred hands.  allow  1920.0
          221532                      A short guide to New Zealand.  allow  1943.0
          62432        Kinematics and dynamics of plane mechanisms.  allow  1962.0
          69360   Modern California houses; case study houses, 1...  allow  1962.0
          71061   Quintus Curtius [History of Alexander] with an...  allow  1946.0
```

```
103043               The lesson of Japanese architecture.  allow  1954.0
118999                                   [Publications]  allow  9999.0
1497    Consumption of the lungs and kindred diseases,...  allow  1914.0
238087  Roster of the Confederate soldiers of Georgia,...  allow  9999.0
60758   Propaganda technique in the World War [by] Har...  allow  1938.0
200351  Circuit analysis of A-C power systems; symmetr...  allow  1950.0
238089  Roster of the Confederate soldiers of Georgia,...  allow  9999.0
45970   Pennsylvania German pioneers; a publication of...  allow  1934.0

                    oclc  Unnamed: 0     count
105427           7332881           0  101702.0
75199            1724215           1   69754.0
65550            2133694           2   55418.0
172780           3095972           3   48835.0
61310            1493297           4   48363.0
111279            823935           5   44557.0
61058             326807           6   41721.0
71062             685637           7   38700.0
189012           2794726           8   38244.0
99427            3165474           9   31078.0
174170          34759005          10   30437.0
182455           1293631          14   30432.0
78781             227380          11   29935.0
221532            937704          12   29618.0
62432             562906          13   28314.0
69360            1349332          15   27706.0
71061             685637          16   26467.0
103043           1243958          17   26350.0
118999         426038752          18   25769.0
1497            36830491          19   24974.0
238087  1624676,27030216          20   24594.0
60758            9086269          21   24484.0
200351           1563693          22   24095.0
238089  1624676,27030216          23   22818.0
45970            1850127          24   22688.0
```

```
In [190]: #Top 25 limited view titles in Hathi
          counts_ids[counts_ids.access == 'deny'].sort_values('count', ascending=False).head(25)
```

```
Out[190]:                 id                                              title  \
          113655  mdp.39015066789838  Theogony ; and, Works and days / Hesiod ; tran...
          47544   inu.30000103012815                   Kasaita / na Maryam Kabir Mashi.
          48369   inu.30000124268446      Rufaida ko mufida? / na Hadiza Salisu Sharif.
          189432  uc1.32106012042997                               A treatise on money,
          102444  mdp.39015046422120                        Nectar in a sieve, a novel.
          75692   mdp.39015011482067  Württembergisches Adels- und Wappenbuch / im A...
          130536  mdp.39076006350719  The theory of spherical and ellipsoidal harmon...
          119201  mdp.39015072611786            The war of the worlds / by H. G. Wells.
```

```
203099         uc1.b4906221  The competent manager : a model for effective ...
114964  mdp.39015068290124                 The advanced theory of statistics
74097   mdp.39015010576356  Objects of daily use, with over 1800 figures f...
74095   mdp.39015010574575                         Catalogue of Alexandrian coins,
67439   mdp.39015005323111                                                Proust.
104877  mdp.39015052047589  American archival studies : readings in theory...
125309  mdp.39015079728443  Men at war : the best war stories of all time ...
79266   mdp.39015014559135  My experiences in the world war, by John J. Pe...
87022   mdp.39015023388500  The anatomy of the root-canals of the teeth of...
68718   mdp.39015006079035  Linear circuits. With the editorial assistance...
130383  mdp.39076005361576  The idea of reform; its impact on Christian th...
49529   inu.32000009618820                               A brighter sun, a novel.
63752   mdp.39015002699810  Aristotle dictionary / Edited by Thomas P. Kie...
60672   mdp.39015000143266  Coral gardens and their magic : a study of the...
118884  mdp.39015071754159                                    The Michigan daily.
72605   mdp.39015009106751  The mothers : a study of the origins of sentim...
80228   mdp.39015015725156  The advanced theory of statistics, by Maurice ...

       access    date       oclc  Unnamed: 0   count
113655   deny  2006.0   63122803         170  6450.0
47544    deny  9999.0   64193309         175  6395.0
48369    deny  9999.0  179404851         304  4715.0
189432   deny  1930.0     721781         460  3739.0
102444   deny  1954.0     733922         512  3557.0
75692    deny  1975.0    4832917         537  3477.0
130536   deny  1931.0    1379672         583  3309.0
119201   deny  1926.0   17861207         932  2602.0
203099   deny  1982.0    7740141         990  2525.0
114964   deny  9999.0     527103        1146  2346.0
74097    deny  1927.0    3553454        1253  2247.0
74095    deny  1933.0    6342337        1395  2131.0
67439    deny  1957.0     188645        1407  2120.0
104877   deny  2000.0   44391683        1417  2112.0
125309   deny  1942.0     319365        1499  2056.0
79266    deny  1931.0     394688        1527  2032.0
87022    deny  1925.0    5969802        1532  2029.0
68718    deny  1960.0     986383        1645  1964.0
130383   deny  1959.0    1210563        1681  1947.0
49529    deny  1952.0    1211314        1823  1864.0
63752    deny  1962.0    1388152        1890  1829.0
60672    deny  1935.0    6174779        2008  1775.0
118884   deny  1969.0    9651208        2266  1668.0
72605    deny  1927.0     530511        2277  1665.0
80228    deny  9999.0    6583484        2512  1580.0
```

```
In [191]: #Top 25 Hathi volumes scanned from NYPL collections
          counts_ids[counts_ids['id'].str.startswith('nyp') == True].sort_values('count', ascend
```

