Homework 1: Crawling

INF 558 BUILDING KNOWLEDGE GRAPH

DUE DATE: Friday, 08/31/2018 @ 11:59pm on Blackboard.

Ground Rules
This homework must be done individually. You can ask others for help with the tools, but the submitted homework has to be your own work.

Summary
In this homework, you will create/use Web crawlers to collect webpages from **Smithsonian American Art Museum** (SAAM). A Web crawler is a program that explores the Web, typically for the purpose of Web indexing. It starts with a list of seed URLs to visit, and as it visits each webpage, it finds the links in that web page, and then visits those links and repeats the entire process.

Two Web crawlers you have to use in this homework are:

- ACHE (https://github.com/ViDA-NYU/ache)
- Scrapy (https://scrapy.org)

Task 1 (3 points)

Crawl at least 3000 webpages of artworks in SAAM. The sample webpage is in figure 1. You have to submit two sets of webpages, which are obtained from ACHE and Scrapy respectively.

We provide a list of artwork URLs (mandatory_artworks.txt), and your result set **must include** these webpages. **All of the collected webpages should be artwork pages**. We will sample the webpages to check whether they are correct. For example, the page shown in Figure 2 should not be in your result set since it is not artwork page.

Task 2 (3 points)

Similar to task 1, you will crawl at least 5000 webpages of artists in SAAM using ACHE and Scrapy. The sample webpage is in figure 3. The list of required artists is in mandatory_artists.txt.

All of the collected webpages should be artist pages. We will sample the webpages to check whether they are correct. For example, the page shown in Figure 2 should not be in your result set since it is not artist page.

Task 3 (2 points)

Answer the following questions (maximum 2 sentences for each question):

- What is the seed URL(s)?
- How did you manage to only collect artwork/artist pages? How did you discard irrelevant pages?
- If you were not able to collect 3000/5000 pages, describe and explain your issues.

Task 4 (2 points)
Store your crawled webpages into CDR files. CDR files follow JSON Lines (http://jsonlines.org/) format. Each line in a CDR file is a valid JSON object that represents the information about one crawled webpage. The JSON object has the following attributes:

- doc_id: unique id for the webpage
- url: url of the webpage
- raw_content: html content of the webpage
- timestamp_crawl: when did you crawl

You can check the attached file sample_cdr.jl to understand how CDR format looks like. You should validate your JSON objects in your CDR file to ensure they have the correct format, especially the string value of 'raw_content' attribute.

