**Uncovering Art History |** Web Scraping Documentation

# **Table of Contents**

* [**Project Information**](https://docs.google.com/document/d/1hGm6dCzjn_5NuhgRv_6cCLHZeB1xPZ6BtcMKWwEEupQ/edit?pli=1&tab=t.0#heading=h.u0y615hbaaxb)
* [**Deliverables**](https://docs.google.com/document/d/1hGm6dCzjn_5NuhgRv_6cCLHZeB1xPZ6BtcMKWwEEupQ/edit?pli=1&tab=t.0#heading=h.ugwlvgu4w1te)
* [**Notes**](#_srk1u0jvpmx)
* [**Archives of American Art Collections Scraping**](https://docs.google.com/document/d/1hGm6dCzjn_5NuhgRv_6cCLHZeB1xPZ6BtcMKWwEEupQ/edit?pli=1&tab=t.0#heading=h.onrejm8r5jlw)
* [**Archives of American Art Journal Scraping**](https://docs.google.com/document/d/1hGm6dCzjn_5NuhgRv_6cCLHZeB1xPZ6BtcMKWwEEupQ/edit?pli=1&tab=t.0#heading=h.23ncxjbd0z0j)
* [**Data Analysis**](#_omuendbplonl)

# **Project Information**

* Team: [Abby Sypniewski](mailto:abbysyp@umich.edu) [Annabel Zhuang](mailto:annazjf@umich.edu) [Tara Dorje](mailto:tydorje@umich.edu) | University of Michigan School of Information
  + Project Manager: [Abby Sypniewski](mailto:abbysyp@umich.edu)
  + AAA Supervisor: [Bourgoin, Marisa](mailto:bourgoinm@si.edu)
* Project Term: Mar 3, 2025–Mar 6, 2025
* Research Questions
  + *How do we uncover what Archives of American Art collections are cited?*
  + *Who has been citing articles from the Archives of American Art Journal?*

# **Deliverables**

* **Folder** | [Uncovering Art History | Web Scraping](https://drive.google.com/drive/folders/10ofF8VP0xccel0hXVd-XBkDX2dkmw52q?usp=drive_link)
  + **Documentation** | [Uncovering Art History | Web Scraping Documentation](https://docs.google.com/document/d/1hGm6dCzjn_5NuhgRv_6cCLHZeB1xPZ6BtcMKWwEEupQ/edit?pli=1&tab=t.0)
  + **Scripts** | [scripts](https://drive.google.com/drive/u/0/folders/1N3cqDPfHD3Ob-wOpU8VF5zcGvB8Eg7qz)
    - **Collection Titles** | [collection-titles](https://drive.google.com/drive/folders/1chH6Ol2c13TST-zlzAEdS-g-U0Bjx5h-?usp=drive_link)
      * [**get-collection-titles.py**](https://drive.google.com/file/d/1BysnJ4j6Ub4KXkMiJpfOhS2QgZR50PQF/view?usp=drive_link)
      * [**collection-titles.txt**](https://drive.google.com/file/d/1swb73jMpiGNznoqwmUDqCV2Ri21zuuLD/view?usp=drive_link)
    - **Collections** | [collections](https://drive.google.com/drive/folders/1BDBT0kJP0K4wuxI2uCslTgofpOZGOCQG?usp=drive_link)
      * [**get-text-contents-from-urls-COLLECTIONS.py**](https://drive.google.com/file/d/1oO4y_eHH8y6C17HpiRc4rSYBpfJ2Zi4n/view?usp=drive_link)
      * [**get-urls-from-open-library-COLLECTIONS.py**](https://drive.google.com/file/d/123Xw37VJcx6JNqH3nFRZoqx0r_PwGbgl/view?usp=drive_link)
      * **c**[**ollections-search-results.csv**](https://drive.google.com/file/d/101JX2gS-Zt8929BBfEBm1texy6yzeZ-w/view?usp=drive_link)
      * [**issues.txt**](https://drive.google.com/file/d/1PVfIJM77EL6lvJFgpf-8vbj6u9pnB5J2/view?usp=drive_link)
    - **Journal** | [journal](https://drive.google.com/drive/folders/1OW4w5D3ITxDwOMiaolBQj65b-vzQ0YqZ?usp=drive_link)
      * [**get-text-contents-from-urls-JOURNAL.py**](https://drive.google.com/file/d/1re3wM3Rc4gpWZEjwfhk8yPfUro-VNlB_/view?usp=drive_link)
      * [**get-urls-from-open-library-JOURNAL.py**](https://drive.google.com/file/d/14irEpoXzmp-JA3sPqBtkVAEgMp4AgJfq/view?usp=drive_link)
      * [**journal-search-results.csv**](https://drive.google.com/file/d/1CyfHnqfNb_g8QL6q1Aqs1RvE3A0NAptV/view?usp=drive_link)
      * [**issues.txt**](https://drive.google.com/file/d/1ir8qyDQ4zukt0XkFd7F_BeLyWdcez7vw/view?usp=drive_link)
  + **Data Analysis** | [data-analysis](https://drive.google.com/drive/folders/1v8JGdYacdrqT94B-KWD4yfw-GdHRBKXX?usp=drive_link)
    - [**top\_20\_collection\_titles.png**](https://drive.google.com/file/d/1GiBvc6t19uHZMi3GbvESFpeG3mTEuH0F/view?usp=drive_link)
    - [**collection\_title\_counts.csv**](https://drive.google.com/file/d/1IUrulAskWndJH29eNoLdH3SfEx0z11ct/view?usp=drive_link)
    - [**collection\_title\_counts\_sorted.csv**](https://drive.google.com/file/d/1VoYGmcQUJxNlduipRpgGeSZkXQcrIWsM/view?usp=drive_link)
    - [**collection\_title&snippets\_analysis.ipynb**](https://drive.google.com/file/d/1y5ro1s2EoyrMAO1b4IV72tU_WsRguXMB/view?usp=drive_link)
    - [**snippets\_similarity\_score\_with\_title.csv**](https://drive.google.com/file/d/1n5GCTYGQ4JSIuMnlNKRuGvZAM2L11lGp/view?usp=drive_link)
    - **s**[**nippets\_with\_title.csv**](https://drive.google.com/file/d/1bEbZDRpJGSzag1J9l9pJ6V0NBM4OhhLN/view?usp=drive_link)
    - [**collection-titles.csv**](https://drive.google.com/file/d/1CBf9uLji9p_ppQiEoKsjkEwZov_oZkwq/view?usp=drive_link)