```
Out[191]:                       id                                                title  \
160119  nyp.33433076064025  Miranda / by Grace Livingston Hill Lutz ... ; ...
157449  nyp.33433069455859  Illustrated trade catalogue and price list : m...
155069  nyp.33433066397708  Illustrated catalogue of hand and power pumps,...
161184  nyp.33433081675450                                 Godey's magazine.
163079  nyp.33433081893293  L'Egypte a L'Exposition universelle de 1867 /...
149895  nyp.33433000335228  Glossarium ad scriptores mediae et infimae Lat...
162670  nyp.33433081844692  A standard history of Stark County, Ohio : an ...
150987  nyp.33433006773448                       Home needlework magazine ...
153382  nyp.33433023615366  Regimental colors of the German armies in the ...
158132  nyp.33433072182490  Art monograms and lettering, by J.M. Bergling,...
163637  nyp.33433082132030  História orgánica de las armas de infantería y...
149897  nyp.33433000335244  Glossarium ad scriptores mediae et infimae Lat...
153405  nyp.33433023758695  Report of the Committee of Secrecy on the Bank...
163263  nyp.33433081921573  A history of Jasper County, Missouri, and its ...
163669  nyp.33433082137914  Wife no. 19, or the story of a life in bondage...
152080  nyp.33433009488465  The law reports,. under the superintendence an...
150773  nyp.33433006349736  A specimen of printing types, and ornaments, c...
166927  nyp.33433090820188                            The Commercial vehicle.
161197  nyp.33433081675583                                 Godey's magazine.
155741  nyp.33433066642897                          American chess magazine.
166940  nyp.33433090821731                                      Power wagon.
166689  nyp.33433090781398              Cycle and automobile trade journal.
153751  nyp.33433037323635  Illustrated catalogue of Seth Thomas, New Have...
167810  nyp.33433112041938       A book of verses / by William Ernest Henley.
166592  nyp.33433090762398               The Paper mill and wood pulp news.

       access    date             oclc  Unnamed: 0   count
160119  allow  1915.0         17553920         181  6304.0
157449  allow  1897.0         64665705         293  4769.0
155069  allow  1903.0         39741465         327  4481.0
161184  allow  1831.0          2133694         329  4462.0
163079  allow  1867.0         37632857         332  4449.0
149895  allow  1736.0          8055999         354  4297.0
162670  allow  1916.0          6430855         359  4268.0
150987  allow  1912.0          9398894         441  3803.0
153382  allow  1911.0         14560353         457  3743.0
158132  allow  1912.0         11611832         521  3543.0
163637  allow  1859.0         36850981         590  3288.0
149897  allow  1736.0          8055999         660  3055.0
153405  allow  1832.0         11597232         729  2929.0
163263  allow  1912.0          2704614         773  2838.0
163669  allow  1875.0  2582100,8064193         798  2801.0
152080  allow  1884.0         53115156         802  2795.0
150773  allow  1828.0         38404282         842  2747.0
166927  allow  1917.0              NaN         853  2732.0
161197  allow  1850.0          2133694         879  2688.0
155741  allow  1899.0          3983478         881  2685.0
166940  allow  1913.0              NaN         921  2612.0
```

```
166689  allow  1904.0           NaN      935  2594.0
153751  allow  1878.0           NaN      998  2509.0
167810  allow  1888.0      13897970     1002  2508.0
166592  allow  1903.0       1369875     1047  2442.0
```

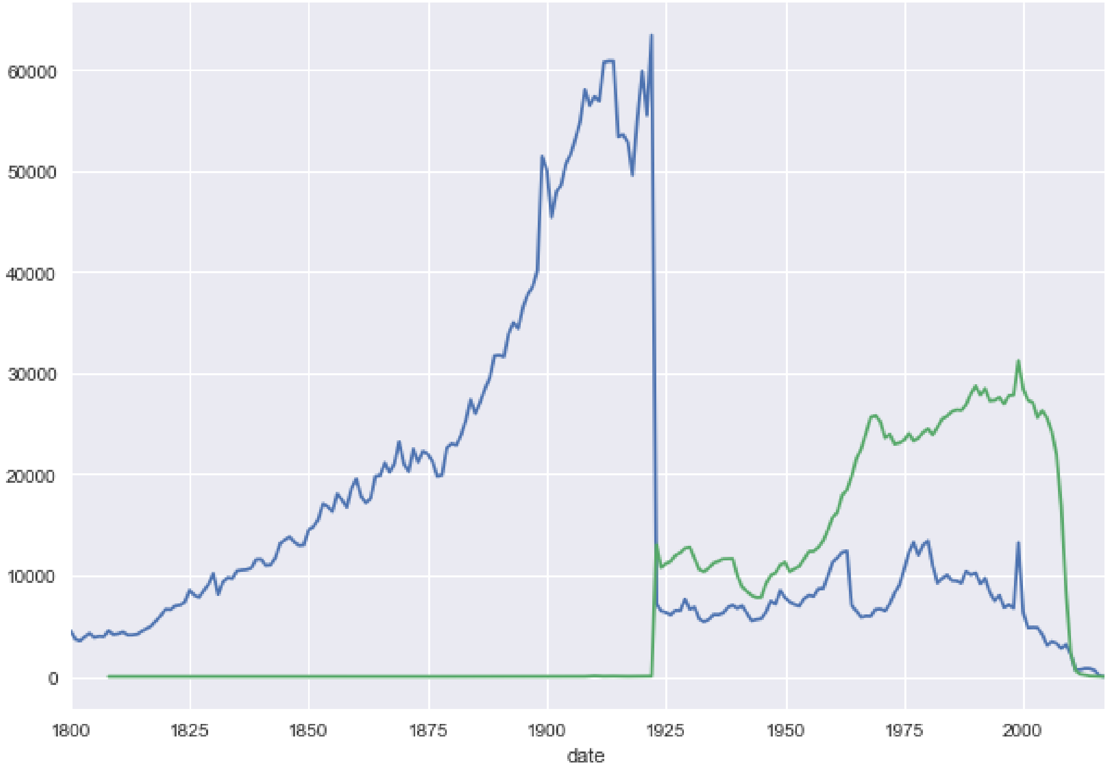
1.7 Publication Year Analysis

