Journal_Names

August 12, 2020

1 Journal Names

Questions in the context of fuzzy matching.

- How many journal names appear more than once?
- What is the average length of the duplicated names vs the unique names?

Input file is a single larger JSON, mapping names to issns.

```
{
  "Acta Orientalia.": [
    "0001-6438"
  ],
  "Acta Orientalia (København)": [
    "0001-6438"
  ],
..
```

[4]:
```
import json
import pandas as pd
```

[5]:
```
with open("../data/name_to_issn.json") as f:
    mapping = json.load(f)
```

We have about 3M keys.

[7]:
```
len(mapping)
```

[7]: 2929727

[21]:
```
df = pd.DataFrame(((k, len(v)) for k, v in mapping.items()), columns=["name",␣
↪"issn_count"])
```

[25]:
```
len(df)
```

[25]: 2929727

[26]:
```
df.head()
```

[26]:
```
                          name  issn_count
0             Acta Orientalia.           1
1  Acta Orientalia (København)           1
2     The  publishers weekly.           1
3            Publishers weekly           1
4                    ASMT news           1
```

[31]:
```
unique_name = df[df.issn_count == 1]
```

[32]:
```
repeated_names = df[df.issn_count > 1]
```

[34]:
```
len(repeated_names)
```

[34]:
```
194241
```

[33]:
```
len(repeated_names) / len(df)
```

[33]:
```
0.06630003409874026
```

About 6% (or 194241) names are repeated.

[35]:
```
repeated_names.describe()
```

[35]:
```
          issn_count
count  194241.000000
mean        3.197523
std        25.081605
min         2.000000
25%         2.000000
50%         2.000000
75%         2.000000
max      8980.000000
```

Which name is shared by over 8000 ISSN?

[40]:
```
repeated_names.iloc[repeated_names.issn_count.argmax()] # Annual report.
```

[40]:
```
name          Annual report.
issn_count              8980
Name: 45907, dtype: object
```

It is the "Annual report."

[42]:
```
mapping["Annual report."][:10]
```

[42]:
```
['0706-537X',
 '1186-7957',
 '2324-1926',
```

```
 '1445-9248',
 '0872-3982',
 '1714-1524',
 '1037-8812',
 '0225-0241',
 '1327-6344',
 '0702-7702']
```

On average a repeated name will point to 3 ISSN. About 24k names point to more than 3 ISSN.

[45]:
```
len(repeated_names[repeated_names.issn_count > 3])
```

[45]:
```
24107
```

[49]:
```
repeated_names[repeated_names.issn_count > 3].sample(n=10)
```

[49]:

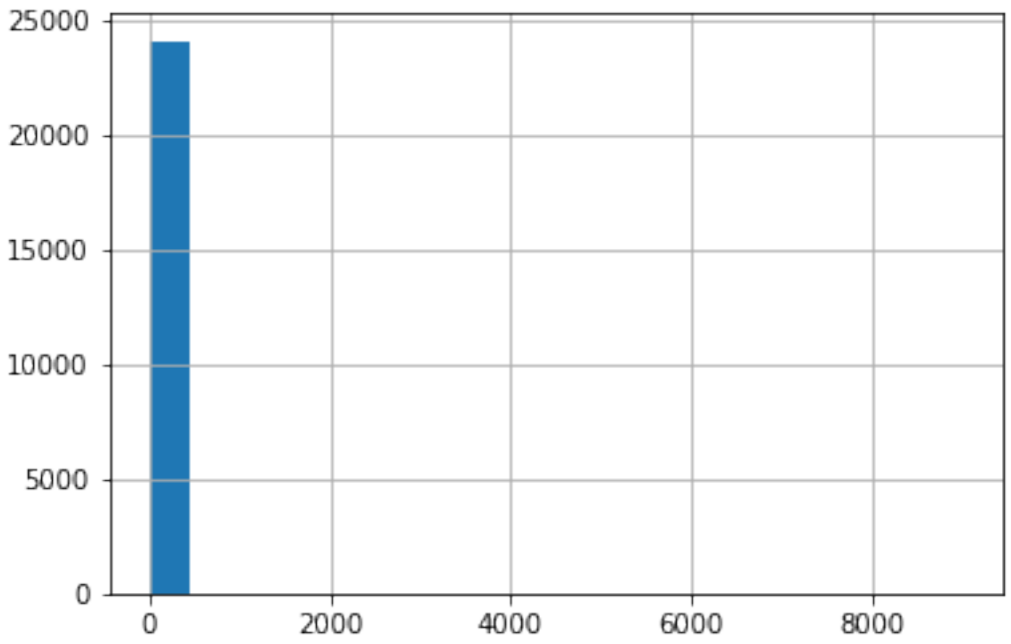
	name	issn_count
322100	Philosophica.	17
183928	Edad de oro.	4
294309	Horoskop.	10
517039	Le Grand journal.	11
1664616	Caleidoscop şcolar.	4
258430	La Feuille.	34
309546	The Wilson quarterly.	4
795859	Introductory research essay	4
1470838	Publicaciones del SEMYR.	4
657041	Le Kiosque.	14