# **Notes**

* **Field Names**
  + In both [**journal-search-results.csv**](https://drive.google.com/file/d/1CyfHnqfNb_g8QL6q1Aqs1RvE3A0NAptV/view?usp=drive_link) and [**collection-search-results.csv**](https://drive.google.com/file/d/1ZoLWiFvsCpaZayqQhOvEW_tGzXyb7gGs/view?usp=drive_link), the field names are as follows:

| **Field Name** | **Value** |
| --- | --- |
| title | Title of work |
| url | Link to the work and its metadata in Open Library |
| author | Author of work |
| date\_published | Date published (note that the formatting is pretty inconsistent in Open Library, so it may be difficult to sort before doing some sort of data cleanup) |
| publisher | Publisher of work |
| snippets\_url | Link to interactive ereader of work in Archive-It and the snippets that include “Archives of American Art” (or “Archives of American Art Journal”) |
| number\_of\_snippets | Number of references to “Archives of American Art” (or “Archives of American Art Journal”) found in the work |
| snippets | A list of page numbers and context for each reference to “Archives of American Art” (or “Archives of American Art Journal”) found in the work |

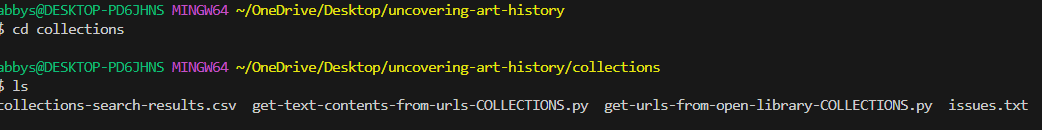
* **Limitations**
  + Open Library search result batches aren’t super consistent day-to-day, but the results that are being returned all do contain “Archives of American Art”
    - How we’ve accounted for it: The data in search-result.csv contains cumulative and unique results
  + Archive-It’s text recognition AI is ~mostly~ reliable
    - This means that some of the snippets we’re capturing have limited context surrounding them
    - How we’ve accounted for it: Our data analysis only
  + Relying on Internet Archive + Open Library alone for finding matches, not Google Scholar
    - How we’ve accounted for it: We confirmed through manual search that many searches for collections on Google Scholar and Open Library were identical
* **Next Steps + Future Projects**
  + Compare existing AAA Citation Zotero Library/AAA Journal Citations Document with our dataset
  + Archives of American Art Journal results is not on article-level, just returning matches for “Archives of American Art Journal” at this time – glancing at the citations, there may be an opportunity to identify specific articles and/or volumes
  + Further Data Analysis on the context of those citations
  + Scrape more citations from other repositories to cover a wider range of research