```
In [9]: all_years = full[(full.date > 1799) & (full.date < 2018)].groupby('date')['id'].count()
        allow_years = allow[(allow.date > 1799) & (allow.date < 2018)].groupby('date')['id'].cou
        deny_years = deny[(deny.date > 1799) & (deny.date < 2018)].groupby('date')['id'].count()
```

```
In [131]: #Plots publication date of volumes with analytics events, as full-view vs limited vie

          #all_years.plot()
          allow_years.plot()
          deny_years.plot(figsize=(10,7))
```

```
Out[131]: <matplotlib.axes._subplots.AxesSubplot at 0x1929b7410>
```



```
In [13]: #This grabs all dates from the postgres DB to some data analysis and historgrams, etc.

         all_dates = x = pd.read_sql_query("SELECT DISTINCT publication_date, count(publication_
                   from hathifiles GROUP BY publication_date ORDER BY publication_date ASC", co
         all_dates_a = x = pd.read_sql_query("SELECT DISTINCT publication_date, count(publicatio
                      from hathifiles WHERE access = 'allow' GROUP BY publication_date ORDER BY pu

In [14]: dates = all_dates[(all_dates.publication_date < 2021) & (all_dates.publication_date > 1
         dates_a = all_dates_a[(all_dates_a.publication_date < 2018) & (all_dates_a.publication_

In [16]: #dates.index = dates.publication_date
         dates_a.index = dates_a.publication_date
         #dates.loc[:,'accessed'] = all_years
         dates_a.loc[:,'accessed'] = allow_years
         dates_a

Out[16]:                   publication_date  count  accessed
         publication_date
         1800.0                      1800.0   6303      4552
         1801.0                      1801.0   5044      3696
         1802.0                      1802.0   4917      3535
         1803.0                      1803.0   5526      3918
         1804.0                      1804.0   6044      4276
         1805.0                      1805.0   5275      3859
         1806.0                      1806.0   5421      3981
         1807.0                      1807.0   5243      3938
         1808.0                      1808.0   6491      4547
         1809.0                      1809.0   5587      4131
         1810.0                      1810.0   5923      4220
         1811.0                      1811.0   6129      4416
         1812.0                      1812.0   5528      4089
         1813.0                      1813.0   5765      4095
         1814.0                      1814.0   6035      4168
         1815.0                      1815.0   6218      4493
         1816.0                      1816.0   6570      4731
         1817.0                      1817.0   7113      5003
         1818.0                      1818.0   7883      5533
         1819.0                      1819.0   9146      6049
         1820.0                      1820.0  10086      6656
         1821.0                      1821.0   9482      6617
         1822.0                      1822.0  10389      6997
         1823.0                      1823.0  10417      7071
         1824.0                      1824.0  10661      7339
         1825.0                      1825.0  12323      8537
         1826.0                      1826.0  11429      8037
         1827.0                      1827.0  11301      7834
         1828.0                      1828.0  12498      8474
         1829.0                      1829.0  13462      9134
         ...                            ...    ...       ...
         1988.0                      1988.0  18228     10406
         1989.0                      1989.0  17658     10043
         1990.0                      1990.0  17880     10240
```

```
1991.0                          1991.0  17124       9173
1992.0                          1992.0  17786       9708
1993.0                          1993.0  15630       8329
1994.0                          1994.0  14859       7484
1995.0                          1995.0  15755       8025
1996.0                          1996.0  13742       6825
1997.0                          1997.0  13434       7101
1998.0                          1998.0  12938       6770
1999.0                          1999.0  27529      13217
2000.0                          2000.0  11329       6323
2001.0                          2001.0   8303       4822
2002.0                          2002.0   8515       4860
2003.0                          2003.0   8464       4835
2004.0                          2004.0   7574       4167
2005.0                          2005.0   5513       3081
2006.0                          2006.0   5694       3465
2007.0                          2007.0   5868       3281
2008.0                          2008.0   4467       2776
2009.0                          2009.0   5389       3171
2010.0                          2010.0   3561       2148
2011.0                          2011.0   1379        644
2012.0                          2012.0   1148        677
2013.0                          2013.0   1129        806
2014.0                          2014.0    994        800
2015.0                          2015.0    716        590
2016.0                          2016.0     86         83
2017.0                          2017.0      9          9

[218 rows x 3 columns]
```

1.7.1 Utilization analysis

Below takes a look at the total number of openly available volumes, and what percentage have been accessed since mid-2013

```
In [205]: #Calculate raw utilization rate

          #util = dates.accessed / dates['count']
          dates_a.loc[:, 'percent'] = dates_a.accessed / dates_a['count']

          print ("The average utlization rate for open volumes 1800-2017 is: %s" % dates_a['perc
          print ("The average utlization rate for open volumes 1800-1875 is: %s" % dates_a[1800:
          print ("The average utlization rate for open volumes 1876-1922 is: %s" % dates_a[1876:
          print ("The average utlization rate for open volumes 1923-1962 is: %s" % dates_a[1923:
          print ("The average utlization rate for open volumes 1963-2017 is: %s" % dates_a[1963:


          dates_a
```

```
The average utlization rate for open volumes 1800-2017 is: 0.667521454608
The average utlization rate for open volumes 1800-1875 is: 0.713943612836
The average utlization rate for open volumes 1876-1922 is: 0.614021150498
The average utlization rate for open volumes 1923-1962 is: 0.744973594259
The average utlization rate for open volumes 1963-2017 is: 0.592764085188
```

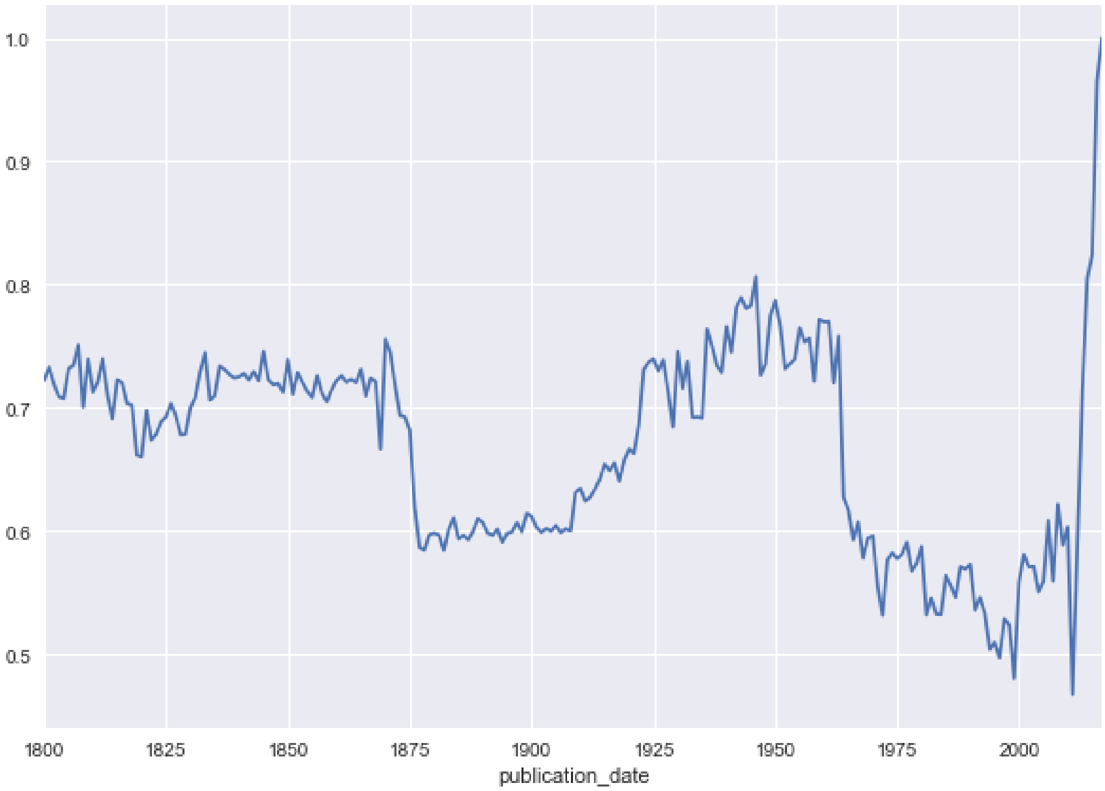
```
Out[205]:                  publication_date  count  accessed   percent
          publication_date
          1800.0                      1800.0   6303      4552  0.722196
          1801.0                      1801.0   5044      3696  0.732752
          1802.0                      1802.0   4917      3535  0.718934
          1803.0                      1803.0   5526      3918  0.709012
          1804.0                      1804.0   6044      4276  0.707478
          1805.0                      1805.0   5275      3859  0.731564
          1806.0                      1806.0   5421      3981  0.734366
          1807.0                      1807.0   5243      3938  0.751097
          1808.0                      1808.0   6491      4547  0.700508
          1809.0                      1809.0   5587      4131  0.739395
          1810.0                      1810.0   5923      4220  0.712477
          1811.0                      1811.0   6129      4416  0.720509
          1812.0                      1812.0   5528      4089  0.739689
          1813.0                      1813.0   5765      4095  0.710321
          1814.0                      1814.0   6035      4168  0.690638
          1815.0                      1815.0   6218      4493  0.722580
          1816.0                      1816.0   6570      4731  0.720091
          1817.0                      1817.0   7113      5003  0.703360
          1818.0                      1818.0   7883      5533  0.701890
          1819.0                      1819.0   9146      6049  0.661382
          1820.0                      1820.0  10086      6656  0.659925
          1821.0                      1821.0   9482      6617  0.697849
          1822.0                      1822.0  10389      6997  0.673501
          1823.0                      1823.0  10417      7071  0.678794
          1824.0                      1824.0  10661      7339  0.688397
          1825.0                      1825.0  12323      8537  0.692770
          1826.0                      1826.0  11429      8037  0.703211
          1827.0                      1827.0  11301      7834  0.693213
          1828.0                      1828.0  12498      8474  0.678028
          1829.0                      1829.0  13462      9134  0.678502
          ...                            ...    ...       ...       ...
          1988.0                      1988.0  18228     10406  0.570880
          1989.0                      1989.0  17658     10043  0.568751
          1990.0                      1990.0  17880     10240  0.572707
          1991.0                      1991.0  17124      9173  0.535681
          1992.0                      1992.0  17786      9708  0.545823
          1993.0                      1993.0  15630      8329  0.532885
          1994.0                      1994.0  14859      7484  0.503668
          1995.0                      1995.0  15755      8025  0.509362
```

```
1996.0                      1996.0 13742       6825  0.496653
1997.0                      1997.0 13434       7101  0.528584
1998.0                      1998.0 12938       6770  0.523265
1999.0                      1999.0 27529      13217  0.480112
2000.0                      2000.0 11329       6323  0.558125
2001.0                      2001.0  8303       4822  0.580754
2002.0                      2002.0  8515       4860  0.570757
2003.0                      2003.0  8464       4835  0.571243
2004.0                      2004.0  7574       4167  0.550172
2005.0                      2005.0  5513       3081  0.558861
2006.0                      2006.0  5694       3465  0.608535
2007.0                      2007.0  5868       3281  0.559134
2008.0                      2008.0  4467       2776  0.621446
2009.0                      2009.0  5389       3171  0.588421
2010.0                      2010.0  3561       2148  0.603201
2011.0                      2011.0  1379        644  0.467005
2012.0                      2012.0  1148        677  0.589721
2013.0                      2013.0  1129        806  0.713906
2014.0                      2014.0   994        800  0.804829
2015.0                      2015.0   716        590  0.824022
2016.0                      2016.0    86         83  0.965116
2017.0                      2017.0     9          9  1.000000

