

Web Crawling

February 4, 2020

Data Science CSCI 1951A

Brown University

Instructor: Ellie Pavlick

HTAs: Josh Levin, Diane Mutako, Sol Zitter

Announcements

- Sign the collab policy! Do it literally right now...it takes 2 seconds
- Final project pitches due next Monday (2/10)
- If you still need a group.....
- If you have a group but it is the wrong size...
- Thursday's lecture—half day!
- Questions about any of this?

Clicker Question!

Did you sign the collaboration policy?

(a) Yes, of course.

(b) No, because clicking buttons is too much effort, and also I don't mind if I don't receive grades for my assignments.

Today

- Code-along!
- Legal 101

Code-along!



```
html_dump = BeautifulSoup ( html_doc, 'html.parser' )
```

Legal 101

Legal 101

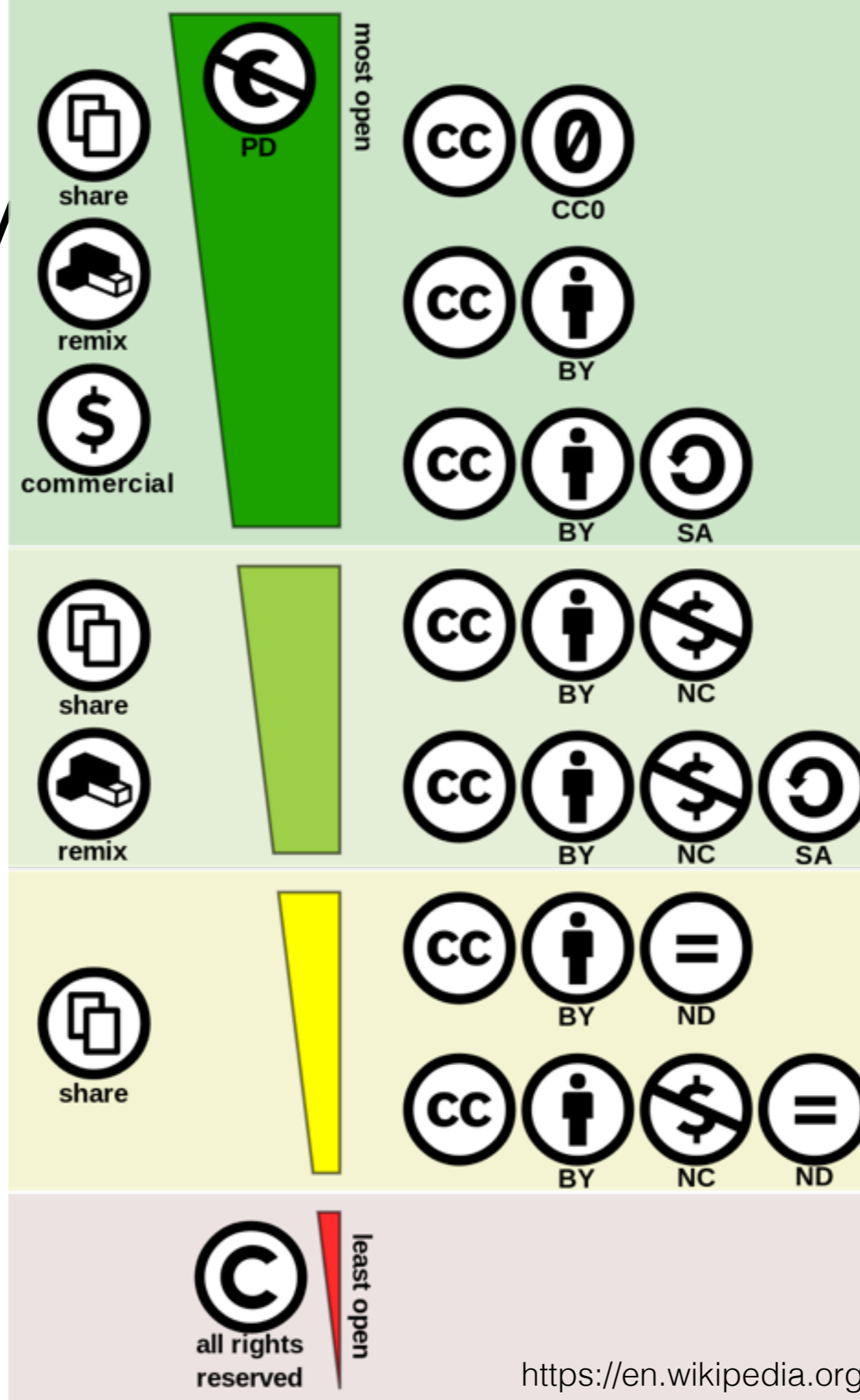
- First, in case its not obvious, I am not a lawyer...
- Licensing—things to look out for
- Privacy/Ethics—things to think about