[50]:
```
mapping["Philosophica."]
```

[50]:
```
['1285-9133',
 '1480-4670',
 '1487-5349',
 '1724-6598',
 '2183-0134',
 '2538-693X',
 '2610-8933',
 '2035-8326',
 '2295-9084',
 '1517-8889',
 '2249-5053',
 '2420-9198',
 '2654-9263',
 '2610-8925',
 '1158-9574',
 '0872-4784',
 '0379-8402']
```

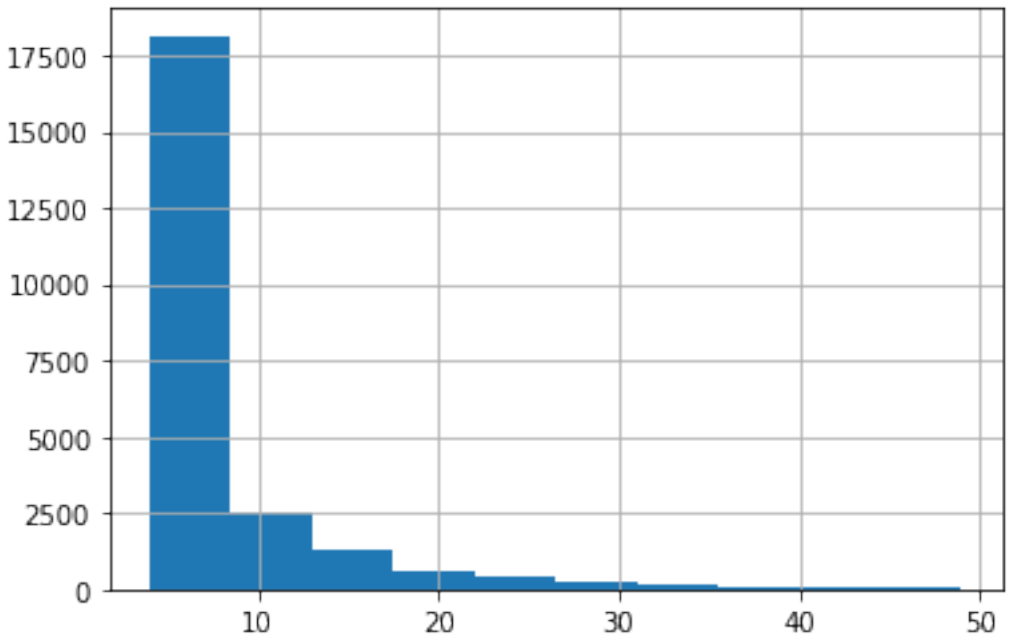
[61]:
```
repeated_names[repeated_names.issn_count > 3].issn_count.hist(bins=20)
```