[218 rows x 4 columns]
```

```
In [202]: #plot utlization per year
          dates_a['percent'].plot(figsize=(10,7))
```

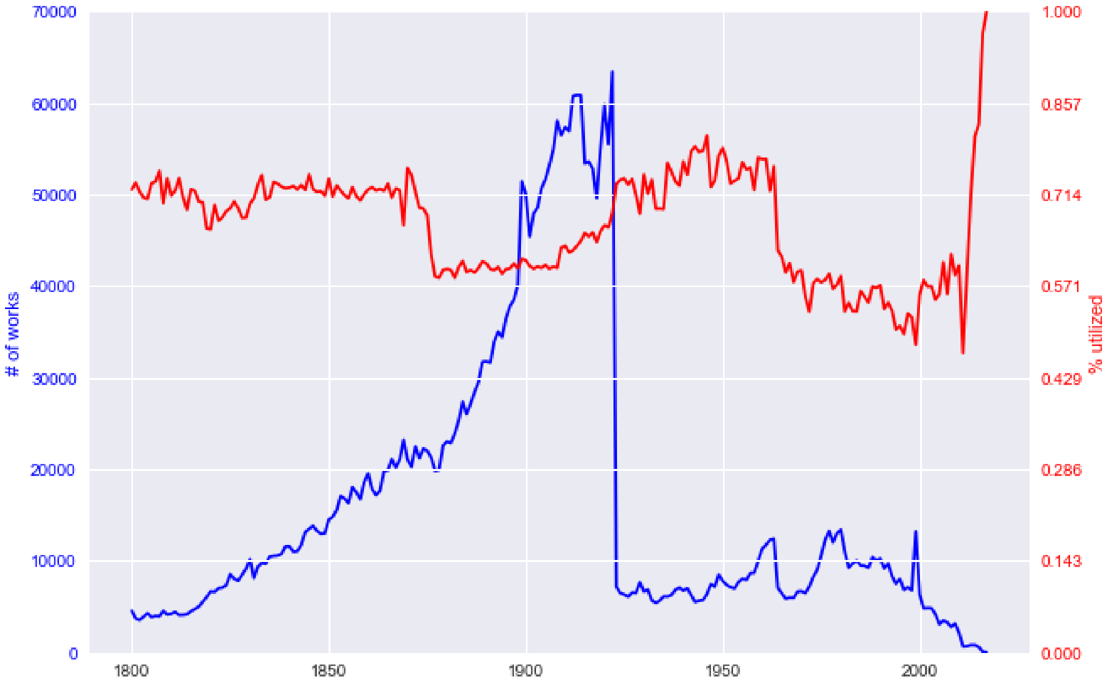
```
Out[202]: <matplotlib.axes._subplots.AxesSubplot at 0x152b9eb50>
```



```
In [174]: #Plot publication year distribution of accessed volumes (blue) against utilization rat
          import matplotlib.ticker as ticker

          fig, ax1 = plt.subplots(figsize=(10,7))
          ax1.plot(allow_years, color='b')
          ax1.set_ylabel('# of works', color='b')
          ax1.tick_params('y', colors='b')
          #ax1.set_yticks([0,10000])
          ax1.set_ylim([0,70000])

          ax2 = ax1.twinx()
          ax2.plot(util_a, color='r')
          ax2.set_ylabel('% utilized', color='r')
          ax2.tick_params('y', colors='r')
          ax2.set_ylim([0,1])
          ax2.yaxis.set_major_locator(ticker.MultipleLocator(1 / 7))
```



1.8 Publication date historgram

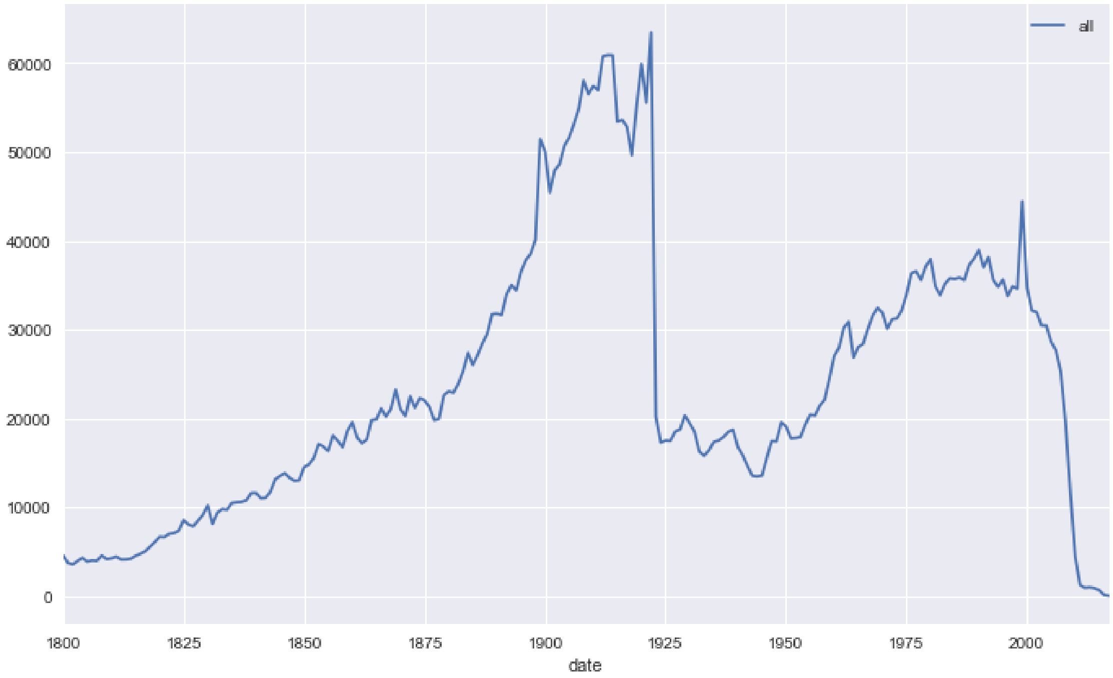
```
In [19]: pdata = pd.DataFrame()

         pdata.loc[:, 'all'] = all_years
         #pdata.loc[:, 'accessed'] = allow_years
         #pdata.loc[:, 'denied'] = deny_years
```

The cell below shows the publication date distribution of everything in the HathiTrust corpus

