Digging Deeper, Reaching Further

Module 2.2: Gathering Textual Data

Bulk retrieval

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

In this module we'll…

- Introduce options for automated data retrieval

  → *Facilitate researcher needs for data access*

- Practice web scraping with `wget` on the command line

  → *Gain confidence automating data collection*

- See how Sam gathered data for his Creativity Corpus

  → *Learn how one scholar used automated data collection*

Where we'll end up

```
19:00 ~ $ wget -l 1 --limit-rate=20k https://en.wikisource.org/wiki/George_Washington%27s_Fourth_
State_of_the_Union_Address --output-document=washington_4.txt
--2017-07-06 19:00:46--  https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the
_Union_Address
Resolving proxy.server (proxy.server)... 10.0.0.82
Connecting to proxy.server (proxy.server)|10.0.0.82|:3128... connected.
Proxy request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: 'washington_4.txt'

    [          <=>

2017-07-06 19:00:48 (20.0 KB/s) - 'washingto

19:00 ~ $
```

Run command from command line interface…

…to scrape the text from a webpage.

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>George Washington's Fourth State of the Union Address - Wikisource, the free online library</title>
<script>document.documentElement.className = document.documentElement.className.replace( /(^|\s)client-nojs(\s|$)/, "$1client-js$2" );</script>
<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"George_Washington's_Fourth_Sta
mw.user.tokens.set({"editToken":"+\\","patrolToken":"+\\","watchToken":"+\\","csrfToken":"+\\"});/*@nomin*/;

});mw.loader.load(["site","mediawiki.page.startup","mediawiki.user","mediawiki.hidpi","mediawiki.page.ready","mediawiki.searchSuggest","ext.gadget.charinsert","ext.gadget.Easy_LST","ext.gadget.d
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=ext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wikimediaBadges%7Cmediawiki.legacy.commo
<script async="" src="/w/load.php?debug=false&amp;lang=en&amp;modules=startup&amp;only=scripts&amp;skin=vector"></script>
<meta name="ResourceLoaderDynamicStyles" content=""/>
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=ext.gadget.Site&amp;only=styles&amp;skin=vector"/>
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vector"/>
<meta name="generator" content="MediaWiki 1.30.0-wmf.7"/>
<meta name="referrer" content="origin-when-cross-origin"/>
<link rel="alternate" type="application/x-wiki" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_Union_Address&amp;action=edit"/>
<link rel="edit" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_Union_Address&amp;action=edit"/>
<link rel="shortcut icon" href="/static/favicon/wikisource.ico"/>
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikisource (en)"/>
<link rel="EditURI" type="application/rsd+xml" href="//en.wikisource.org/w/api.php?action=rsd"/>
<link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/"/>
<link rel="alternate" type="application/atom+xml" title="Wikisource Atom feed" href="/w/index.php?title=Special:RecentChanges&amp;feed=atom"/>
<link rel="canonical" href="https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address"/>
<link rel="dns-prefetch" href="//login.wikimedia.org"/>
<link rel="dns-prefetch" href="//meta.wikimedia.org" />
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-George_Washington_s_Fourth_State_of_the_Union_Address rootpage-George_Washington_s_Fourth_State_of_the_Union_Address
        <div id="mw-head-base" class="noprint"></div>
        <div id="content" class="mw-body" role="main">
            <a id="top"></a>