[61]: <AxesSubplot:>



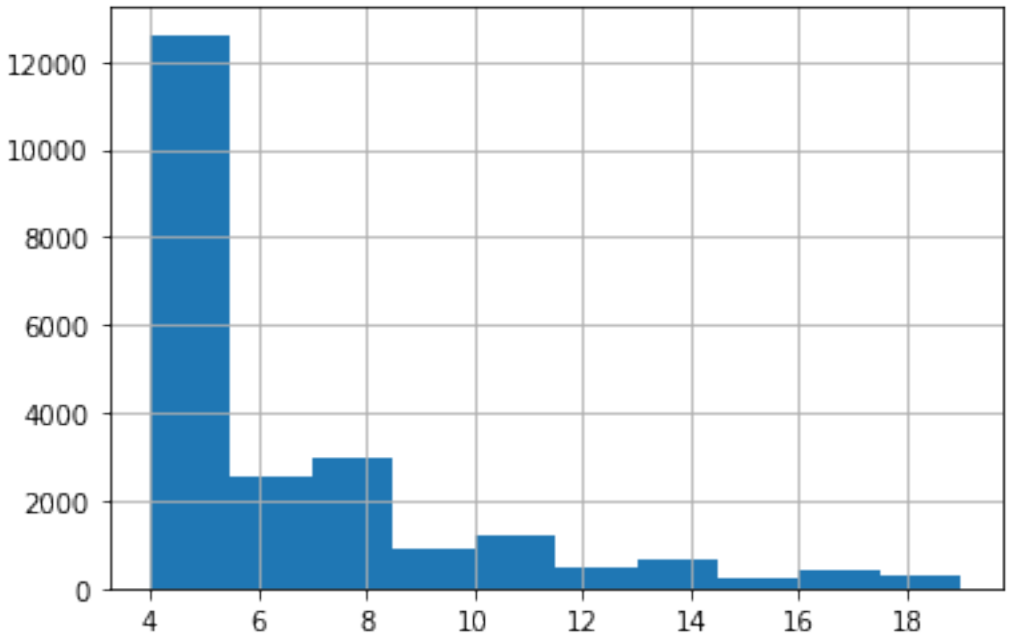
[60]:
```
repeated_names[(repeated_names.issn_count > 3) & (repeated_names.issn_count <␣
↪50)].issn_count.hist(bins=10)
```

[60]: <AxesSubplot:>



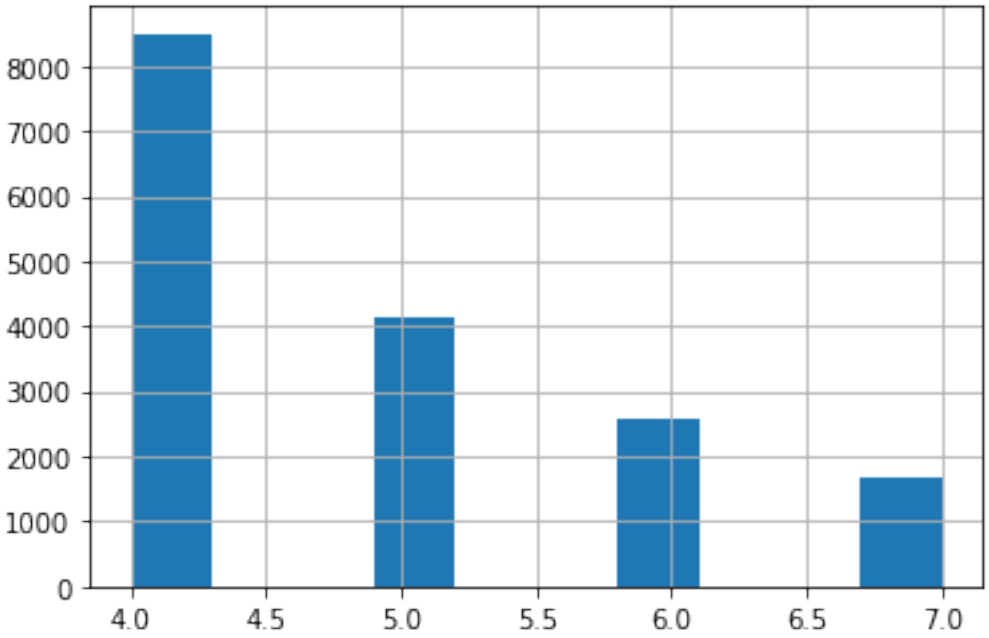
[62]:
```
repeated_names[(repeated_names.issn_count > 3) & (repeated_names.issn_count <
↪20)].issn_count.hist(bins=10)
```