Creative Commons Licenses

- Attribution (by): All CC licenses require that others who use your work in any way must give you credit the way you request, but not in a way that suggests you endorse them or their use. If they want to use your work without giving you credit or for endorsement purposes, they must get your permission first.
- ShareAlike (sa): You let others copy, distribute, display, perform, and modify your work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get your permission first.
- NonCommercial (nc): You let others copy, distribute, display, perform, and (unless you have chosen NoDerivatives) modify and use your work for any purpose other than commercially unless they get your permission first.
- NoDerivatives (nd): You let others copy, distribute, display and perform only original copies of your work. If they want to modify your work, they must get your permission first.
- Public Domain (CC0): You waives all rights that are legally possible to waive.

Creative

enses



WIKIPEDIA

The Free Encyclopedia

English

5 769 000+ articles

日本語

1 132 000+ 記事

Deutsch

2 249 000+ Artikel

Русский

1 515 000+ статей

Español

1 494 000+ artículos

Français

2 065 000+ articles

Italiano

1 486 000+ voci

中文

1 036 000+ 條目

Português

1 012 000+ artigos

Polski

1 312 000+ hasel

EN ▾🔍

🗺️ Read Wikipedia in your language ▾



Wikipedia is hosted by the [Wikimedia Foundation](#), a non-profit organization that also hosts a range of other projects.



Commons
Freely usable photos & more



Wikivoyage
Free travel guide



Wiktionary
Free dictionary



Wikipedia apps are now available:
[Download for iOS on the App Store](#)
[Download for Android on Google Play](#)
[View full list of available Wikipedia apps](#)



Wikibooks
Free textbooks



Wikinews
Free news source



Wikidata
Free knowledge base



Wikiversity
Free course materials



Wikiquote
Free quote compendium



MediaWiki
Free & open wiki application



Wikisource
Free library



Wikispecies
Free species directory



Meta-Wiki
Community coordination & documentation

WIKIPEDIA

The Free Encyclopedia

English

5 769 000+ articles

日本語

1 132 000+ 記事

Deutsch

2 249 000+ Artikel

Русский

1 515 000+ статей

Español

1 494 000+ artículos

Français

2 065 000+ articles

Italiano

1 486 000+ voci

中文

1 036 000+ 條目

Português

1 012 000+ artigos

Polski

1 312 000+ hasel

EN ▾🔍

🗺️ Read Wikipedia in your language ▾



Wikipedia is hosted by the [Wikimedia Foundation](#), a non-profit organization that also hosts a range of other projects.



Commons
Freely usable photos & more



Wikivoyage
Free travel guide



Wiktionary
Free dictionary



Wikipedia apps are now available:
[Download for iOS on the App Store](#)
[Download for Android on Google Play](#)
[View full list of available Wikipedia apps](#)



Wikibooks
Free textbooks



Wikinews
Free news source



Wikidata
Free knowledge base



Wikiversity
Free course materials



Wikiquote
Free quote compendium



MediaWiki
Free & open wiki application



Wikisource
Free library



Wikispecies
Free species directory



Meta-Wiki
Community coordination & documentation

WIKIPEDIA

The Free Encyclopedia

English
5 769 000+ articles

日本語
1 132 000+ 記事

Deutsch
2 249 000+ Artikel

Русский
1 515 000+ статей

Español
1 494 000+ artículos

Français
2 065 000+ articles

Italiano
1 486 000+ voci

中文
1 036 000+ 條目

Português
1 012 000+ artigos

Polski
1 312 000+ hasel

EN ▾

Read Wikipedia in your language ▾



Wikipedia is hosted by the [Wikimedia Foundation](#), a non-profit organization that also hosts a range of other projects.