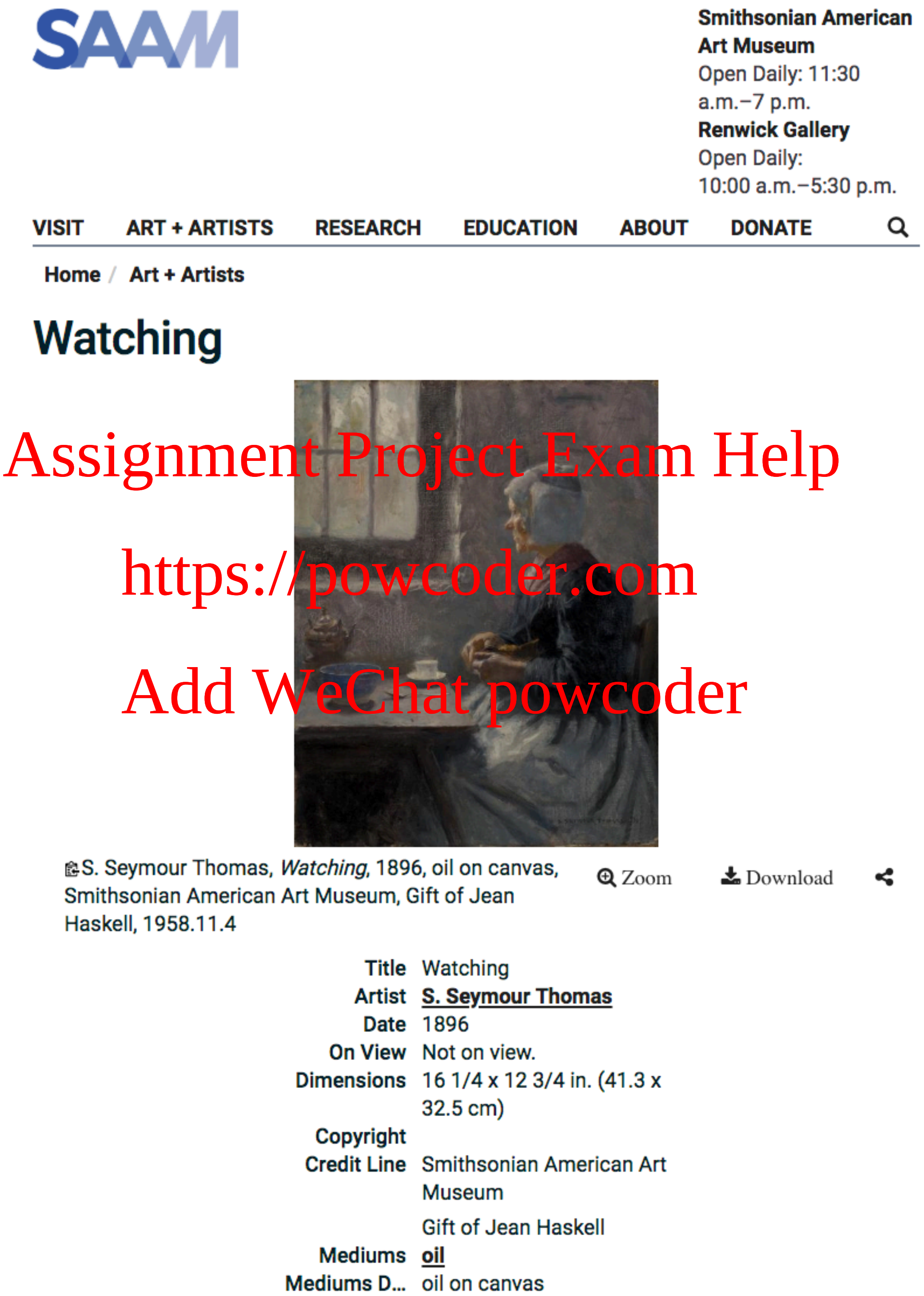
After you build your CDR files, use the provided script post_processing.py to reduce size of your files. The script takes one argument `<cdr_path>` which is path of your CDR file, and outputs a new cdr file at: `<cdr_path>.processed`. For example: `python post_processing.py /home/users/sample_cdr.jl` will output `/home/users/sample_cdr.jl.processed`. Refer to post_processing_usage.pdf for more information.

You will **submit** the new CDR files instead of your original files.

Submission Instructions
You must submit the following files/folders in a single .zip archive named Firstname_Lastname_hw1.zip and submit it via Blackboard:

- **Firstname_Lastname_hw1_report.pdf**: A pdf file containing your answers to the Task 1.
- CDR files contain all the web pages that you crawled:
  - **Firstname_Lastname_artist_ache_cdr.jl.processed**: Artist web pages you got from ACHE.
  - **Firstname_Lastname_artist_scrapy_cdr.jl.processed**: Artist web pages you got from Scrapy.
  - **Firstname_Lastname_artwork_ache_cdr.jl.processed**: Artwork web pages you got from ACHE.
  - **Firstname_Lastname_artwork_scrapy_cdr.jl.processed**: Artwork web pages you got from Scrapy.

- **source:** This folder includes all the code you wrote to accomplish Task 1, 2 and 4, for example, your Scrapy crawler, or your script/program to eliminate unwanted pages and store webpages into CDR format.