[62]: <AxesSubplot:>



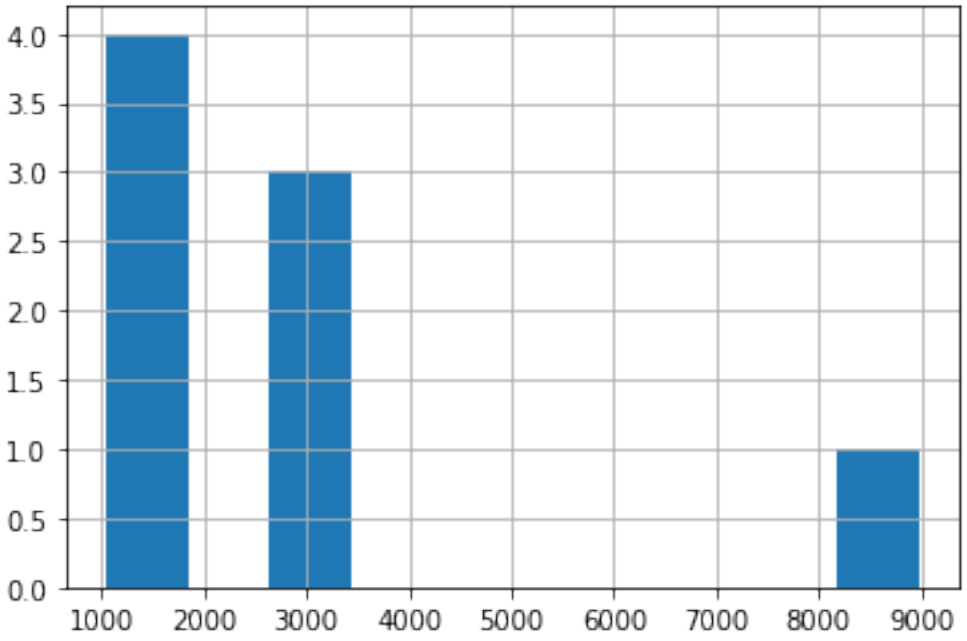
[64]:
```
repeated_names[(repeated_names.issn_count > 3) & (repeated_names.issn_count <␣
↪8)].issn_count.hist()
```

[64]: <AxesSubplot:>



[70]:
```
repeated_names[repeated_names.issn_count > 1000].issn_count.hist(bins=10)
```

[70]: <AxesSubplot:>



[71]:
```
repeated_names[repeated_names.issn_count > 1000]
```

[71]:

	name	issn_count
3499	Bulletin.	2752
7632	Newsletter.	2715
8317	Rapport.	1050
23662	Proceedings.	1403
45839	Annual report /	1090
45907	Annual report.	8980
45964	Annuaire.	1260
47217	Rapport annuel.	2656

[72]:
```
repeated_names[repeated_names.issn_count > 500]
```

[72]:

	name	issn_count
102	Bulletin d'information.	693
3218	Bulletin de liaison.	510
3499	Bulletin.	2752
7632	Newsletter.	2715
8317	Rapport.	1050
23662	Proceedings.	1403
45794	Report.	743
45839	Annual report /	1090
45907	Annual report.	8980
45964	Annuaire.	1260
46370	Jaarverslag.	675

```
47142          Rapport d'activité.         660
47217              Rapport annuel.        2656
49289               Jahresbericht.         518
57558                Annual report         760
121599          Alumni directory /         511
128827         Bulletin municipal.         521
150529                La  Lettre.         623
168933  Local climatological data.         613
269004                  Estimates.         535
```

[75]:
```
repeated_names[repeated_names.issn_count > 200]
```

[75]:
```
                                name  issn_count
102          Bulletin d'information.         693
2665                    Newsletter /         259
3218            Bulletin de liaison.         510
3499                        Bulletin.        2752
3926                         Boletín.         216
…                                …         …
425644        Rapport d'activité …           394
532500           Relatório e contas.         247
603144   Bildung und Beruf regional.         292
1006131             Vies de famille.         222
1110247        Country risk service.         271

[66 rows x 2 columns]
```

[76]:
```
repeated_names[repeated_names.issn_count > 100]
```

[76]:
```
                                name  issn_count
102          Bulletin d'information.         693
2665                    Newsletter /         259
3218            Bulletin de liaison.         510
3499                        Bulletin.        2752
3926                         Boletín.         216
…                                …         …
1306798           Country commerce.          120
1318569     Bible studies for life.          159
1796742  LexisNexis practice guide.          101
2628387    Operational risk report.          119
2650557          Interempresas net.          108

[191 rows x 2 columns]
```

[82]:
```
repeated_names
```

[82]:

	name	issn_count
5	Activitas Nervosa Superior.	2
11	Library journal.	2
23	Acta cardiologica.	2
26	Actualidad económica.	3
31	Acta Ornithologica.	3
…	…	…
2929626	Modern machine shop México.	2
2929635	Lecture notes in control and information scien…	2
2929646	Critical Studies in Dance Leadership and Inclu…	2
2929691	Nigerian Journal of Wildlife Management	2
2929702	Verzeichniss der Kunstwerke lebender Künstler,…	2

[194241 rows x 2 columns]

If a name matches a repeated name exactly or fuzzy matches to a repeated name and there is not other information available, the match status must be ambigious.

[ ]: