



**BC COMS 2710:
Computational Text Analysis**

**Lecture 12 – Data Collection:
Webscrapping**



- Readings 04:
 - link posted to course site later today
 - due Sunday


- HW 02:
 - Due Wednesday night

- Tutorial 4.1:
 - Releasing tonight or tomorrow
 - Topic Modeling for today's data

- Office hours
 - Today 1:30-3:00 pm

Solutions to assignments



 Pinned by you



Adam Poliak 1:22 PM

Each post in this channel will correspond to a specific assignment. Please add any reaction to the post if you want me to grant you access to the solutions for that assignment.

Tutorial 1.1

Tutorial 1.2

Tutorial 1.3

Tutorial 2.1

Homework 01





- Mid-semester anonymous brief survey
- What have you learned so far and how comfortable do you feel with the material?
- What has been going well in the course so far? What are things you are enjoying about the course?
- What has not been going well in the course so far? What are things you are not enjoying about the course?
- What can we (the course staff) be doing better?



- Python Overview **Week 1**

- Lexical based analysis methods **Week 2 - 3**
 - Text Processing
 - Document Representation
 - Topic Modeling

- **Data Collection** **Week 4**
 - **Web Scraping**
 - **APIs**

- Machine Learning **Week 5**
 - Regression & Classification
 - Clustering

- Advanced Topics & Final Projects **Week 6**
 - Dimensionality Reduction
 - Word Representations



- Project ideation – Friday May 28st
- Project proposal – Friday June 4th
- Project presentations – Monday June 14th
- Project submissions – Friday June 18th
- http://coms2710.barnard.edu/final_project



- **Project ideation – Friday May 28st**
 - <https://www.overleaf.com/read/yzpgxcgsqdvq>
- roughly 250 word overview of what you are interested in



Data Collection



- New York Times Corpus (1987 – 2007)
 - <https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/GZC6PL>
- LexisNexis: need an account
 - You can experiment with:
<https://github.com/ahalterman/cloacina>



- EuroParl:
 - <http://www.talkofeurope.eu/data/>
- UK
 - <https://www.hansard-corpus.org/>
- US Congress
 - <https://www.congress.gov/>



- Manifesto project:
<https://manifestoproject.wzb.eu/>
- US Presidency
<http://www.presidency.ucsb.edu/>
- Web Archives
<https://www.loc.gov/collections/united-states-elections-web-archive/>



- Project Gutenberg
 - <https://www.gutenberg.org/>

- HathiTrust Digital Library -
<https://www.hathitrust.org/>
 - Python tool
 - <https://github.com/htrc/htrc-feature-reader>
 - Tutorial
 - <https://programminghistorian.org/en/lessons/text-mining-with-extracted-features>
 - Examples
 - <https://github.com/htrc/htrc-feature-reader/tree/master/examples>

Common Crawl



The web:

- largest & most diverse collection of information in history
- Rich corpus for scientific research, technological advancement, and innovative new businesses
- A digital copy of our world

Using the web to:

- insight into politics, art, economics, health, culture and almost every other aspects of life.

Purpose of Common Crawl:

- Make the web be openly accessible to anyone who desires to utilize it

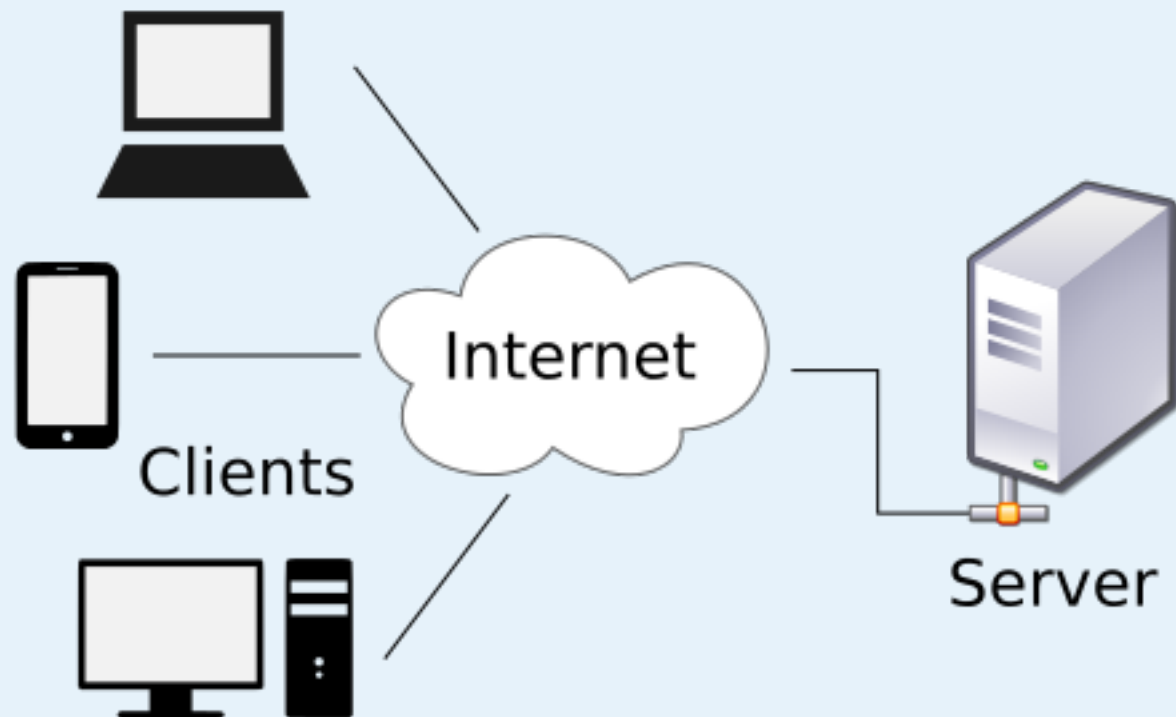


—

Interacting with the Internet

—

- HTTP
 - *HyperText Transfer Protocol*





Interact with websites via Request & Responses:

- Request:
 - *an operation to be performed on a URL*
- Response:
 - Message from server based on a request



— Web Scraping & Crawling —

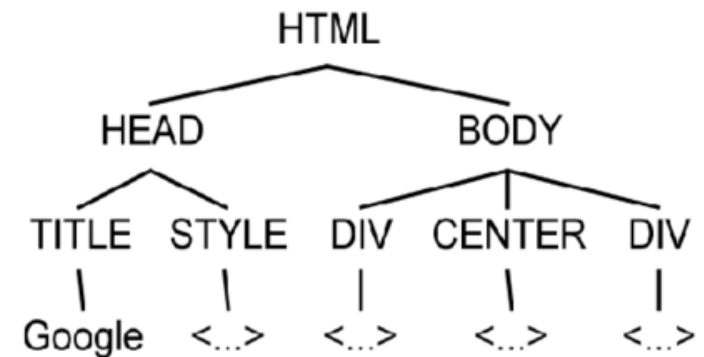


- Scraping:
 - *Using tools to gather data you can see on a webpage*
- Crawling:
 - *Moving across or through a website in an attempt to gather data from more than one URL or page*
- *HTML:*
 - *HyperText Markup Language*
 - The standard markup language on the Web



- HTML **tags** to represent different elements on a web page
- Structured as a tree

```
<html>
  <head>
    <title>Google</title>
    <style>...</style>
  </head>
  <body>
    <div>...</div>
    <center>...</center>
    <div>...</div>
  </body>
</html>
```





HTML Tag	Explanation
<!DOCTYPE>	Defines document type
<html>	Defines HTML document
<head>	Main information about document
<title>	Title for document
<body>	Document body
<h1> to <h6>	Headings
<p>	Paragraph
 	Line break
<!--comment here-->	Comment
	Image
<a>	Hyperlink
	Unordered list
	Ordered list
	List item
<style>	Style information for a document
<div>	Section in a document
	Section in a document



- Selectors:
 - Class – can apply to multiple elements
 - Id – unique to an element

- Attributes
 - url: `<href>`
 - Image
 - Style



Beautiful Soup



- Python library for parsing HTML (and XML)
- Documentation:
 - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>