# **Archives of American Art Collections Scraping**

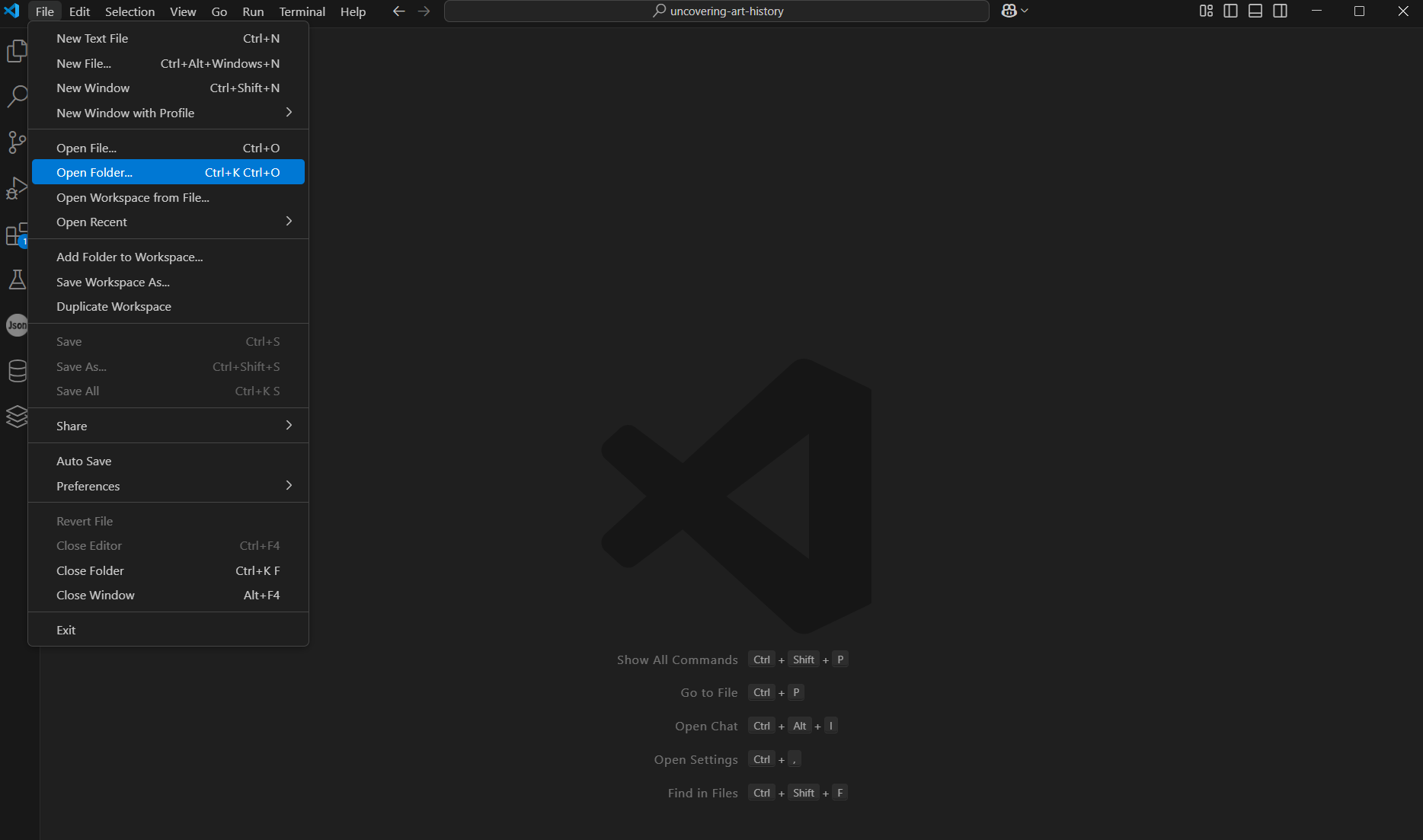
To run the scripts, we suggest downloading the entire [collections](https://drive.google.com/drive/folders/1BDBT0kJP0K4wuxI2uCslTgofpOZGOCQG?usp=drive_link) folder then operating the terminal from that folder. This could look like using **cd** if you’re familiar with the [command line](https://gomakethings.com/navigating-the-file-system-with-terminal/), or manually opening the folder in [VS Code](https://code.visualstudio.com/docs/python/run) then hitting the **Play** button. Instructions on how to do both are linked respectively.

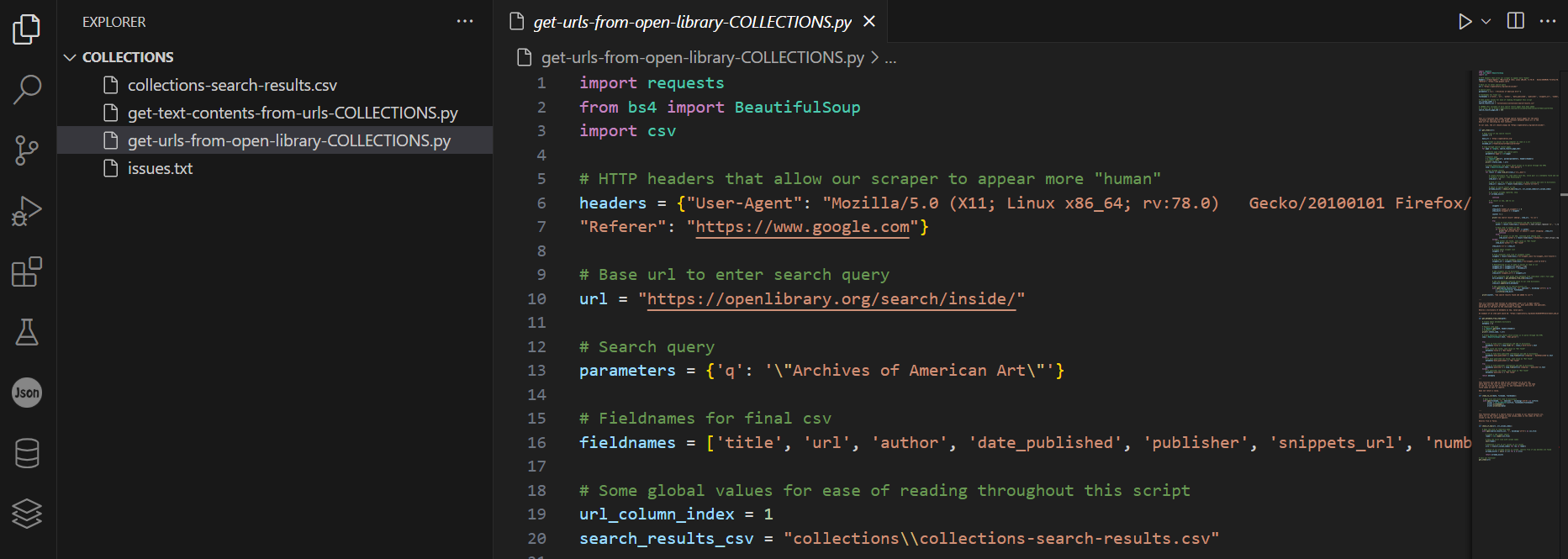
For example, in your terminal you will know you are in the right pathway if it ends with “collections” and by running the command **ls** you see the files:

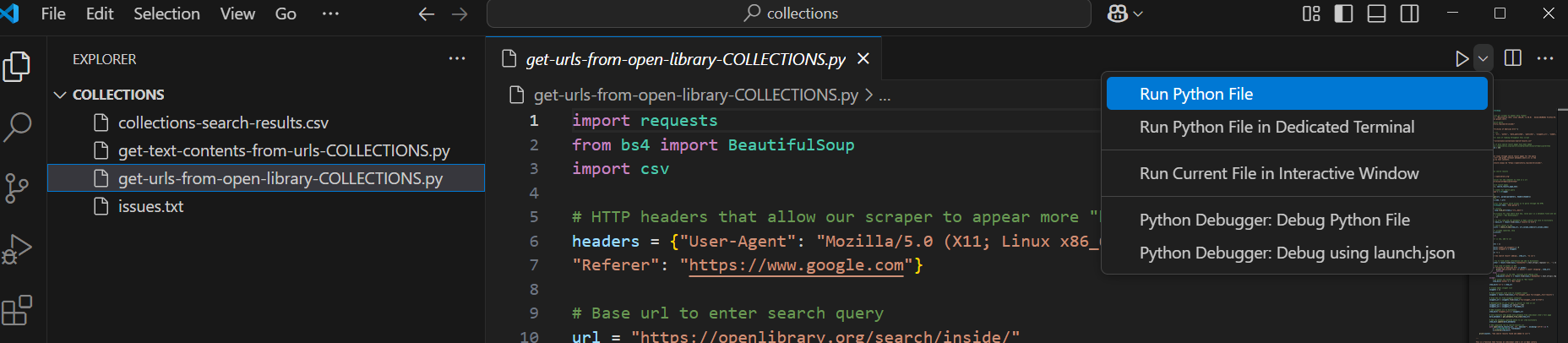
* **get-text-contents-from-urls-COLLECTIONS.py**
* **get-urls-from-open-library-COLLECTIONS.py**
* **collections-search-results.csv,** and
* **issues.txt**



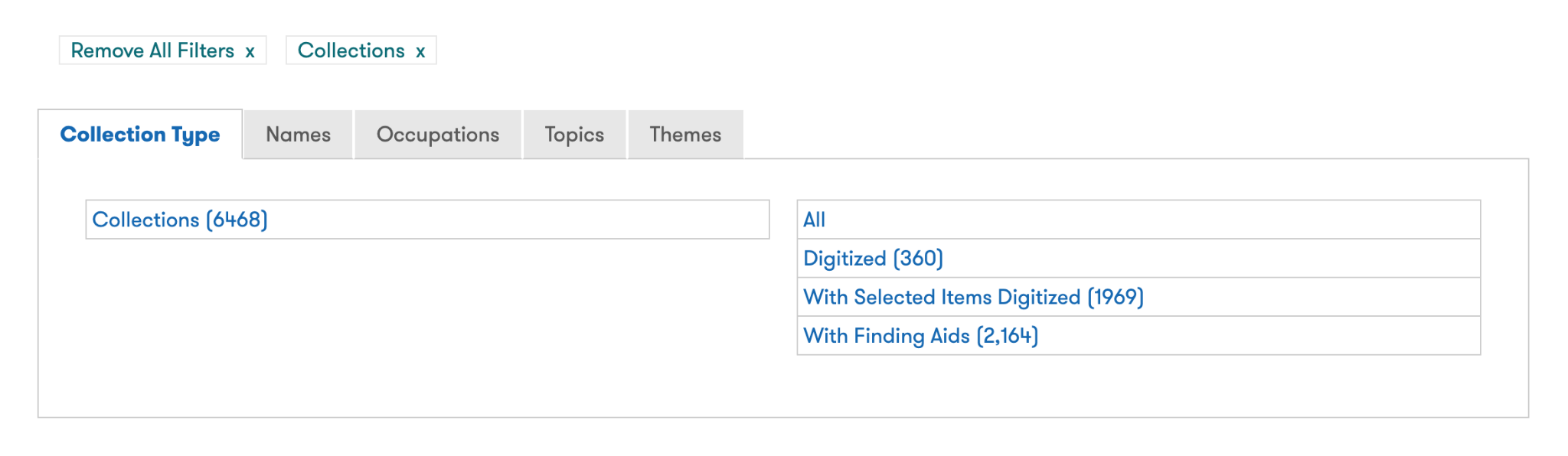
Alternatively, you can open VS Code, navigate to **File -> Open Folder**, then choose your downloaded collections folder. To run specific files, you can click on one to open it and click **Play -> Run Python File**. A terminal will automatically pop up at the bottom of VS Code.



