Scaling up: Using web scraping to augment research in the humanities

John R. Gallagher
Assistant Professor
University of Illinois, Urbana-Champaign
English Department

Schedule for Workshop

- Introduction (9:00-9:05 am)
- Definitions and significance of web-scraping (Facebook example) (9:05-9:15)
- The uses of web-scraping (9:15-9:30)
- Ethical, political, and economic implications (9:30-9:45)
- Guided tutorial of Wikipedia and YouTube (9:45-10:15)
- Data analysis via Google sheets (10:15-10:30)
- Brainstorming uses and deployment for current research projects (10:30-10:50)
- Web-scraping tools, resources, and advice (10:50-11:00)
  - See http://publish.illinois.edu/johnrgallagher/scrape/

Web-scraping conceptually

Definition

- Formally: Pulling information from a website into a format
- Informally: copying text from a webpage into a text or CSV file

Rose to prominence in the late 2000s

Significance

- Allows digital researchers to scale up the scope of their inquiries
- Reduces human error(s)
- Allows research to see broader patterns outside of initial or anecdotal perspectives
- Archive creation

Different types of web-scraping

- Copy and paste!
- Using XML query language to pull information based on (X)HTML tags (what we're doing today!)
- APIs (application programming interface…not quite coding but typically requires some knowledge)
- Computer coding to engage full automation (Python or R are most common)

The uses of web-scraping

How do I use web-scraping?