                        <div id="siteNotice" class="mw-body-content"><!-- CentralNotice --></div>
                    <div class="mw-indicators mw-body-content">
</div>
            <h1 id="firstHeading" class="firstHeading" lang="en">George Washington's Fourth State of the Union Address</h1>
                            <div id="bodyContent" class="mw-body-content">
                            <div id="siteSub">From Wikisource</div>
                        <div id="contentSub"></div>
                                <div id="jump-to-nav" class="mw-jump">
```

Bulk retrieval

- Most researchers will need more than 1 or 10 texts
  - Hundreds, thousands, or millions of texts
- Getting lots of data could take lots of time!
  - Point-and-click is inefficient
  - Automate when possible

Automating retrieval

Transferring files

- FTP or SFTP: (Secure) File Transfer Protocol
  - moves files from one place to another on the internet
- rsync
  - Efficient: sends only the differences
  - Run from command line
  - Used by HathiTrust, can be used to download Extracted Features data

Automating retrieval

Web scraping (grabbing text on the web)

- Avoids tedious copying-and-pasting
- Some ways to scrape text from the web:
  - Run commands such as wget or curl in the command line
  - Write and run a script (a file of programming statements)
  - Use software such as webscraper.io or Kimono

We'll do an activity with this one later!

Web scraping for the wise

- Web scraping puts a large workload on targeted server
  - This can upset the data holder
- Some data providers are more than willing to share
  - Ask for access
  - Check for an API
- Otherwise, time your requests to add a delay between server hits
  - It’s polite
  - Also signifies you are not a malicious attacker

Automating retrieval

APIs (Application Programming Interfaces)

- Digital pathways to or from content
  - Sometimes need a “key” for access
- Can be used to gain programmatic access
  - Usually need to write code to retrieve content
  - Sometimes have graphical user interface (GUI)
- Examples: A number of digital content providers have APIs
  - Twitter API: display tweets on a non-Twitter website