```
In [130]: #import seaborn as sns
          pdata.plot(figsize=(11.5,7))

Out[130]: <matplotlib.axes._subplots.AxesSubplot at 0x168eb6b90>
```

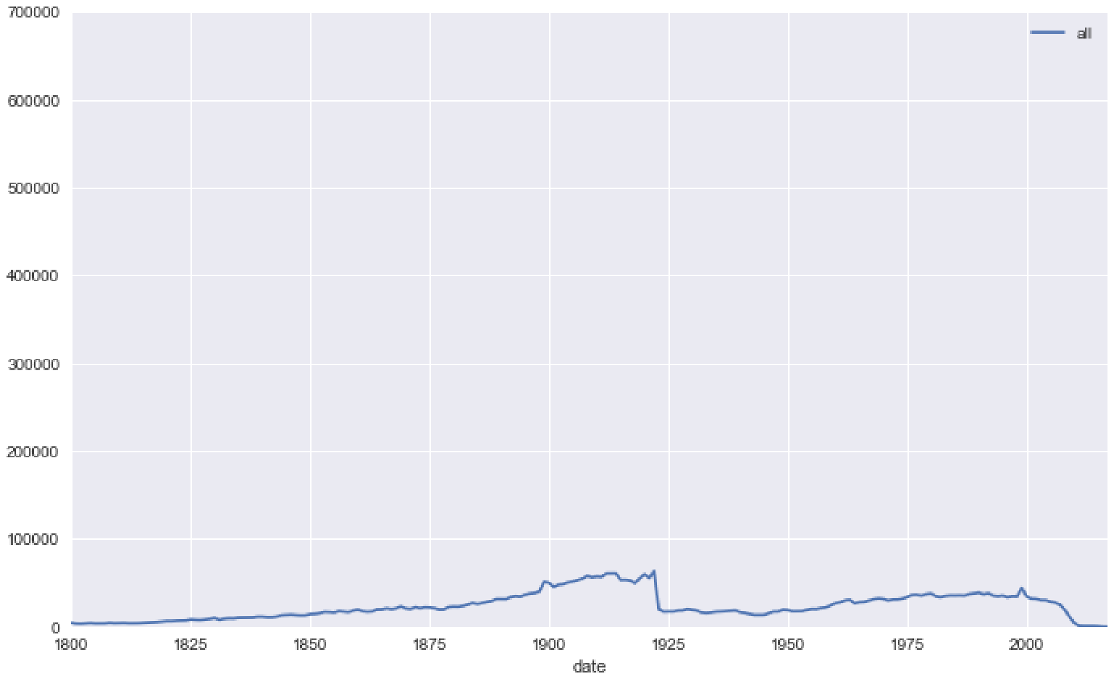


For comparison, here is the distribution of all publications in WorldCat, as estimates by Brian Lavoie and Lorcan Dempsey in an 2009 D-Lib article titled "Beyond 1923: Characteristics of Potentially In-copyright Print Books in Library Collections"

For context, here's the above Hathi publication date plot on the same scale as the OCLC data:

```
In [129]: pdata.plot(figsize=(11.5,7)).set_ylim(0, 700000)

Out[129]: (0, 700000)
```



1.9 Ongoing publications (serials) analysis

```
In [22]: ongoing = full[(full.date == 9999)].groupby('date')['id'].count()
         ongoing_a = full[(full.date == 9999) & (full.access == 'allow')].groupby('date')['id'].
         ongoing_d = full[(full.date == 9999) & (full.access == 'deny')].groupby('date')['id'].c