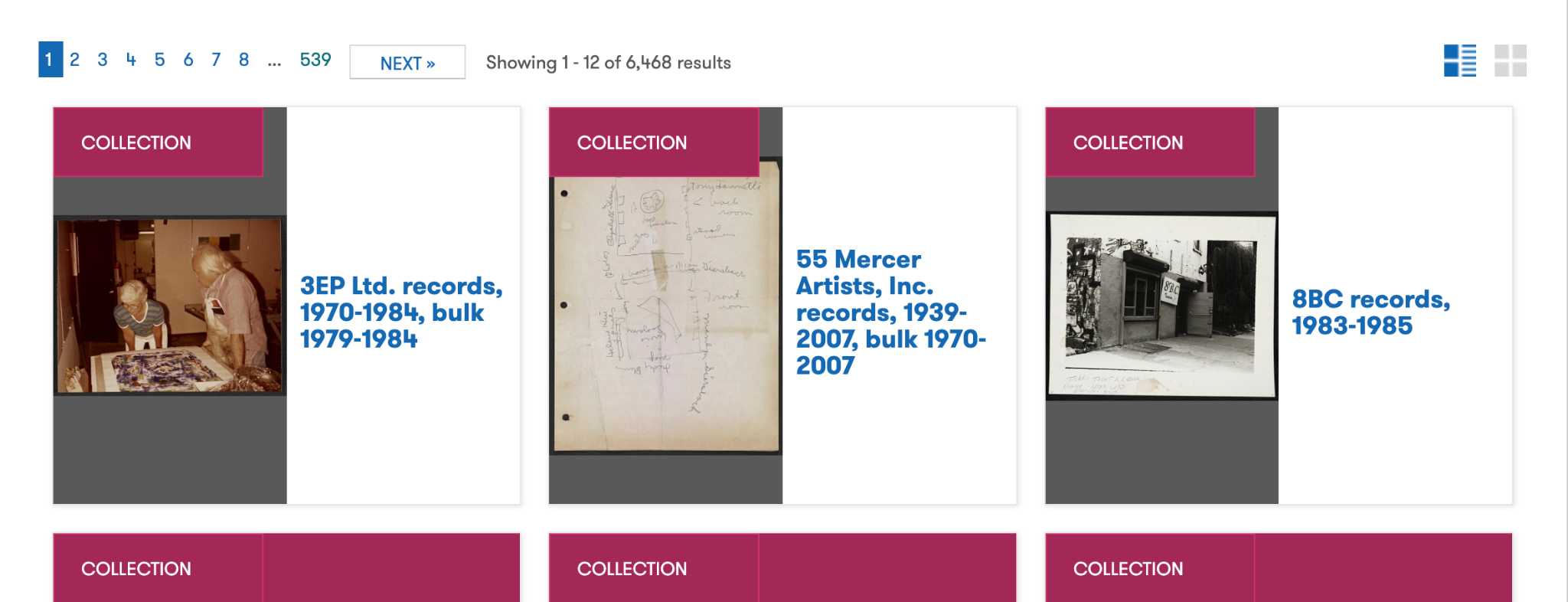




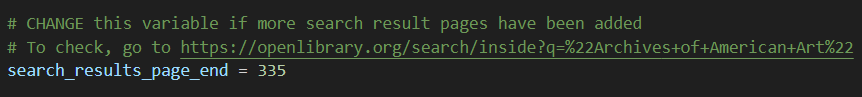
1. **Navigate to Archives of American Art** [**Search**](https://www.aaa.si.edu/search/collections?&edan_fq%5B%5D=p.edanmdm.descriptivenonrepeating.record_id:AAADCD_coll_*) **page**
   1. Filter to Collections only | Collection Types > Collections > All | [**Link**](https://www.aaa.si.edu/search/collections?&edan_fq%5B%5D=p.edanmdm.descriptivenonrepeating.record_id:AAADCD_coll_*)