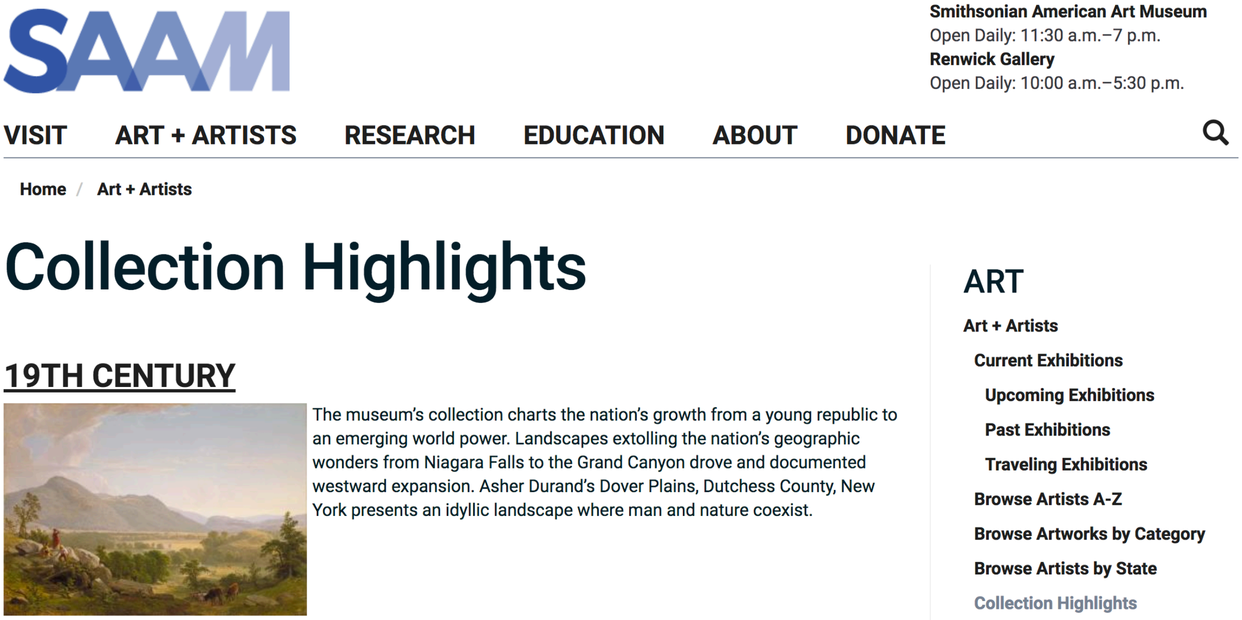


Watching

S. Seymour Thomas, *Watching*, 1896, oil on canvas, Smithsonian American Art Museum, Gift of Jean Haskell, 1958.11.4

Title	Watching
Artist	**S. Seymour Thomas**
Date	1896
On View	Not on view.
Dimensions	16 1/4 x 12 3/4 in. (41.3 x 32.5 cm)
Copyright	
Credit Line	Smithsonian American Art Museum Gift of Jean Haskell
Mediums	**oil**
Mediums D...	oil on canvas

Figure 1. An artwork in SAAM



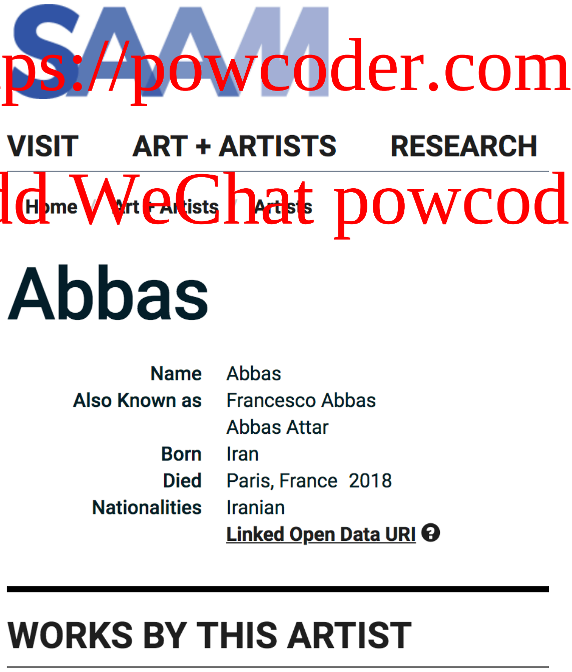
Smithsonian American Art Museum
Open Daily: 11:30 a.m.–7 p.m.
Renwick Gallery
Open Daily: 10:00 a.m.–5:30 p.m.

Collection Highlights

19TH CENTURY

The museum's collection charts the nation's growth from a young republic to an emerging world power. Landscapes extolling the nation's geographic wonders from Niagara Falls to the Grand Canyon drove and documented westward expansion. Asher Durand's Dover Plains, Dutchess County, New York presents an idyllic landscape where man and nature coexist.

Figure 2. An irrelevant page



Abbas

Name	Abbas
Also Known as	Francesco Abbas Abbas Attar
Born	Iran
Died	Paris, France 2018
Nationalities	Iranian
	Linked Open Data URI

WORKS BY THIS ARTIST

Figure 3. An artist in SAAM