Commons
Freely usable photos & more



Wikivoyage
Free travel guide



Wiktionary
Free dictionary

Wikipedia apps are now available:



Wikibooks



Wikinews



Wikidata

This page is available under the [Creative Commons Attribution-ShareAlike License](#) • [Terms of Use](#)



wikiversity
Free course materials



Wikiquote
Free quote compendium



MediaWiki
Free & open wiki application



Wikisource
Free library



Wikispecies
Free species directory



Meta-Wiki
Community coordination & documentation



Warm Up with Ree's Simple, Perfect Chili



Warm Up with Ree's Simple, Perfect Chili



Warm Up with Ree's Simple, Perfect Chili

ons ^ | See all Scripps Networks Digital ^ | © 2019 Television Food Network, G.P. All rights reserved.

Twitter

- “Get the user’s express consent before you do any of the following...Republish Twitter Content accessed by means other than via the Twitter API or other Twitter tools....Use a user’s Twitter Content to promote a commercial product or service, either on a commercial durable good or as part of an advertisement.”
- “If Twitter Content is deleted, gains protected status, or is otherwise suspended, withheld, modified, or removed from the Twitter Service (including removal of location information), you will make all reasonable efforts to delete or modify such Twitter Content (as applicable) as soon as reasonably possible...”

GDPR

- User-Side: Your rights
 - information about the processing of your personal data;
 - obtain access to the personal data held about you;
 - ask for incorrect, inaccurate or incomplete personal data to be corrected;
 - request that personal data be erased when it's no longer needed or if processing it is unlawful;
 - object to the processing of your personal data for marketing purposes or on grounds relating to your particular situation;
 - request the restriction of the processing of your personal data in specific cases;
 - receive your personal data in a machine-readable format and send it to another controller ('data portability');
 - request that decisions based on automated processing concerning you or significantly affecting you and based on your personal data are made by natural persons, not only by computers. You also have the right in this case to express your point of view and to contest the decision.

GDPR

- Business Side: The type and amount of personal data you may process depends on the reason you're processing it
 - personal data must be processed in a lawful and transparent manner, ensuring fairness towards the individuals whose personal data you're processing ('lawfulness, fairness and transparency').
 - you must have specific purposes for processing the data and you must indicate those purposes to individuals when collecting their personal data. You can't simply collect personal data for undefined purposes ('purpose limitation').
 - you must collect and process only the personal data that is necessary to fulfil that purpose ('data minimisation').
 - you must ensure the personal data is accurate and up-to-date, having regard to the purposes for which it's processed, and correct it if not ('accuracy').
 - you can't further use the personal data for other purposes that aren't compatible with the original purpose of collection.
 - you must ensure that personal data is stored for no longer than necessary for the purposes for which it was collected ('storage limitation').

Research and IRBs



BROWN UNIVERSITY

Research at Brown



Human subject: means a living individual about whom an investigator (whether professional or student) conducting research:

- (i) Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or
 - (ii) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.
- (2) Intervention includes both physical procedures by which information or biospecimens are gathered (e.g., venipuncture) and manipulations of the subject or the subject's environment that are performed for research purposes.
 - (3) Interaction includes communication or interpersonal contact between investigator and subject.
 - (4) Private information includes information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information that has been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public (e.g., a medical record).
 - (5) Identifiable private information is private information for which the identity of the subject is or may readily be ascertained by the investigator or associated with the information.
 - (6) An identifiable biospecimen is a biospecimen for which the identity of the subject is or may readily be ascertained by the investigator or associated with the biospecimen.

Ethical Dilemmas

- Twitter for public health: All tweets from a single user over an extended period of time. Reasonable expectation of privacy?
- Netflix challenge: Released was “anonymized” but could be cross-referenced with de-anonymized data online.

Ethical Dilemmas

- You are building an app that uses computer vision to do cool filters (make you look older/younger/thinner/fuller/etc). Scraping google images for faces to train your CV algorithm?
- You are building an app to help people manage their overall health. As an easy initial “ingest” they can upload pictures of health records and you’ll populate your database. Storing these pics/the database on the CIT server?

Clicker Question!

Did you sign the collaboration policy?

(a) Yes, of course.

(b) No, because I don't understand how to use the internet. What does this phrase "course web page" mean?

Okay, leave now.