Two key elements when turning to web-scraping

-A question to answer
-A mentor to guide

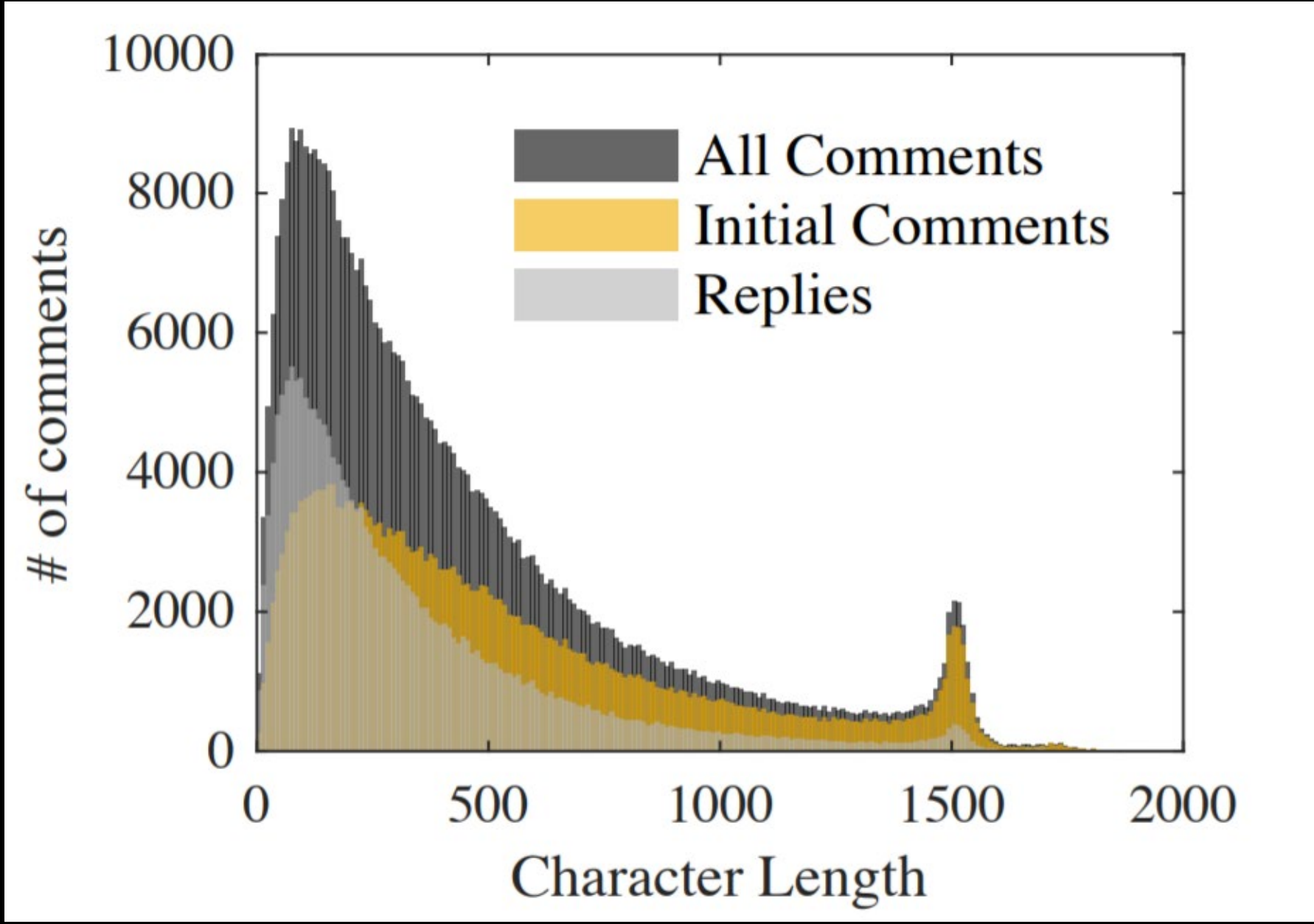
A personal story

:	No, that's 1984 ;)	https://www.facebook.com/groups/
:	Big Brother Godzilla!	https://www.facebook.com/groups/
:	I don't necessarily buy into al	https://www.facebook.com/groups/
:	This is an interesting segment	https://www.facebook.com/groups/
:	That's very, very disturbing. H	https://www.facebook.com/groups/
:	We can start with this one - b	https://www.facebook.com/groups/
:	Yeah this is the Presidency th	https://www.facebook.com/groups/
:	Remember Clinton?	https://www.facebook.com/groups/
:	Obama's the first one I didn't	https://www.facebook.com/groups/

Total Activity	
Total initial posts and comments	5622
Total by Tracy	1468 (26.1%)
Tracy's Initial Texts	
Total initial posts by entire group:	847
Total initial posts by Tracy:	238 (28.1%)
Tracy's Comments:	
Total number of comments	4775
Total number of Tracy's comments	1230 (25.7%)
Tracy's Initial Texts:	
Initial Texts	238
Total Comments	1568
Comments by other writers	1058
Comments by Tracy	509
Comments per post	6.6
Comments by other writers per post	4.4
Comments by Tracy per post	2.2
Average character length (ACL) of comments	
ACL	282
ACL of Monroe's comments	377
ACL of other's comments	250

Using web-scraping for large scale projects

I’m working with a web-scraped dataset of ~450,000 comments from the NYT (using their API)


10000
8000
6000
4000
2000
0
# of comments
All Comments
Initial Comments
Replies
0
500
1000
1500
2000
Character Length

Ethical, political, and economic considerations

Ethical

- Web-scraping can be incomplete
- All automated web-scraping “pings” servers (makes packet requests)
  - If too many pings are made, this can simulate a distributed denial of service (DDoS) attack
  - Requests need to be timed
- Which information should be scraped? Could it harm individuals?
- The way we look at a page is very different from the way it is built—web-scraping requires looking at the building blocks and not the pleasant aesthetics
- What does it mean if a website blocks web-scraping? Or has rules against it?
- Who owns the data if you web-scrape it?

Political

- Web-scraping can create archives for marginalized groups
- Create records-of-proof when websites are constantly changed (for instance, government priorities can change and emphases will differ)
- Identify group mentalities and majority viewpoints (including manipulation)
- Companies choose to treat correlation as causation due to massive amounts of data
- Who owns the data if you web-scrape it?
- What does it mean to scrape? Is it equivalent of “viewing”?

Economic

- Web-scraping can…
  - automate processes that are expensive
  - be outsourced
  - Save time and improve accuracy
- Some websites have disclaimers that web-scrapers are not allowed according to terms & conditions
- Web-scraping avoids the black boxing of past information (social media companies do not collate the past and intentionally try to manipulate the present/future in their interfaces in order to make money)

Guided tutorial

See webpage & worksheet

http://publish.illinois.edu/johnrgallagher/scrape/

Brainstorming web-scraping applications for research projects

Conceptual considerations

- What is your current research? How could it benefit from automating certain processes?
- What kind of text(s) are you interested in collecting?
- Practical elements, including time to devote (see other side)

Technical considerations

- What websites do you currently study? What is the structure?
- Who can help you learn advanced techniques? Who can be a mentor?
- What resources does your institution offer?

Resources

http://publish.illinois.edu/johnrgallagher/scrape/