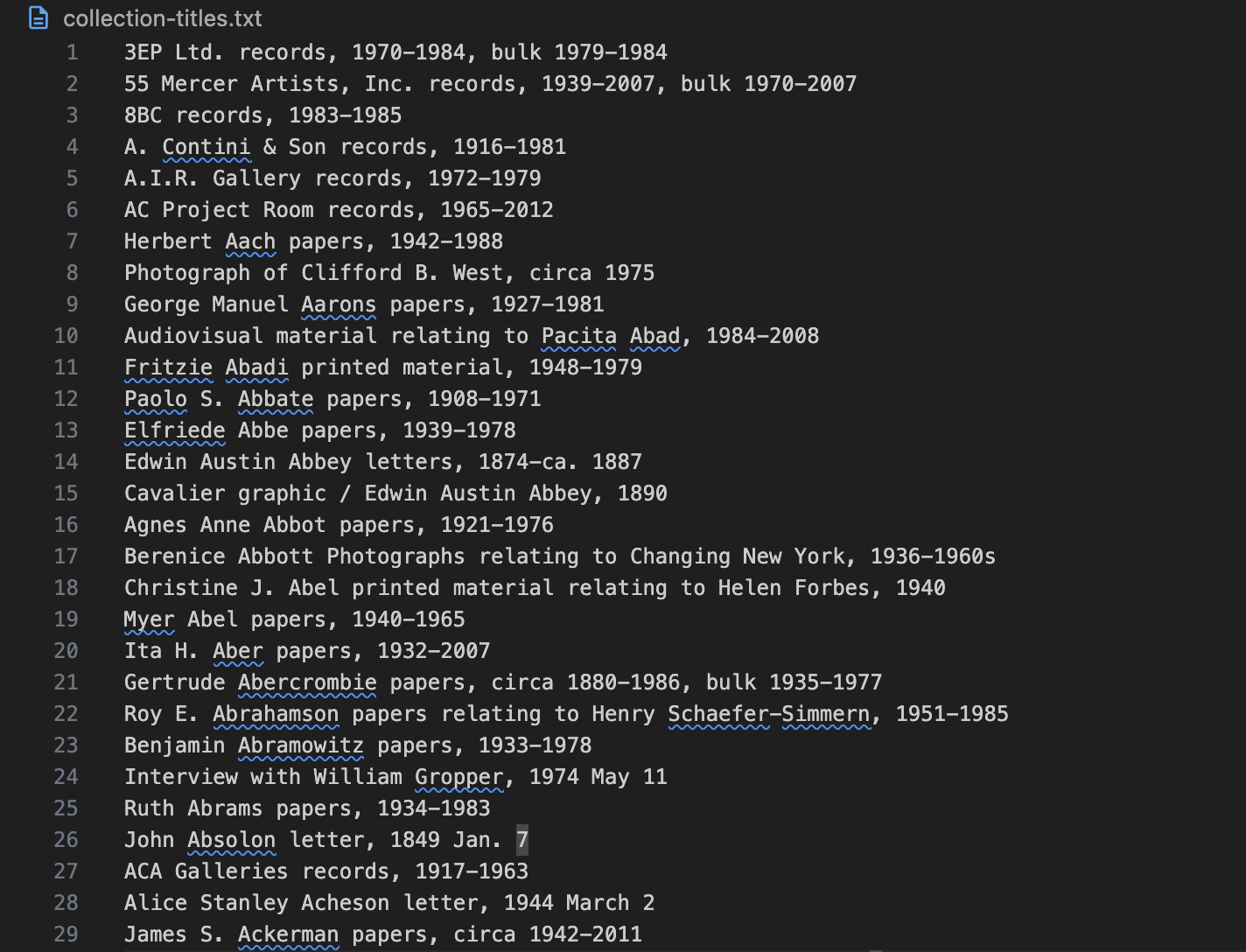
* 1. Check the total number of pages of results | Example: 539



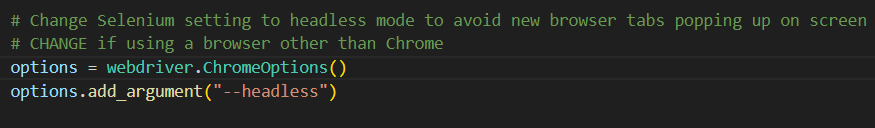
1. **Open get-collection-titles.py**
   1. Adjust **search\_results\_end\_page** to the total number of pages of results + 1
      1. Example: If there are 334 search result pages, **search\_results\_end\_page = 335**



* 1. Enter the command **python3 get-collection-titles.py** into your terminal and press ‘Enter’ to run it
     1. Note: Make sure to navigate to the correct folder where **get-collection-titles.py** is located using your terminal before running the command
     2. Note: Run the command using either **python3 get-collection-titles.py** or **python get-collection-titles.py** depending on the Python version available on your machine
     3. Note: Running this command will create the document **collection-titles.txt**, which contains a listing of all Collections of the Archives of American Art available through the Search page
     4. Example: The first 29 rows of **collection-titles.txt**