         print ("There are %s volumes from ongoing publications that triggered analytics events"
         print ("There are %s  openly available volumes from ongoing publications that triggered
         print ("There are %s 'Limited View' volumes from ongoing publications that triggered an
```

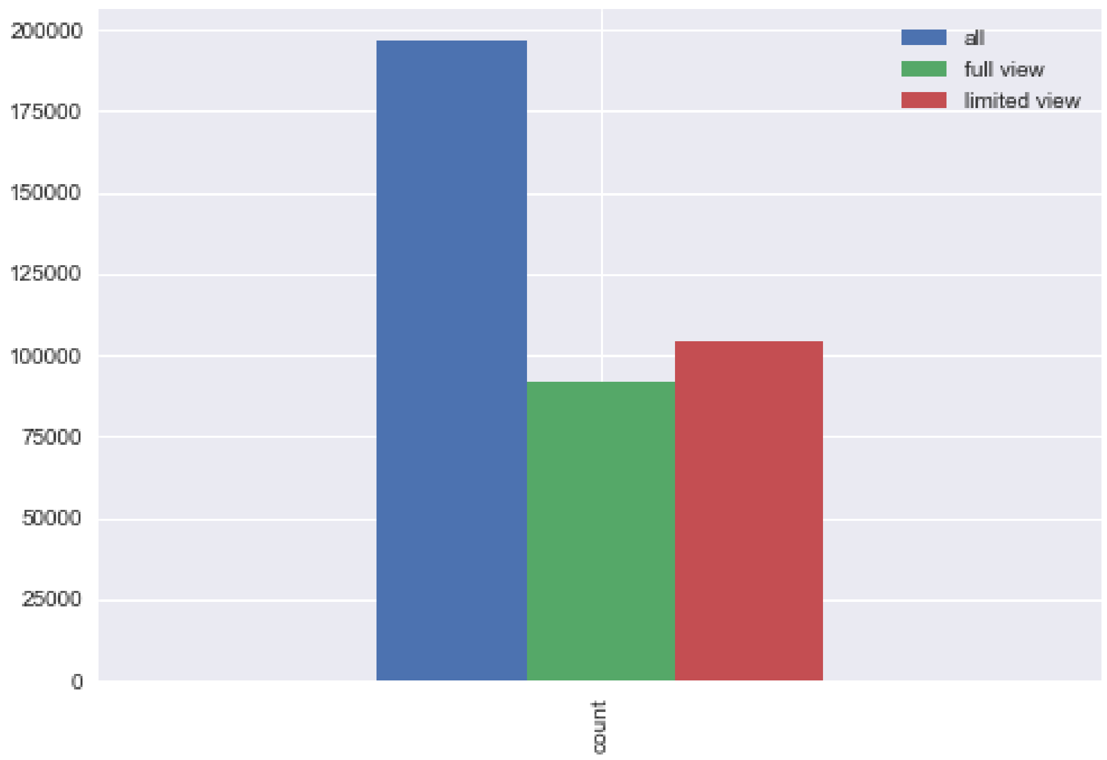
```
There are 196345 volumes from ongoing publications that triggered analytics events
There are 91847  openly available volumes from ongoing publications that triggered analytics eve
There are 104498 'Limited View' volumes from ongoing publications that triggered analytics event
```

```
In [176]: import seaborn as sns

          ongoing_all = pd.DataFrame()
          ongoing_all.loc['all', 'count'] = ongoing.iloc[0]
          ongoing_all.loc['full view', 'count'] = ongoing_a.iloc[0]
          ongoing_all.loc['limited view', 'count'] = ongoing_d.iloc[0]

          ongoing_all.T.plot(kind='bar', legend=True, width=.4, linewidth=.4)

Out[176]: <matplotlib.axes._subplots.AxesSubplot at 0x128ea1b50>
```



```
In [128]: from __future__ import division
          # Determine ongoing serials utilization rate

          accessed_ongoing = len(counts_ids[(counts_ids.date == 9999) & (counts_ids.access == 'a

          print( "%s of all possible 9999s are open volumes" % ongoing_a.iloc[0])
          print( "%s of all open 9999 volumes have been accessed" % accessed_ongoing)

          print("The overall utilization rate of ongoing serials in HathiTrust is {0:.2f}%".form

          counts_ids[(counts_ids.date == 9999) & (counts_ids.access == 'allow') & (counts_ids['c
```

```
91847 of all possible 9999s are open volumes
60144 of all open 9999 volumes have been accessed
The overall utilization rate of ongoing serials in HathiTrust is 65.48%
```

```
Out[128]:                         id  \
          118999   mdp.39015071886035
          238087       wu.89059402255
          238089       wu.89059402289
          63199    mdp.39015002304221
          238088       wu.89059402263
```

```
238091        wu.89059402313
238090        wu.89059402297
118956   mdp.39015071884410
76054    mdp.39015011819037
129828   mdp.39015095766716
119031   mdp.39015071888437
104763   mdp.39015051447657
171949     pst.000023992122
119019   mdp.39015071887850
239456        wu.89062853635
118961   mdp.39015071884634
118990   mdp.39015071885722
94909    mdp.39015031297883
63218    mdp.39015002314204
120280   mdp.39015074096499
62337    mdp.39015002000597
102351   mdp.39015046361062
118985   mdp.39015071885433
139625   njp.32101067262574
118939   mdp.39015071883651
115697   mdp.39015068806333
189062   uc1.32106007869669
75704    mdp.39015011488064
99969    mdp.39015038929710
118989   mdp.39015071885714
...                     ...
3280915          hvd.hwf7lg
3280882          hvd.hwekyi
3280772          hvd.hwcwlg
3280728          hvd.hwbn2x
3280708          hvd.hwbar4
3280687          hvd.hwaxa8
3280667          hvd.hwat25
3280665          hvd.hwarzd
3280613          hvd.hwabp2
3280556          hvd.hw6arg
3280425          hvd.hw3dys
3281283          hvd.hwjve4
3281426          hvd.hwkmla
3281629          hvd.hwmmsk
3283277          hvd.hxcnn2
3284298          hvd.li5lqp
3284266          hvd.li46l6
3284090          hvd.hy1ghh
3283735          hvd.hxj9jv
3283708          hvd.hxj2ya
3283374          hvd.hxdhvg
3283250          hvd.hx81h5
```

```
3282071          hvd.hwrgg2
3283173          hvd.hx5tr9
3282880          hvd.hx4ac4
3282655          hvd.hx27c4
3282615          hvd.hx1cis
3282598          hvd.hx17n4
3282597          hvd.hx17n1
4246464      wu.89107728693