  - Chronicling America API: https://chroniclingamerica.loc.gov/about/api/

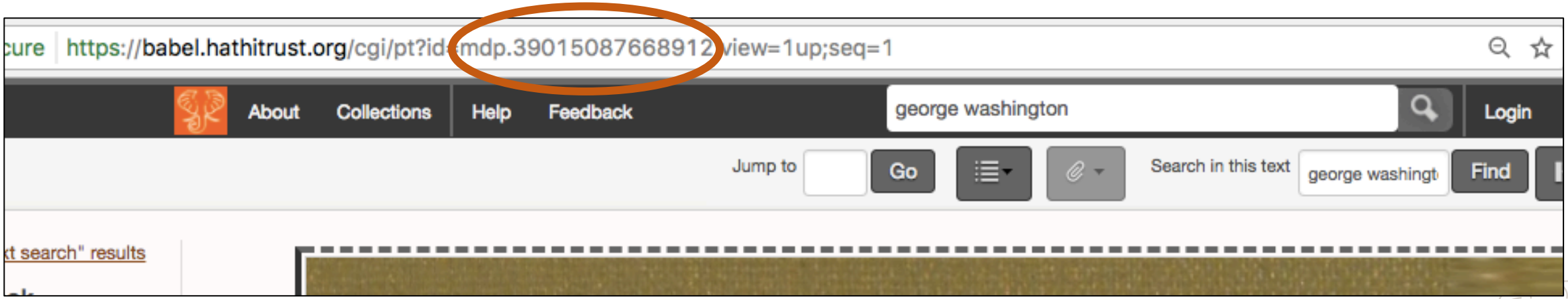
HathiTrust Bibliographic API

- Provides programmatic access to bibliographic metadata for volumes in the HathiTrust
- Metadata retrieved using a specially formatted URL and the volume ID number
- More information: https://www.hathitrust.org/bib_api

Hands-on activity

☞ *See Handout p. 1*

- Use the HathiTrust Bibliographic API to retrieve the metadata for a volume of your choice
  - Search HTDL for a volume (hathitrust.org)
  - Click on the “Full view” or “Limited (search-only)”option of the volume
  - Find the volume ID in the URL (all characters between “id=” and “;”, including all numbers and letters)



- Retrieve the volume’s metadata with an API call

Anatomy of an API call

- Types of volume ID: htid, OCLC, record number, lccn, isbn, issn
- A structured URL

https://catalog.hathitrust.org/api/volumes

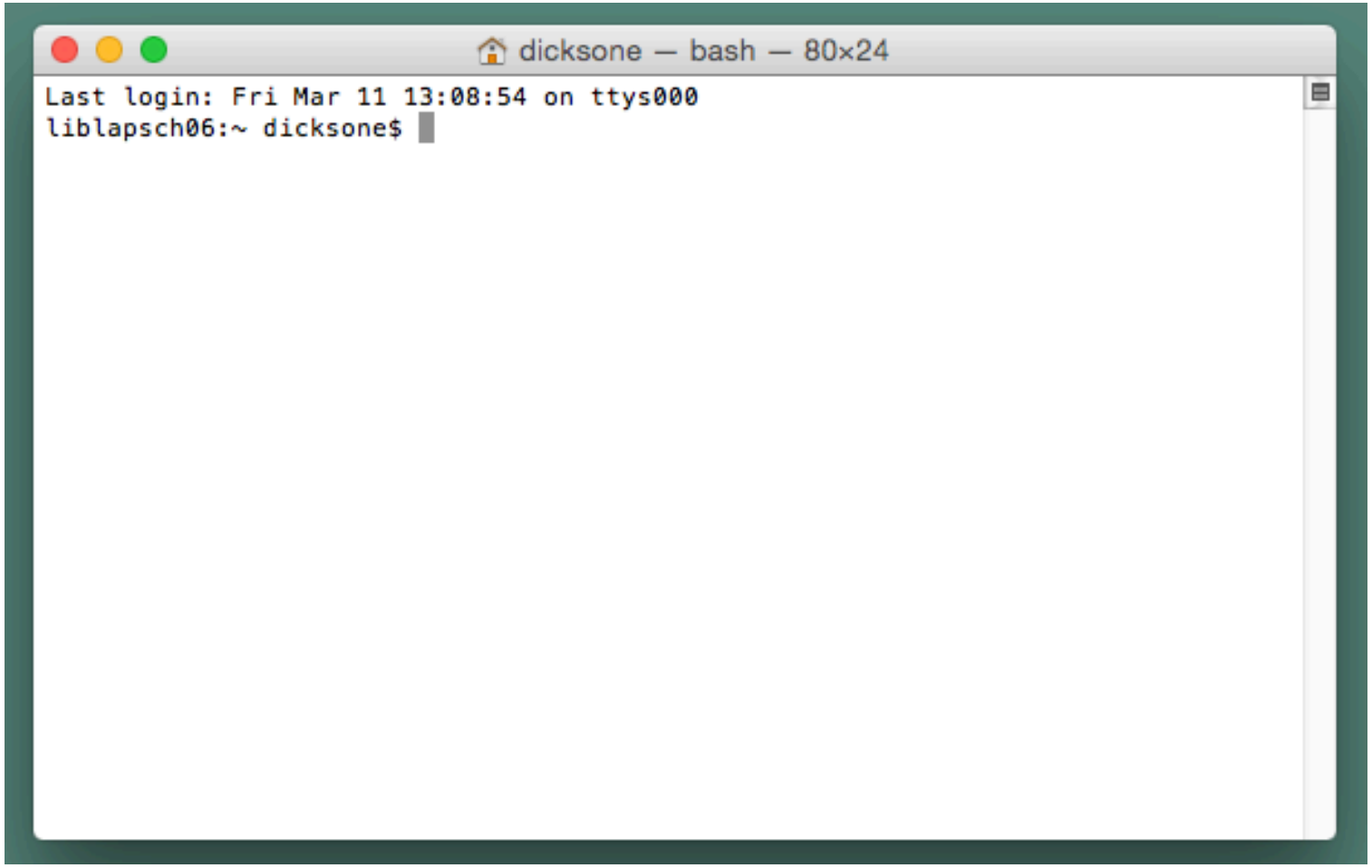
/brief **OR** /full

/<ID type>

/<ID.json>

https://catalog.hathitrust.org/api/volumes/full/htid/mdp.39015005337046.json

Intro to the command line



Background

- Command line interfaces
  - On Mac: Terminal
  - On Windows: Command Prompt
  - They do not speak the same language!
- Bash shell = Unix shell
  - Default on Mac
  - To use on Windows, install GitBash or Cygwin
    - *Bash on Ubuntu on Windows* comes installed on Windows 10

We're going to use the built-in Bash shell environment in PythonAnywhere

What’s PythonAnywhere?

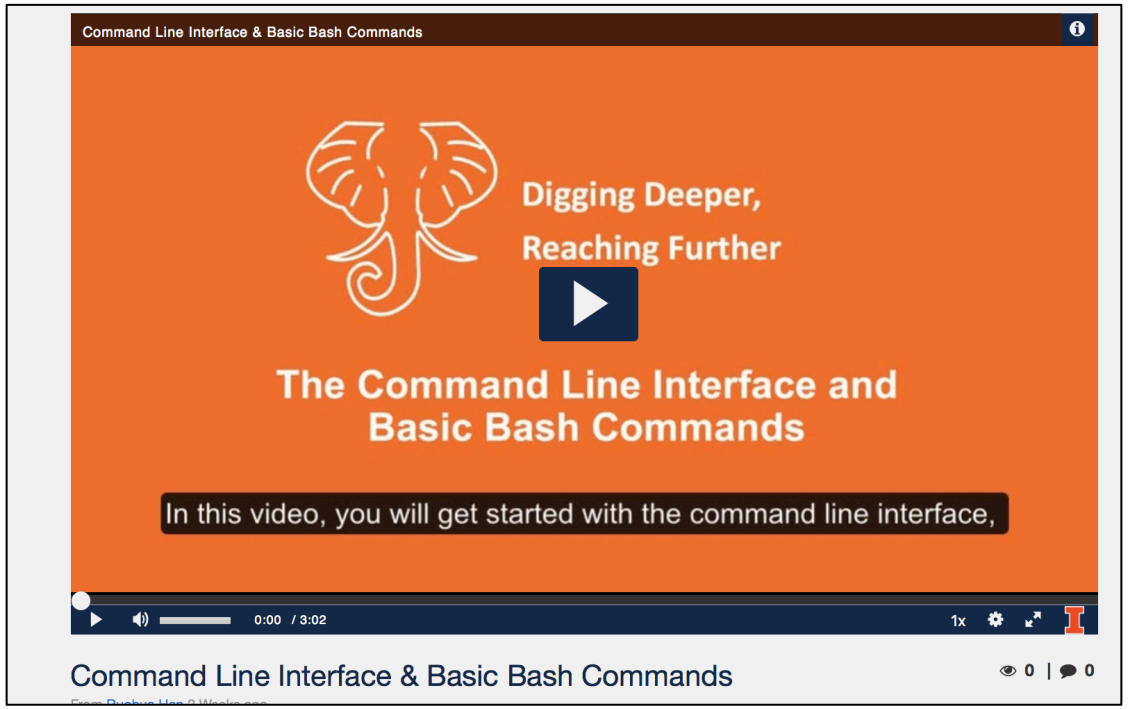
- PythonAnywhere is a browser-based programming environment
  - Write code
  - Save files
- Comes with a built-in Bash console
  - Run commands
- Does not interact with your local file system

Tips for working in a shell

- Directory = folder
- Case, spaces, and punctuation matter
- Tab to autocomplete a line
- Hit up/down arrow to see last commands entered
- Use “q” to quit viewing a file

Basic Bash commands

- Introductory video:



http://go.illinois.edu/ddrf-bash-video

Hands-on activity

☞ *See Handout p. 2*

Open a Bash console in PythonAnywhere, and practice the basic commands to:

- See your current directory