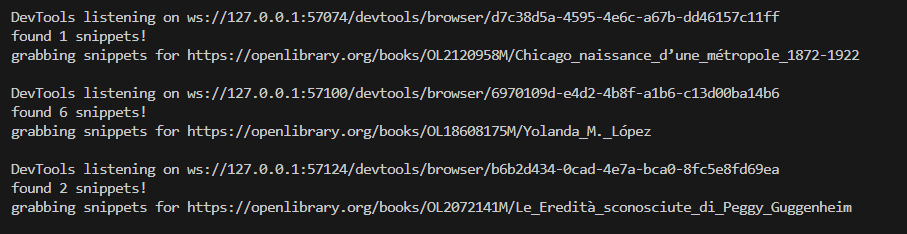


1. **Open collections-get-text-contents-from-urls.py**
   1. If this is your first time running the scripts, you may need to first install [Python](https://www.python.org/downloads/). Additionally, you will need to install the following modules by running these commands in your terminal:
      1. **pip3 (or pip) install selenium**
      2. **pip3 (or pip) install requests**
      3. **pip3 (or pip) install beautifulsoup4**
      4. **pip3 (or pip) install furl**
   2. If you’re using a default browser other than Chrome, change line 61 from **options = webdriver.ChromeOptions()** to **options = webdriver.FirefoxOptions()**



* 1. Enter the command **python3 collections-get-text-contents-from-urls.py** into your terminal and press ‘Enter’ to run it

1. **Open collections-get-urls-from-open-library.py**
   1. Enter the command **python3 collections-get-urls-from-open-library.py** into your terminal and press ‘Enter’ to run it
      1. Note: Waiting for this process to complete will be the longest step in the procedure
   2. Example of terminal output once script is done running:

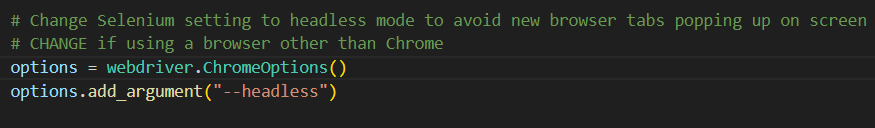


1. **Proceed to** [**Data Analysis**](#_omuendbplonl)

# **Archives of American Art Journal Scraping**

To run the scripts, we suggest downloading the entire [journal](https://drive.google.com/drive/folders/1OW4w5D3ITxDwOMiaolBQj65b-vzQ0YqZ?usp=drive_link) folder then operating the terminal from that folder. This could look like using **cd** if you’re familiar with the [command line](https://gomakethings.com/navigating-the-file-system-with-terminal/), or manually opening the folder in [VS Code](https://code.visualstudio.com/docs/python/run) then hitting the **Play** button. Instructions on how to do both are linked respectively. [See collections section](#_onrejm8r5jlw) for visuals on how to open a folder and run a file in the terminal and in VS Code.

1. **Open get-text-contents-from-urls-JOURNAL.py**
   1. If you’re using a default browser other than Chrome, change line 61 from **options = webdriver.ChromeOptions()** to **options = webdriver.FirefoxOptions()**



* 1. Enter the command **python3 get-journal-text-contents-from-urls-JOURNAL.py** into your terminal and press ‘Enter’ to run it

1. **Open get-urls-from-open-library-JOURNAL.py**
   1. Enter the command **python3 get-urls-from-open-library-JOURNAL.py** into your terminal and press ‘Enter’ to run it
      1. Note: Waiting for this process to complete will be the longest step in the procedure

# **Data Analysis**

**collection\_title&snippets\_analysis.ipyb**

1. Find the most similar collection title and the respective score for each snippet:

Output dataset -> **snippets\_similarity\_score\_with\_title.csv**

This dataset could be used for examining how the machine learning algorithm matches the snippet, which does not have a standardized way for referencing the Archives of American Art collection title.

* 1. Using the ‘SentenceTransformer (all-MiniLM-L6-v2)’ model and ‘cosine similarity matrix’ to match all snippets to the most similar collection title (with the highest similarity score).
  2. The current script uses **6.5** as the similarity threshold for similarity. It could be changed based on the degree of matching we want in the script. The bigger the threshold, the more similarity is required to match the collection title and the snippet. If the threshold is 0, all snippets would match a title.

1. Find the snippet and collection title matching:

Output dataset -> **snippets\_with\_title.csv**

Each publication on internet archives contains multiple snippets, and each snippet can be matched with one collection title. The output dataset contains the summary of all snippets for one publication referencing the Archives of American Art and the list of all collection titles that got paired with those snippets separated by “|”.