                                                     title access    date  \
118999                                      [Publications]  allow  9999.0
238087   Roster of the Confederate soldiers of Georgia,...  allow  9999.0
238089   Roster of the Confederate soldiers of Georgia,...  allow  9999.0
63199    Encyclopedia of American Quaker genealogy, by ...  allow  9999.0
238088   Roster of the Confederate soldiers of Georgia,...  allow  9999.0
238091   Roster of the Confederate soldiers of Georgia,...  allow  9999.0
238090   Roster of the Confederate soldiers of Georgia,...  allow  9999.0
118956                                      [Publications]  allow  9999.0
76054    The abridged compendium of American genealogy;...  allow  9999.0
129828   Physical and biophysical foundations of pharma...  allow  9999.0
119031                                      [Publications]  allow  9999.0
104763   Encyclopedia of American Quaker genealogy, by ...  allow  9999.0
171949   Calendar of inquisitions miscellaneous, Chance...  allow  9999.0
119019                                      [Publications]  allow  9999.0
239456   Bosworth genealogy; a history of the descendan...  allow  9999.0
118961                                      [Publications]  allow  9999.0
118990                                      [Publications]  allow  9999.0
94909    A comprehensive study of Egyptian Arabic / Ern...  allow  9999.0
63218     Coins of the Roman empire in the British museum.  allow  9999.0
120280   The chemical formulary; a condensed collection...  allow  9999.0
62337                Roll pass design ... by W. Trinks ...  allow  9999.0
102351   The abridged compendium of American genealogy;...  allow  9999.0
118985                                      [Publications]  allow  9999.0
139625   Monumenta Ignatiana, ex autographis vel ex ant...  allow  9999.0
118939                                      [Publications]  allow  9999.0
115697    The rise of the Chinese Empire / Chun-shu Chang.  allow  9999.0
189062   Artists' pigments : a handbook of their histor...  allow  9999.0
75704    A lexicon of St. Thomas Aquinas based on the S...  allow  9999.0
99969    Encyclopedia of American Quaker genealogy, by ...  allow  9999.0
118989                                      [Publications]  allow  9999.0
...                                                    ...    ...     ...
3280915                                               Rit.  allow  9999.0
3280882             The poetical works of Robert Browning.  allow  9999.0
3280772  Universal geography : or a description of all ...  allow  9999.0
3280728  Kritisch-exegetischer Kommentar über das Neue ...  allow  9999.0
3280708  Storia di cento anni (1750-1850), narrata da C...  allow  9999.0
3280687  History of the United States from the discover...  allow  9999.0
3280667     Fishing guide : fisherman's friend booklet ...  allow  9999.0
```

```
3280665  Grundriss zur Geschichte der deutschen Dichtun...  allow  9999.0
3280613      Geschichte der neuern philosophie. Band 1-10.  allow  9999.0
3280556  Perepiska Mitropolita Kevskago Evgenia s g...  allow  9999.0
3280425  Paris révolutionnaire : Vieilles maisons, vieu...  allow  9999.0
3281283  Fischerei-Zeitung. Wochenschrift für die inter...  allow  9999.0
3281426                Oeuvres complètes de Clément Marot.  allow  9999.0
3281629               Shire David / me-et Daid Ber irel.  allow  9999.0
3283277  Mirabeau and the French revolution, by Fred Mo...  allow  9999.0
3284298                                          Final act  allow  9999.0
3284266  Constitution making in Indiana; a source book ...  allow  9999.0
3284090          Journal of the American Oriental Society.  allow  9999.0
3283735  Histoire de la vie de Mahomet, législateur de ...  allow  9999.0
3283708  Recueil d'archéologie orientale, par Ch. Clerm...  allow  9999.0
3283374  Das licht in dienste wissenschaftlicher forsch...  allow  9999.0
3283250  Memoirs of the Whig party during my time / by ...  allow  9999.0
3282071  The tales of Chekhov. / from the Russian by Co...  allow  9999.0
3283173  Studien zur Kenntniss des Izbornik Svjatoslava...  allow  9999.0
3282880  Le feld-maréchal prince Paskévitsch; sa vie po...  allow  9999.0
3282655  Die Bevölkerungs- und Wohnungs-Aufnahme [von] ...  allow  9999.0
3282615  A complete collection of the treaties and conv...  allow  9999.0
3282598  Deutsche Reichstagsakten, ältere Reihe. Auf Ve...  allow  9999.0
3282597  Deutsche Reichstagsakten, ältere Reihe. Auf Ve...  allow  9999.0
4246464  Personalhistorisk tidsskrift / udgivet af Samf...  allow  9999.0
```

```
                              oclc  count
118999                   426038752  25769
238087            1624676,27030216  24594
238089            1624676,27030216  22818
63199                       733646  18781
238088            1624676,27030216  18540
238091            1624676,27030216  17279
238090            1624676,27030216  16728
118956                   426038752  14500
76054                     68150295   8769
129828                         NaN   8608
119031                   426038752   8588
104763                      733646   7641
171949                    19432694   6280
119019                   426038752   6089
239456                     5338922   6088
118961                   426038752   5753
118990                   426038752   5298
94909                     23333231   5171
63218                      2061513   5129
120280                     1313469   4936
62337                     12831045   4789
102351                    68150295   4720
118985                   426038752   4694
```

```
139625                      1873136   4626
118939                    426038752   4323
115697                     65400237   4223
189062                     12804059   4196
75704                       2381195   4028
99969                        733646   3886
118989                    426038752   3775
...                             ...    ...
3280915                   236080348      1
3280882                     3209112      1
3280772            27808856,5930653      1
3280728                     4739802      1
3280708                    16126577      1
3280687                    19298298      1
3280667                    50323714      1
3280665                     3352027      1
3280613                    21498035      1
3280556                     6787838      1
3280425                     9116641      1
3281283                   235961492      1
3281426                    12708008      1
3281629  19186170,20005329,25232194      1
3283277                     1848313      1
3284298                   237347123      1
3284266                     3654268      1
3284090    1480509,3649140,47785421      1
3283735                    16949891      1
3283708                     5586639      1
3283374                    27273936      1
3283250                     1486324      1
3282071                     4454067      1
3283173                    13822468      1
3282880                    26657920      1
3282655                    45411933      1
3282615                    28703094      1
3282598                    22265070      1
3282597                    22265070      1
4246464                     1586068      1

[60144 rows x 6 columns]
```

1.10 Query analysis

```
In [1]: # Fun for later:
        # #To extract all possible search queries from the analytics

        # def queries():
        #     #This will extract
```