```
pwd
```

- Are there any files in your directory? List them:

```
ls
```

Hands-on activity

☞ *See Handout p. 2*

In your Bash console, practice the following:

- Unzip activity files (activity_files.zip) in your user directory:

```
unzip activity_files.zip
```

- Files were unzipped to a new directory called activity_files.

- Move the files from activity_files to your user directory. Substituting your Python Anywhere user name for *[your_username]*:

```
mv activity_files/* /home/[your_username]/
```

Hands-on activity

☞ *See Handout p. 2*

Use the ls command to check if files were successfully unzipped and moved:

- List files and directories under your current directory: `ls`

- List should include:

```
1930 README.txt activity_files.zip remove_tag.py
top_adjectives.py 1970 activity_files
remove_stopwords.py stopwords.txt word_count.py
```

Hands-on activity

☞ *See Handout p. 2*

More practice:

- Make a directory called ‘test’

```
mkdir test
```

- Change into the directory you just made, then back out of it

```
cd test
cd ..
```

Note: Please make sure you are back in your main directory after finishing the activity!

Command line tool: wget

- wget is a powerful way to transfer files from a server
  - i.e. web scraping
  - Follows links on the page and grabs that content, too
- You can specify **options**
  - Options tailor the command to do what you want
  - They are often written as single “flagged” letters, such as –d or –r
- Options for wget include:
  - `-l`: how far in the hierarchy you want to go
  - `--limit-rate=___`: sets the file transfer speed

Sample Reference Question

I'm a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I'd like to examine how concepts such as liberty change over time.

Approach:

Download text from an open resource, WikiSource, to start building a corpus.

Hands-on activity

☞ *See Handout p. 3*

In this activity, you will run a command to scrape text from a webpage version of George Washington's Fourth State of the Union Address.

Website Source: https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address

Web scraping

- What You Need:
  - PythonAnywhere: www.pythonanywhere.com
  - Website URL: https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address
  - Command:

```
wget -l 1 --limit-rate=20k
https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address --output-document=washington_4.txt
```

Web Source

Page Discussion Read Edit View history More Search Wikisource

George Washington's Fourth State of the Union Address

←George Washington's Third State of the Union Address

Fourth State of the Union Address (1792)
by George Washington

George Washington's Fifth State of the Union Address→

Delivered in Washington, D.C. on 8 November 1792.

related portals: State of the Union Speeches by United States Presidents.
sister projects: data item.

Fellow-Citizens of the Senate and of the House of Representatives:

It is some abatement of the satisfaction with which I meet you on the present occasion that, in felicitating you on a continuance of the national prosperity generally, I am not able to add to it information that the Indian hostilities which have for some time past distressed our Northwestern frontier have terminated.

You will, I am persuaded, learn with no less concern than I communicate it that reiterated endeavors toward effecting a pacification have hitherto issued only in new and outrageous proofs of persevering hostility on the part of the tribes with whom we are in contest. An earnest desire to procure tranquillity to the frontier, to stop the further effusion of blood, to arrest the progress of expense, to forward the prevalent wish of the nation for peace has led to strenuous efforts through various channels to accomplish these desirable purposes; in making which efforts I consulted less my own anticipations of the event, or the scruples which some considerations were calculated to inspire, than the wish to find the object attainable, or if not attainable, to ascertain unequivocally that such is the case.

A detail of the measures which have been pursued and of their consequences, which will be laid before you, while it will confirm to you the want of success thus far, will, I trust, evince that means as proper and as efficacious as could have been devised have been employed. The issue of some of them, indeed, is still depending, but a favorable one, though not to be despaired of, is not promised by anything that has yet happened.

In the course of the attempts which have been made some valuable citizens have fallen victims to their zeal for the public service. A sanction commonly respected even among savages has been found in this instance insufficient to protect from massacre the emissaries of peace. It will, I presume, be duly considered whether the occasion does not call for an exercise of liberality toward the families of the deceased.

It must add to your concern to be informed that, besides the continuation of hostile appearances among the tribes north of the Ohio, some threatening symptoms have of late been revived among some of those south of it.

https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address

Open shell and enter command

```
19:00 ~ $ wget -l 1 --limit-rate=20k https://en.wikisource.org/wiki/George_Washington%27s_Fourth_
State_of_the_Union_Address --output-document=washington_4.txt
```

Enter:

```
wget -l 1 --limit-rate=20k
https://en.wikisource.org/wiki/George_Washington%27s_Fou
rth_State_of_the_Union_Address --output-
document=washington_4.txt
```

Tip: Copy and paste the URL!

Run command successfully

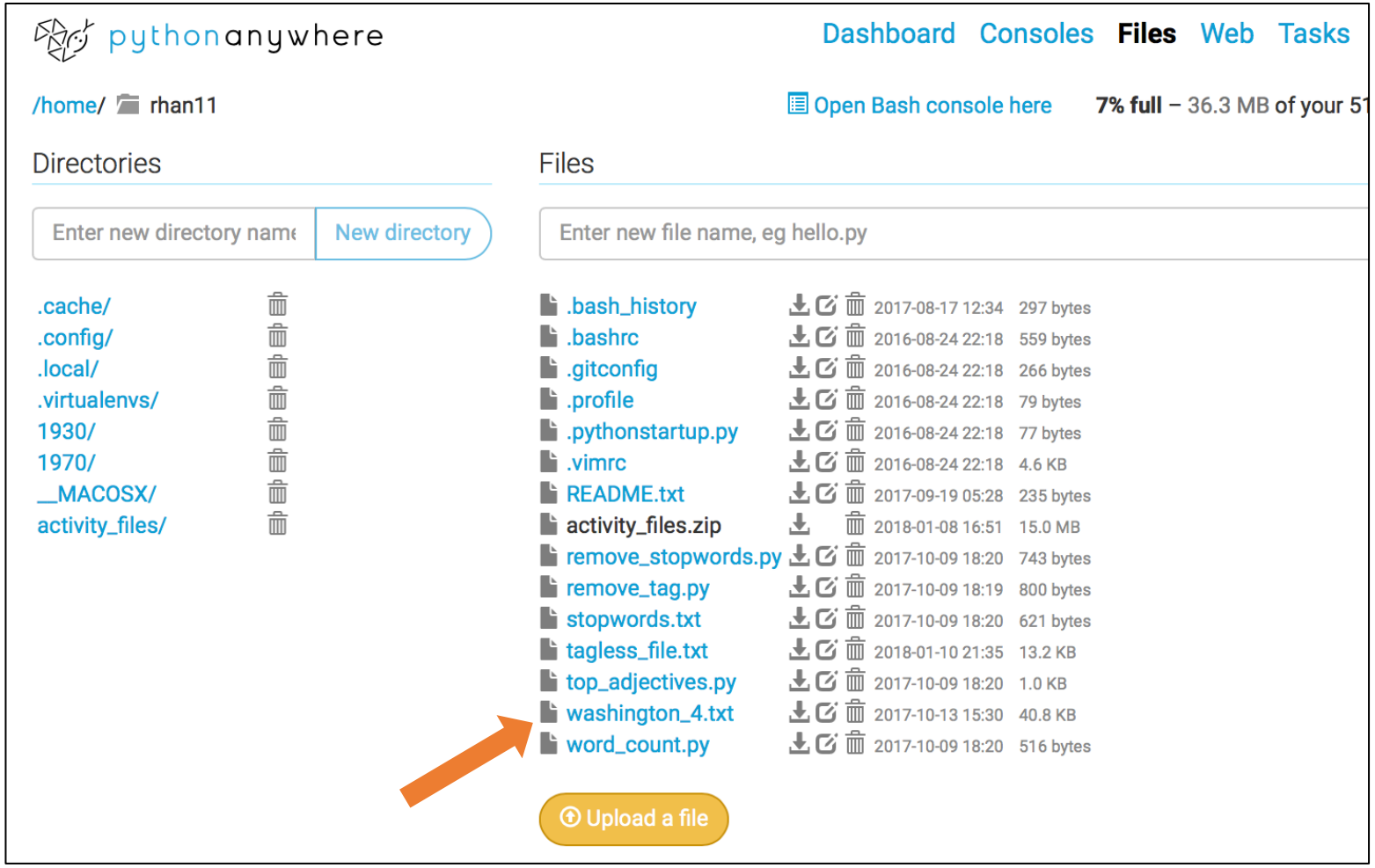
```
19:00 ~ $ wget -l 1 --limit-rate=20k https://en.wikisource.org/wiki/George_Washington%27s_Fourth_
State_of_the_Union_Address --output-document=washington_4.txt
--2017-07-06 19:00:46--  https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the
_Union_Address
Resolving proxy.server (proxy.server)... 10.0.0.82
Connecting to proxy.server (proxy.server)|10.0.0.82|:3128... connected.
Proxy request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: ‘washington_4.txt’

    [          <=>                                          ] 42,088      20.0KB/s   in 2.1s

2017-07-06 19:00:48 (20.0 KB/s) - ‘washington_4.txt’ saved [42088]

19:00 ~ $
```

View new file



View new file

/home/rhan11/washington_4.txt

Keyboard shortcuts: Normal Share Save Save as...

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>George Washington's Fourth State of the Union Address - Wikisource, the free online library</title>
<script>document.documentElement.className = document.documentElement.className.replace( /(^|\s)client-nojs(\s|$)/, "$1client-js
<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,
mw.user.tokens.set({"editToken":"+\\","patrolToken":"+\\","watchToken":"+\\","csrfToken":"+\\"});/*@nomin*/

});mw.loader.load(["site","mediawiki.page.startup","mediawiki.user","mediawiki.hidpi","mediawiki.page.ready","mediawiki.searchSu
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=ext.uls.interlanguage%7Cext.visualEditor.desktopArt
<script async="" src="/w/load.php?debug=false&amp;lang=en&amp;modules=startup&amp;only=scripts&amp;skin=vector"></script>
<meta name="ResourceLoaderDynamicStyles" content=""/>
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=ext.gadget.Site&amp;only=styles&amp;skin=vector"/>
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vector"/>
<meta name="generator" content="MediaWiki 1.31.0-wmf.15"/>
<meta name="referrer" content="origin-when-cross-origin"/>
<link rel="alternate" type="application/x-wiki" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_
<link rel="edit" title="Edit" href="/w/index.php?title=George_Washington%27s_Fourth_State_of_the_Union_Address&amp;action=edit"/
<link rel="shortcut icon" href="/static/favicon/wikisource.ico"/>
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikisource (en)"/>
<link rel="EditURI" type="application/rsd+xml" href="//en.wikisource.org/w/api.php?action=rsd"/>
<link rel="license" href="//creativecommons.org/licenses/by-sa/3.0/"/>
<link rel="alternate" type="application/atom+xml" title="Wikisource Atom feed" href="/w/index.php?title=Special:RecentChanges&am
<link rel="canonical" href="https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address"/>
<link rel="dns-prefetch" href="//login.wikimedia.org"/>
<link rel="dns-prefetch" href="//meta.wikimedia.org" />
<!--[if lt IE 9]><script src="/w/load.php?debug=false&amp;lang=en&amp;modules=html5shiv&amp;only=scripts&amp;skin=vector&amp;syn
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-George_Washington_s_Fourth_State_of_the_Union_Addr
        <div id="mw-head-base" class="noprint"></div>
        <div id="content" class="mw-body" role="main">
            <a id="top"></a>
```

