Digging Deeper, Reaching Further

Module 3: Working with Textual Data

In this module we'll…

- Think about what happens when text is data

  → *Understand best practice in the field*

- Consider common steps to cleaning and preparing text data

  → *Make recommendations to researchers*

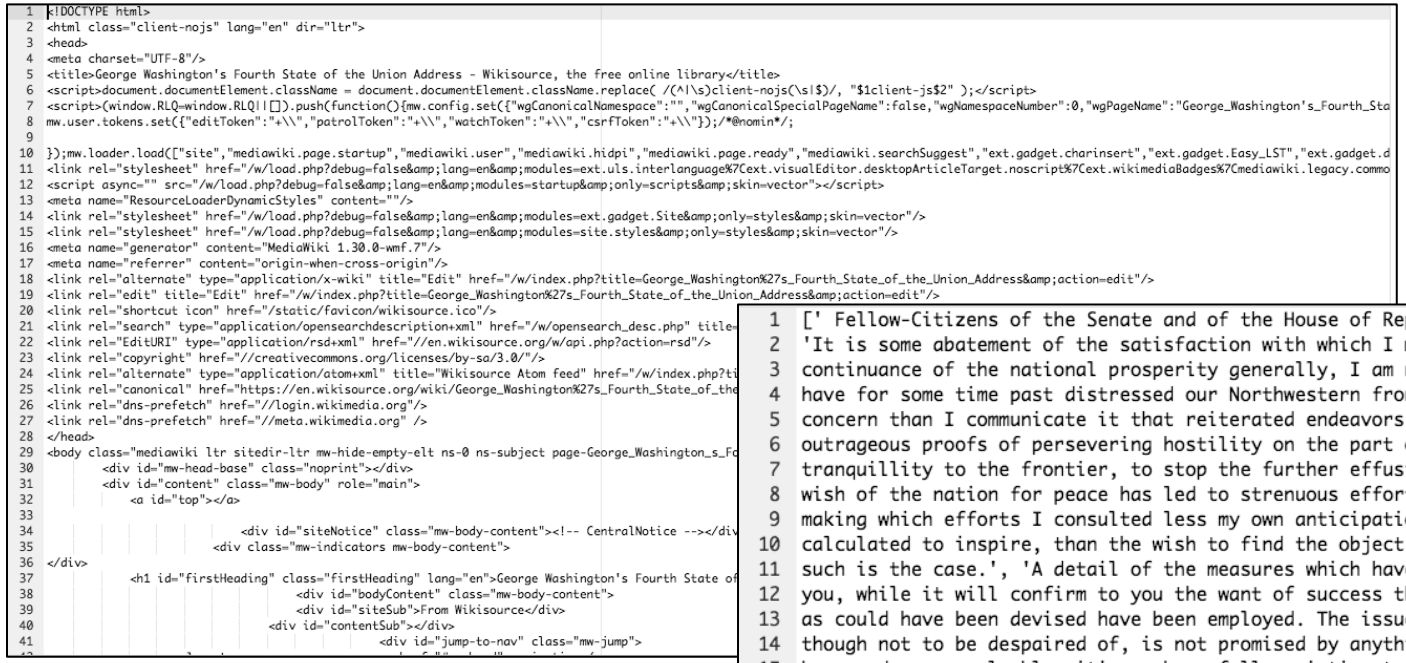
- Practice data cleaning on the speeches we scraped

  → *Gain experience with Python for working with data*

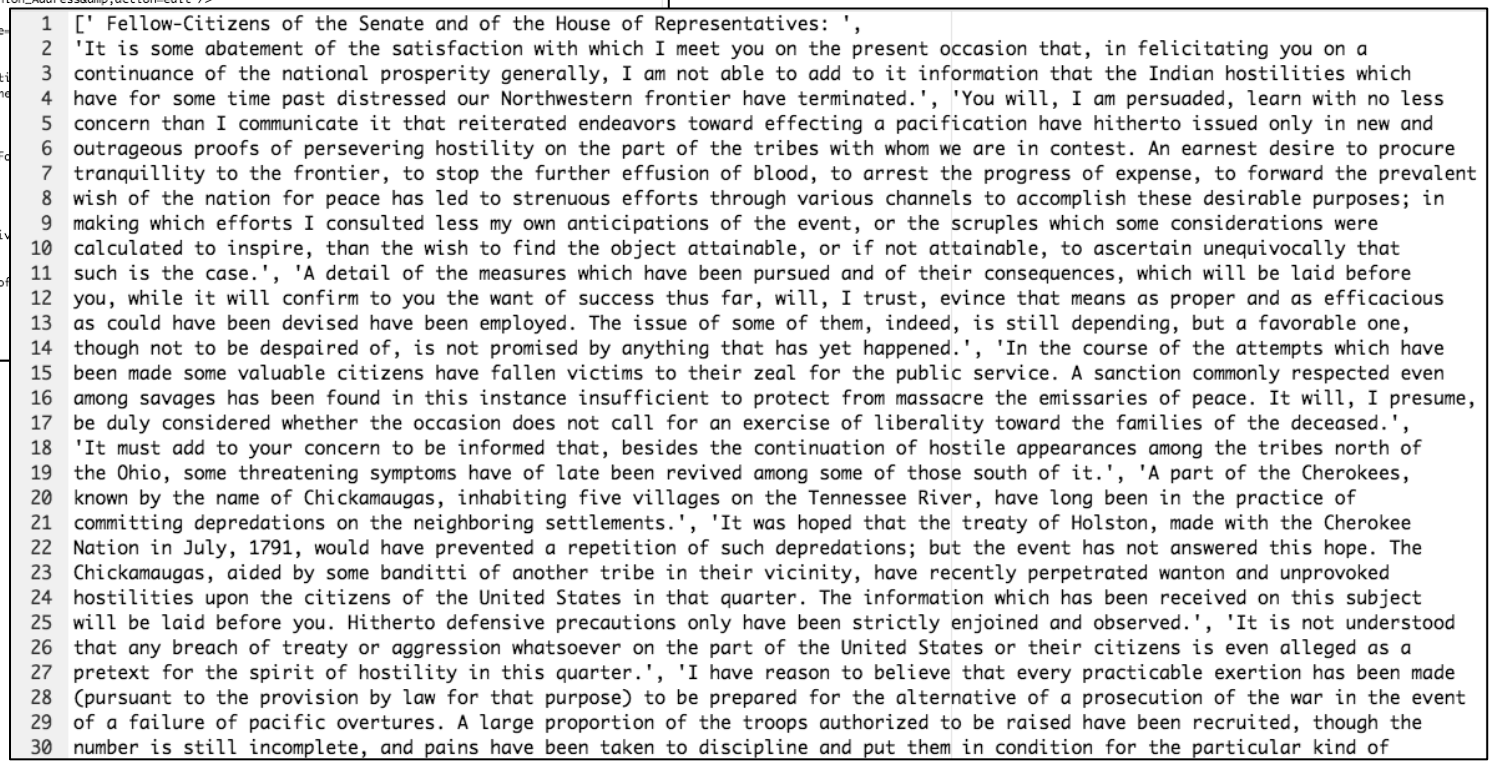
- Learn how Sam prepared his *Creativity Corpus* for analysis

  → See *how one scholar data prepared data*

Where we'll end up



Go from HTML-formatted text...



...to cleaner text ready for analysis.

Humanities data

- Data is material generated or collected while conducting research
- Examples of humanities data:
  - Citations
  - Code/Algorithms
  - Databases
  - Geospatial coordinates

Can you think of others?

Definition adapted from: NEH, https://www.neh.gov/files/grants/data_management_plans_2018.pdf

Text as data

- Data quality
  - Clean vs. dirty OCR
  - HathiTrust OCR is dirty (uncorrected)
- Analyzed by corpus/corpora
  - Text corpus: a digital collection OR an individual's research text dataset
  - Text corpora: "bodies" of text
- Text decomposition/recomposition (Rockwell, 2003)
  - Cleaning data involves discarding data
  - Prepared text may be illegible to the human reader

Preparing data

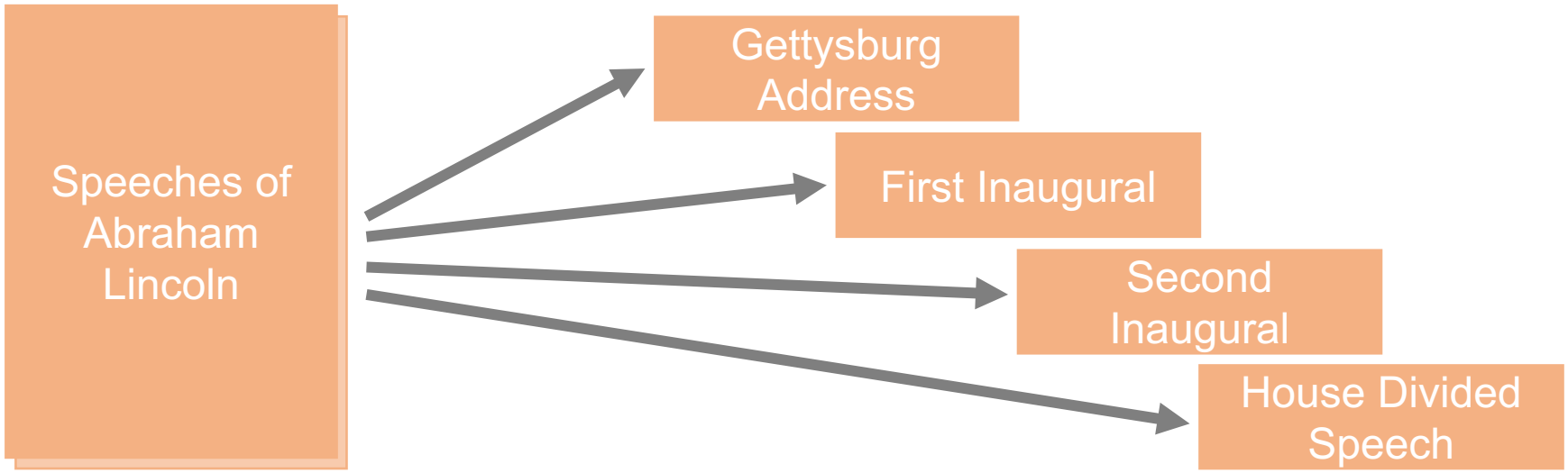
A researcher may:

- Correct OCR errors
- Remove title, header information
- Remove html or xml tags
- Split or combine files
- Remove certain words, punctuation marks
- Lowercase text
- Tokenize the words

Key concepts

Chunking text

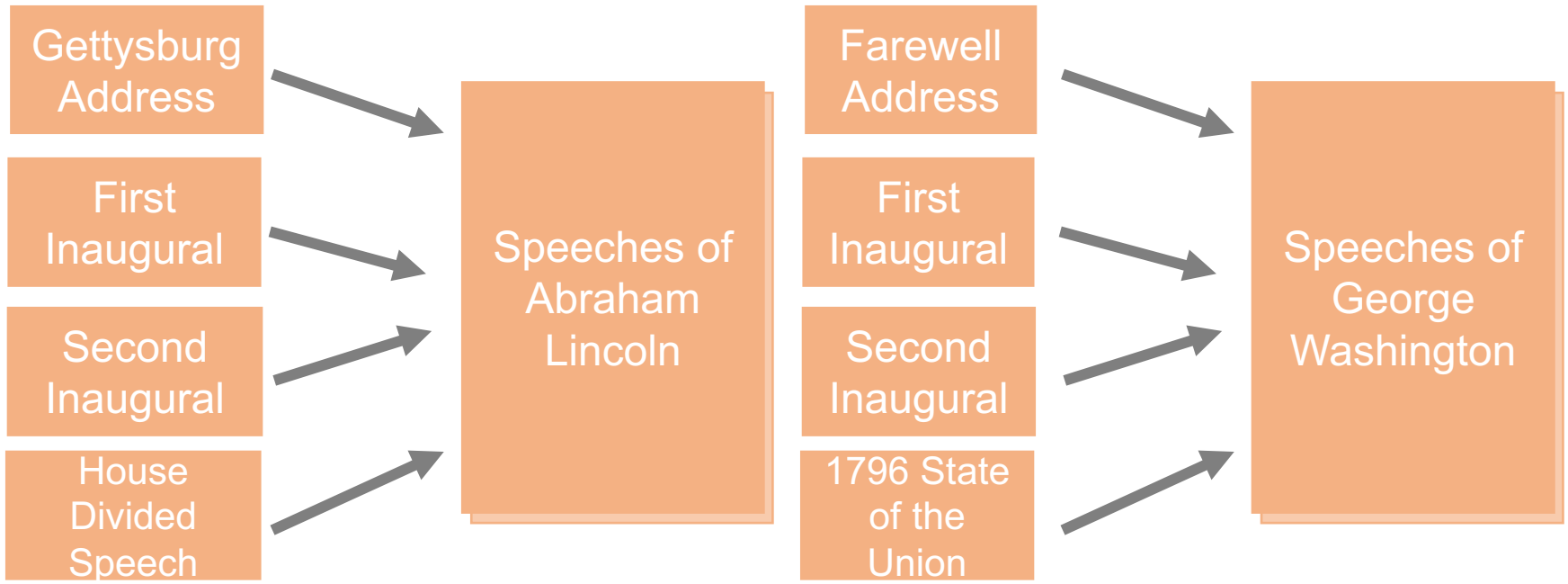
Splitting text into smaller pieces before analysis. May be divided by paragraph, chapter, or a chosen number of words (e.g. 1000 word chunks).



Key concepts

Grouping text

Combining text into larger pieces before analysis.



Key concepts

Tokenization

Breaking text into pieces called tokens. Often certain characters, such as punctuation marks, are discarded in the process

[four], [score], [and], [seven], [years], [ago], [our], [fathers], [brought], [forth], [on], [this], [continent], [a], [new], [nation], [conceived], [in], [liberty], [and], [dedicated], [to], [the], [proposition], [that], [all], [men], [are], [created], [equal]

Preparing data

- Preparation affects results
  - Amount of text and size of chunks
  - Which stop words removed; which characters are included
  - Whether to lowercase and normalize words
- Preparation for analysis takes time, effort
  - This is where scripting becomes useful!

Hands-on Activity

☞ *See Handout pp. 1-2*

- In groups of 2 or 3, assign each person several of the text preparation actions seen in the table to the right (Denny and Spirling, 2017).
- Read the descriptions. Then take turns explaining each to your group.

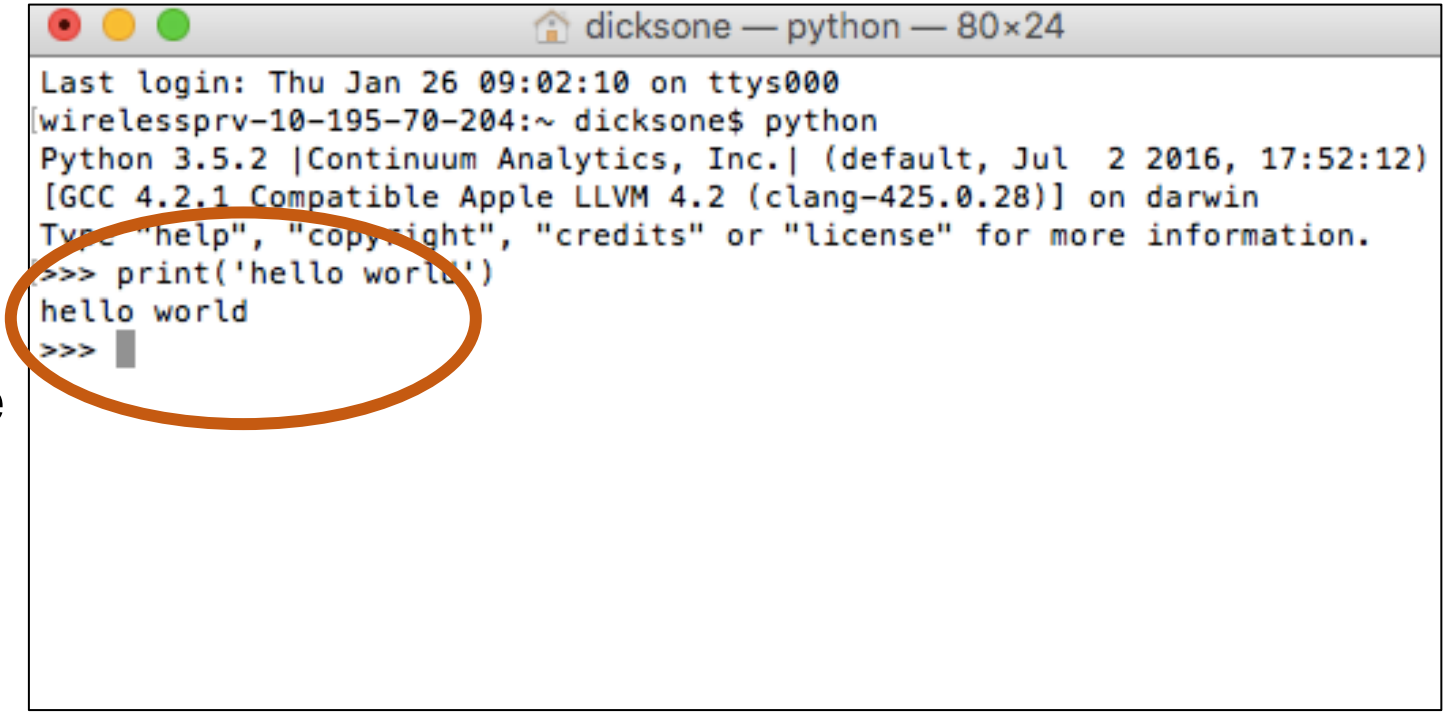
Term
Punctuation
Numbers
Lowercasing
Stemming
Stopword Removal
n-gram Inclusion
Infrequently Used Terms

Scripting with Python

- Python is a scripting language
- Good for working with data
  - Interpreted language → follows step-by-step directions
- Relatively straightforward syntax
  - Avoids excess punctuation

Using Python: Interactive programming

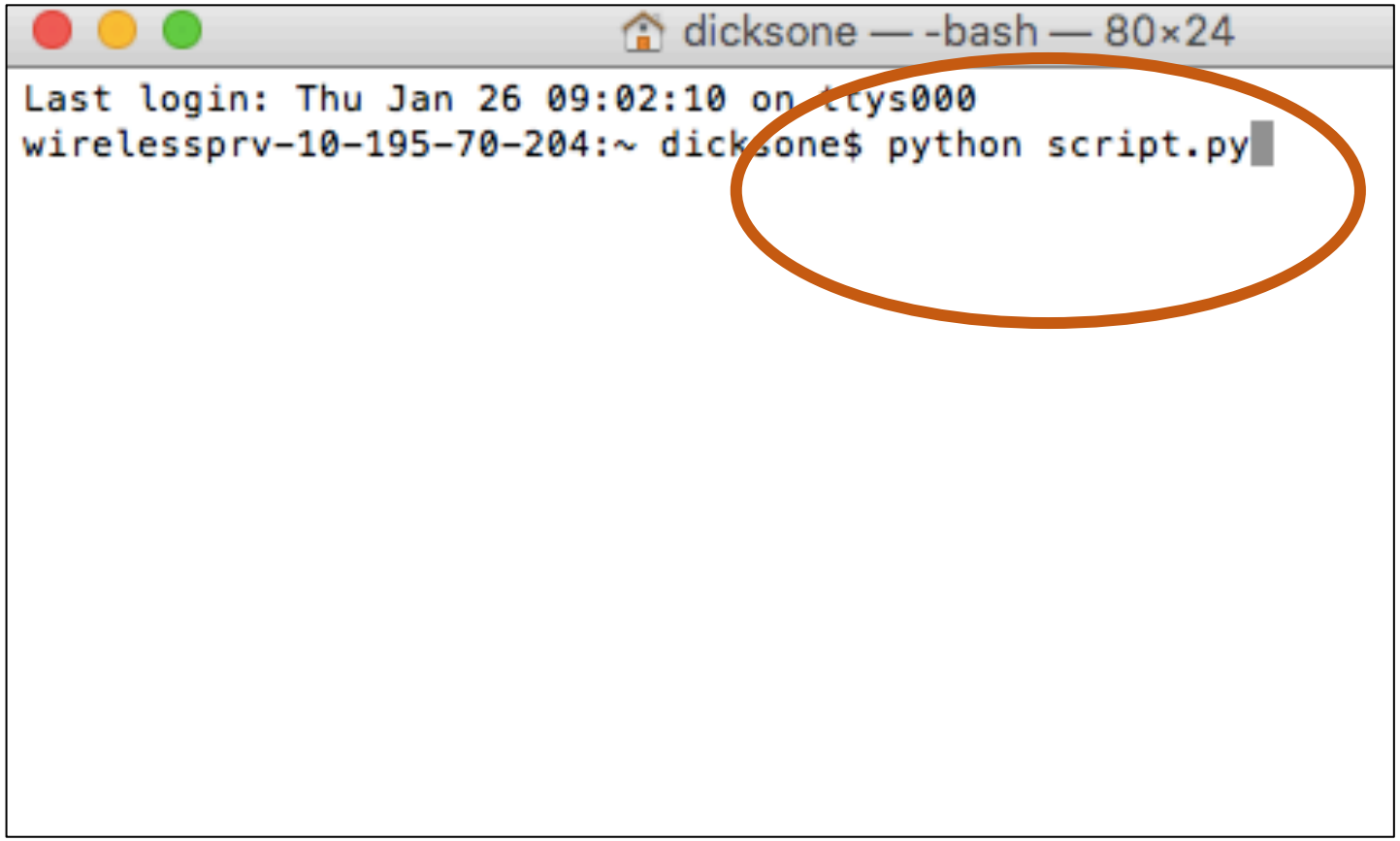
- Using a **python interpreter**
- Run each step at the prompt
- If you enter “Python” on the command line, you will start the interpreter
- *We aren’t using it today!*



Using Python: Write & run scripts

- Scripts are directions for your computer to follow
- Save the script as a file ending in .py
- On the command line, run the script

→This is how we'll do it!



Anatomy of running a Python script

```
python <script filename> <arguments>
```

1. Tells the computer to be ready for Python
2. Directs which file to run
3. The file may ask for additional information to be given at runtime, called arguments

Example: `python myscript.py newfile.txt`

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:

Use a Python script to prepare the text data we scraped from Wikisource.

Hands-on activity

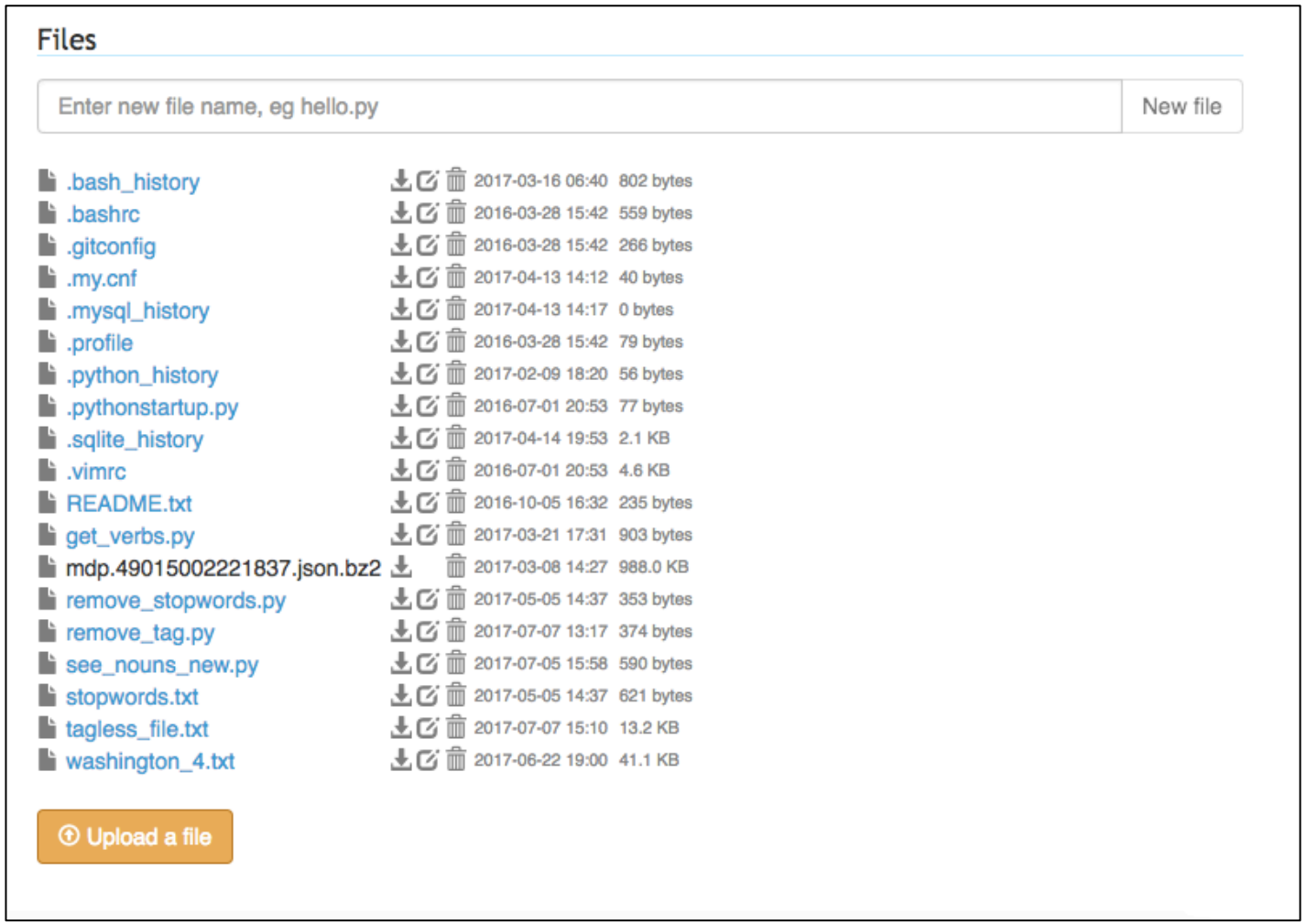
☞ *See Handout p. 3*

In this activity, you will run a Python script that will remove the HTML tags from George Washington's *Fourth State of the Union Address* you scraped earlier from Wikisource.

Strip HTML tags

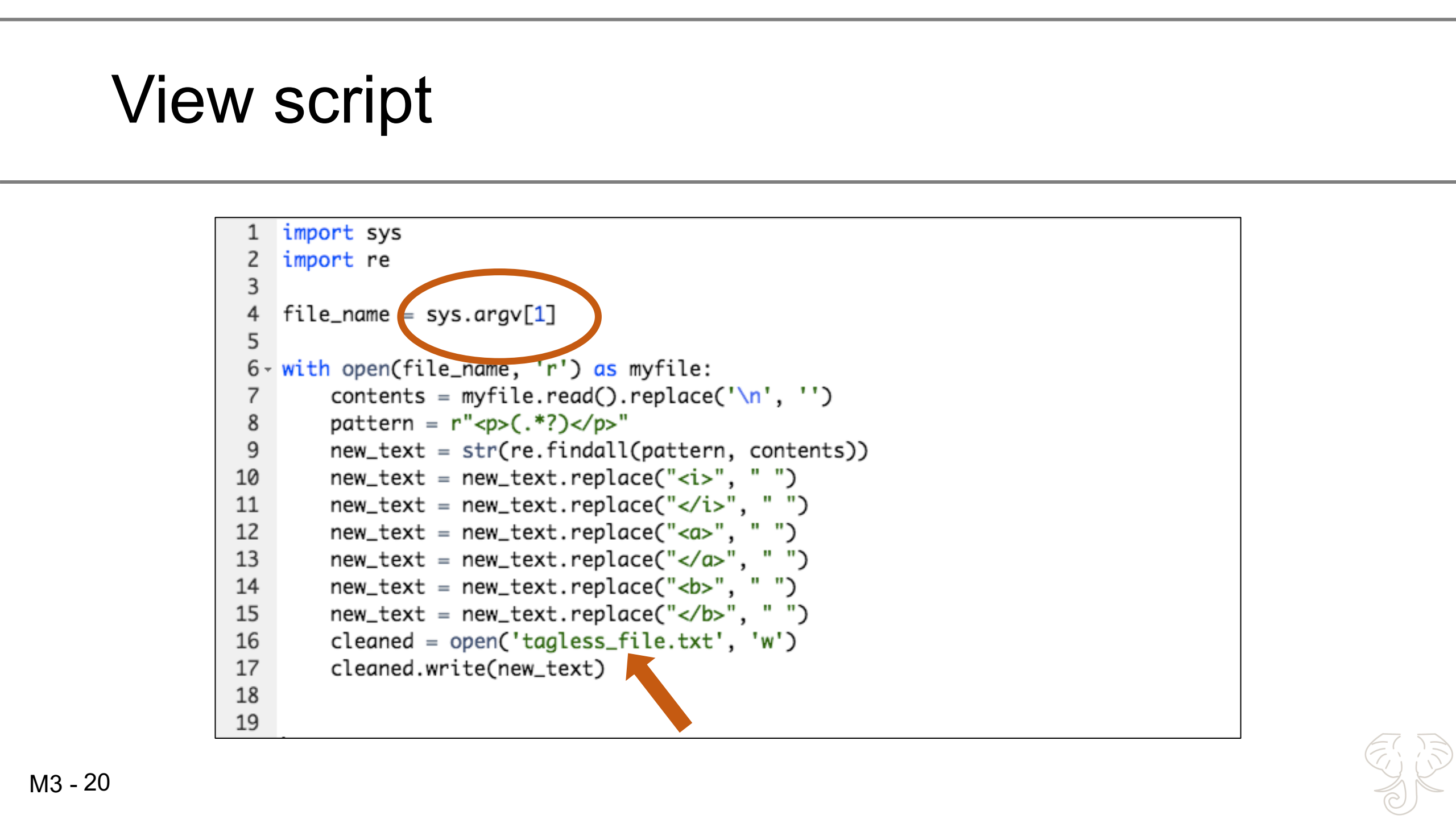
- What You Need:
  - PythonAnywhere bash console
  - washington_4.txt
  - remove_tag.py script

Go to Files to view script

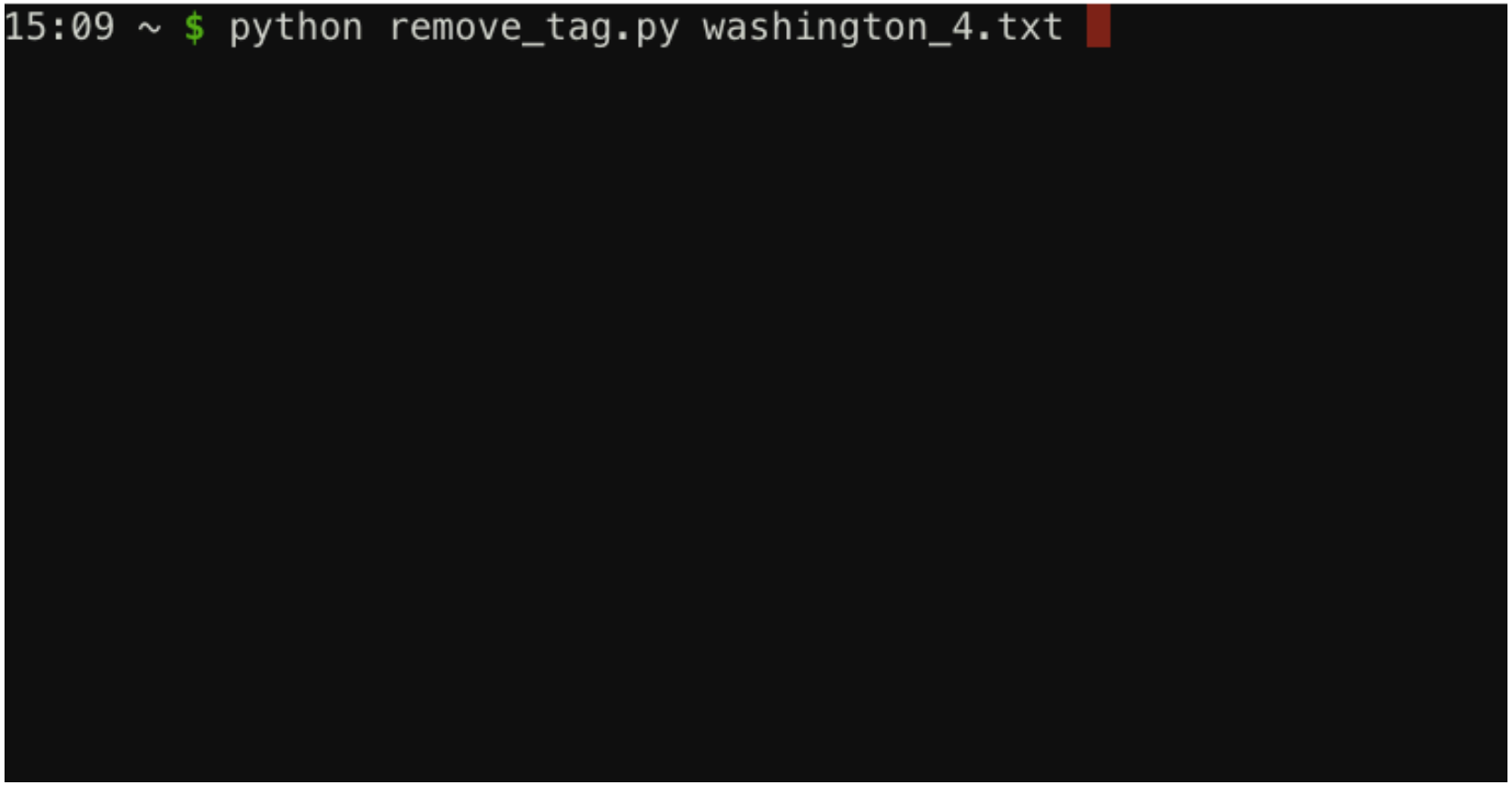


View script

```
1  import sys
2  import re
3
4  file_name = sys.argv[1]
5
6  with open(file_name, 'r') as myfile:
7      contents = myfile.read().replace('\n', '')
8      pattern = r"<p>(.*?)</p>"
9      new_text = str(re.findall(pattern, contents))
10     new_text = new_text.replace("<i>", " ")
11     new_text = new_text.replace("</i>", " ")
12     new_text = new_text.replace("<a>", " ")
13     new_text = new_text.replace("</a>", " ")
14     new_text = new_text.replace("<b>", " ")
15     new_text = new_text.replace("</b>", " ")
16     cleaned = open('tagless_file.txt', 'w')
17     cleaned.write(new_text)
18
19
```



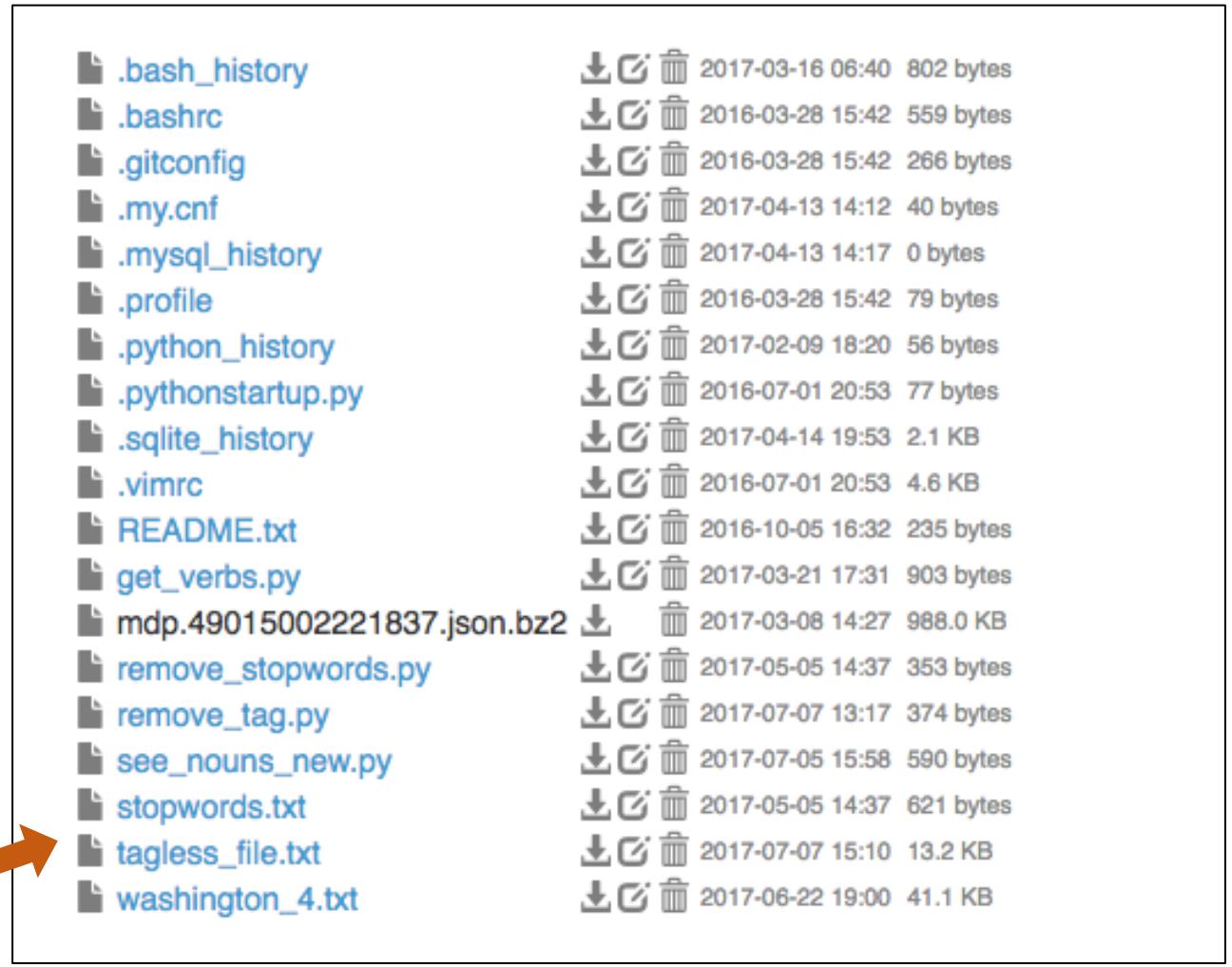
Open shell and run script



Enter command:

```
python remove_tag.py washington_4.txt
```

Find new file



Review results

```
[' 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.', 'A part of the Cherokees,
known by the name of Chickamaugas, inhabiting five villages on the Tennessee River, have long been in the practice of
committing depredations on the neighboring settlements.', 'It was hoped that the treaty of Holston, made with the Cherokee
Nation in July, 1791, would have prevented a repetition of such depredations; but the event has not answered this hope. The
Chickamaugas, aided by some banditti of another tribe in their vicinity, have recently perpetrated wanton and unprovoked
hostilities upon the citizens of the United States in that quarter. The information which has been received on this subject
will be laid before you. Hitherto defensive precautions only have been strictly enjoined and observed.', 'It is not understood
that any breach of treaty or aggression whatsoever on the part of the United States or their citizens is even alleged as a
pretext for the spirit of hostility in this quarter.', 'I have reason to believe that every practicable exertion has been made
(pursuant to the provision by law for that purpose) to be prepared for the alternative of a prosecution of the war in the event
of a failure of pacific overtures. A large proportion of the troops authorized to be raised have been recruited, though the
number is still incomplete, and pains have been taken to discipline and put them in condition for the particular kind of
```

What happened

- Looks for < > to denote html tags
  - Using regular expressions and Python's replace()
- Writes the characters that are NOT html tags and their contents to a file
- Note: this script is simple, but not the most robust way to remove tags

On your own

☞ *See Handout p. 3*

Can you run a script to remove stop words?

- The script is called remove_stopwords.py
- Hint: The remove_stopwords.py script requires three arguments:
  - The input filename of your tagless file
  - The filename of the list of stop words
  - The output filename you make up
- Edit the stopwords.txt file to customize your list

Case Study: *Inside the Creativity Boom*

After downloading the Extracted Features data for the relevant volumes, used scripting to:

- Narrow corpus to individual pages that contained creativ*
  - Discarded all other pages
- Discard certain tokens such as pronouns and conjunctions
  - To keep only to most ”meaningful” terms

Read and Reflect…

☞ *See Handout p. 4*

- Passage from "Against Cleaning" by Katie Rawson and Trevor Muñoz
- They suggest a strategy for dealing with humanities data:
  - Shared authority control across data sets
  - Indexes for nuance
  - Tidy, not clean data

Read and Reflect…

☞ *See Handout p. 4*

"When humanities scholars recoil at data-driven research, they are often responding to the reductiveness inherent in this form of scholarship. This reductiveness can feel intellectually impoverishing to scholars who have spent their careers working through particular kinds of historical and cultural complexity… From within this worldview, data cleaning is then maligned because it is understood as a step that inscribes a normative order by wiping away what is different. The term "cleaning" implies that a data set is 'messy.' "Messy" suggests an underlying order. It supposes things already have a rightful place, but they're not in it—like socks on the bedroom floor rather than in the wardrobe or the laundry hamper."

- "Against Cleaning" (Rawson and Muñoz, 2016)

Discussion

- *What does this excerpt suggest about the nuances of data cleaning?*
- *What does “clean” imply?*
- *How might you talk to researchers on your campus who would be uncomfortable with the idea of clean v. messy data?*

Questions?

References

- Denny, M. J. and Spirling, A. (2017). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. https://ssrn.com/abstract=2849145 .
- National Endowment for the Humanties. (2017) *Data Management Plans for NEH Office of Digital Humanities Proposals and Awards*. Retrieved October 1, 2017, from https://www.neh.gov/files/grants/data_management_plans_2018.pdf .
- Rawson, K., & Muñoz, T. (2016). Against Cleaning. Retrieved August 16, 2017, from http://curatingmenus.org/articles/against-cleaning/ .
- Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing*, *18*(2), 209–219. https://doi.org/10.1093/llc/18.2.209 .