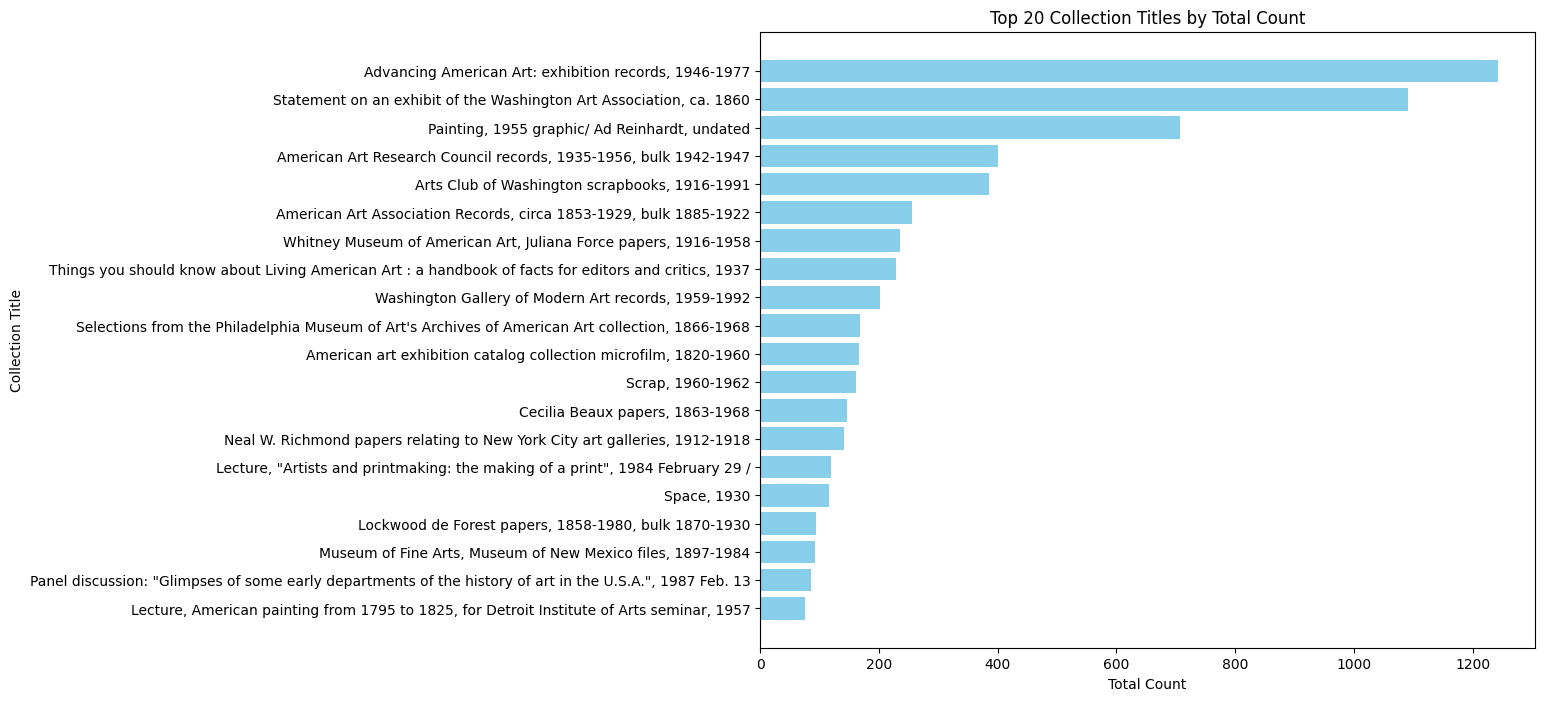
* 1. Using the same model and threshold as the previous session
  2. For snippets that did not get a match by using the similarity model, we will try to match the collection title with it by checking if the collection title is directly stated in the snippet. If so, we will match that snippet with the respective collection. This will increase the chance of the snippet being matched while preserving accuracy.

1. Find the occurrence of collection title on citations:

Output dataset -> **collection\_title\_counts\_sorted.csv**

This dataset contains the total occurrence of each collection title in publications.

1. “Total Count”: the number of occurrences of that title in the whole corpus of internet archive publications.
2. “Unique Count”: As some publications have cited a single Archives of American Art collection multiple times, this “Unique Count” column only counts those multiple citations in one publication once.
3. Limitation:
   1. The similarity model is not 100% accurate for matching the collection title with the snippet, as people reference collections differently and may not be in a standard recognizable format. It might have some errors matching with the correct collection title. However, it does increase the chance that the snippet will get a match with the existing collection title compared to directly finding the collection title inside the snippet. Using the similarity model method captures a large portion of citations that did not include the direct collection name. The balance between matching accuracy and the number of matches could be tailored by adjusting the threshold of the similarity model.
4. Visualization:



The most snippet in the top 2 collection titles did not get matched to the correct title due to the fact that the snippet did not mention any actual collection.

Here is a sample demonstration of the false matching for the top 2 title:

‘Advancing American Art: exhibition records, 1946-1977’ : “'Page 628': 'archives of american art (Smithsonian)'},” or 'Page 140': 'Archives of American Art, Smithsonian Institution, Washington, D.C.'}]

‘Statement on an exhibit of the Washington Art Association, ca. 1860’ : 'Page 140': 'Archives of American Art, Smithsonian Institution, Washington, D.C.'}]

Starting from the top 3 collection titles, they have meaningful matches.