What happened?

- You executed a command to scrape the text from the URL
- It grabbed everything on the page
- What's next?
  - Prepare the data for analysis

What might need to be removed from the files before analysis?

On your own

☞ *See Handout p. 4*

- *Edit the command to scrape George Washington's second State of the Union speech and save to a new text file.*
- *Can you view your scraped files using the* `less` *command?*

Other options to explore

- *Beautiful Soup*
  - https://www.crummy.com/software/BeautifulSoup/
  - Python-based web scraping
  - Lots of options, very powerful
- Write web scraping into a **script**
  - A script is a list of directions for your computer to follow
  - Direct computer to grab many files to retrieve content in bulk
  - Remember to time your requests!

Bulk HathiTrust data access

HT and HTRC datasets

Dataset	Kind of data	Description	Volumes available
HT Custom data request	Full text	Download page images and plain text OCR	Public domain
HTRC Extracted Features	Abstracted text and metadata	JSON files for eachof 15.7 million volumes in HathiTrust	All
HTRC Data API	Full text	Plain text OCR	All for HT members; public domain for others

Case Study: *Inside the Creativity Boom*

After creating final list of volumes:

- Used rsync to retrieve HTRC Extracted Features for the volumes
- Remember rsync is a command line utility, like wget, that transfers files between computers

Case Study: *Inside the Creativity Boom*

- What exactly is the HTRC Extracted Features dataset?

JSON files for each volume in the HathiTrust that researchers can download


```
{
    "id":"uc1.b3419888",
    "metadata":{
        "schemaVersion":"1.2",
        "dateCreated":"2015-02-12T13:30",
        "title":"Zoonomia = or The laws of organic life / by Erasmus Darwin.",
        "pubDate":"1809",
        "language":"eng",
        "htBibUrl":"http://catalog.hathitrust.org/api/volumes/full/htid/uc1.b3419888.json",
        "handleUrl":"http://hdl.handle.net/2027/uc1.b3419888",
        "oclc":"3679915",
        "imprint":"Thomas and Andrews, 1809."
    },
    "features":{
        "schemaVersion":"2.0",
        "dateCreated":"2015-02-20T23:58",
        "pageCount":616,
        "pages":[
```


Metadata: bibliographic; inferred

Data: words and word counts

Read and reflect

☞ *See Handout p. 4*

- Santa Barbara Statement on Collections as Data (Collections as Data National Forum, 2017) https://collectionsasdata.github.io/statement/
- Provides a set of high level principles to guide collections as data work

Read and reflect

☞ *See Handout p. 4*

- "With a few exceptions, cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use. Thinking about collections as data signals an intention to change that."
- "Any digital material can potentially be made available as data that are amenable to computational use. Use and reuse is encouraged by openly licensed data in non-proprietary formats made accessible via a range of access mechanisms that are designed to meet specific community needs."
- "Ethical concerns are integral to collections as data."
- **Principle 2 for collections as data:** "Collections as data development aims to encourage computational use of digitized and born digital collections."

-- Santa Barbara Statement on Collections as Data

Read and reflect

☞ *See Handout p. 4*

- *Does your library provide access to digital collections as data?*
- *How so? Why not? How could it?*

Questions?

References

- Collections as Data National Forum. (2017, March 3). The Santa Barbara Statement on Collections as Data. Retrieved August 16, 2017, from https://collectionsasdata.github.io/statement/ .