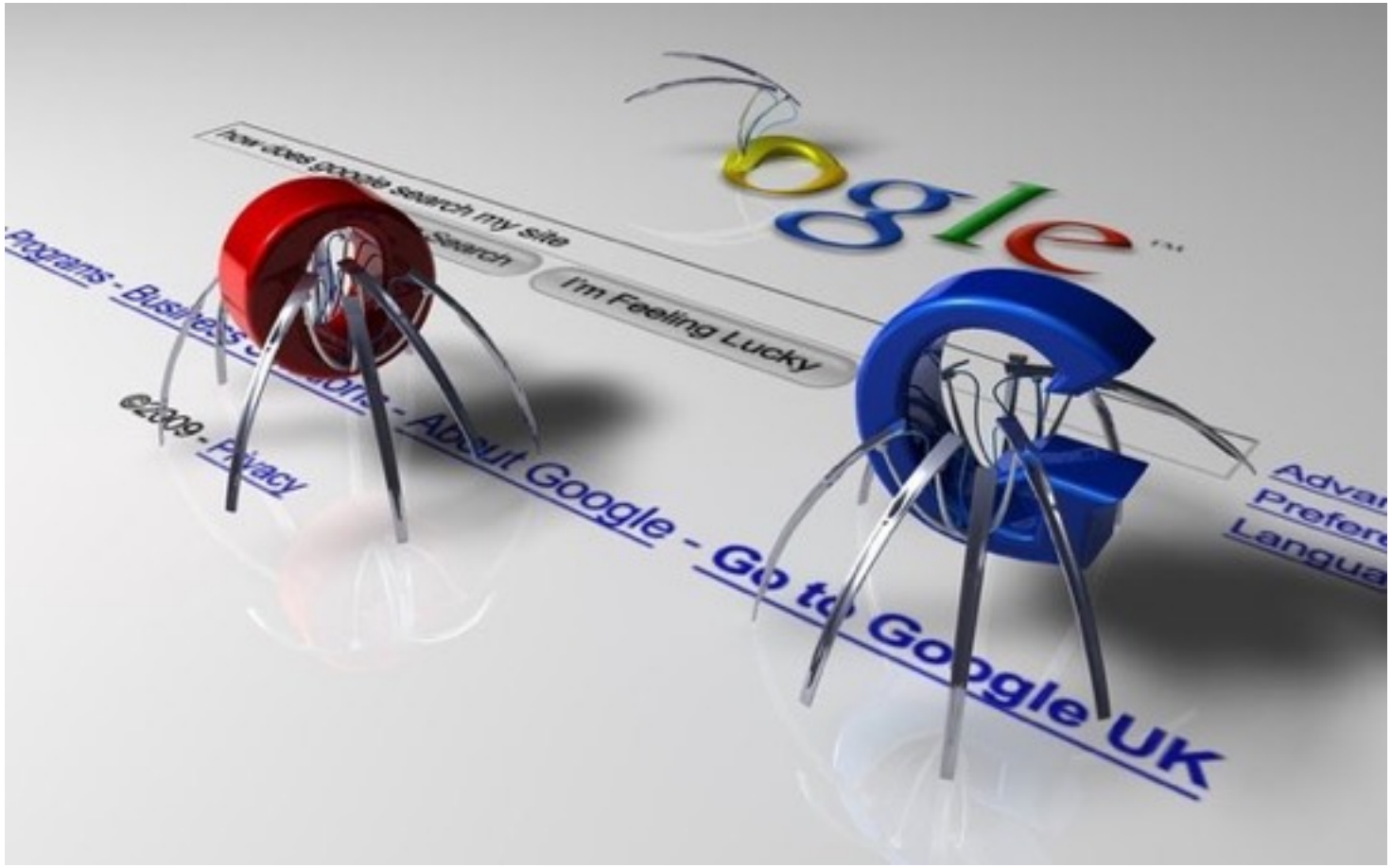
EECS 510: Social Media Mining
Spring 2016

Web Crawling ++

Arindam Paul

arindam.paul@eecs.northwestern.edu

What is crawling ?



Web Crawling Intro

Web Scraping

Web Crawling vs. Web Scraping

- Web Crawling is the process of getting the links in a webpage, and then recursively visiting those links and again, extracting links in those webpages
  - Level 1- > Level 2 -> Level 3 ….

- Web Scraping is the process of getting the data from a webpage

Libraries needed

- Beautiful Soup
- Selenium
- UrlLib

Beautiful Soup

It converts the html into a dictionary like structure

Selenium

We are able to automate web browsing

Some html tags

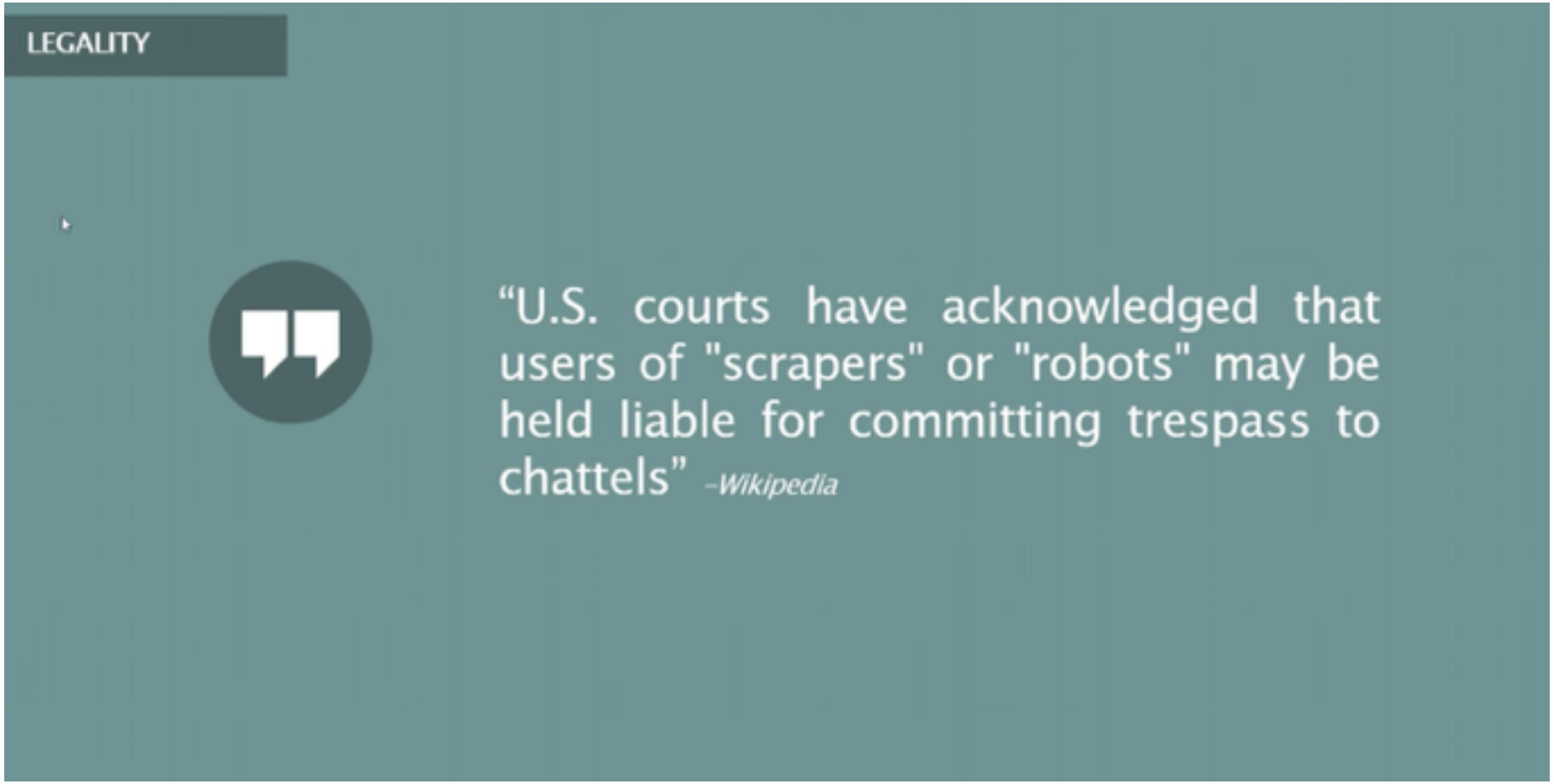
- a = anchor with attribute href for url
- img = for images with attribute src for image url
- p = for paragraphs

Two ways of Scraping

- Selenium
  - from bs4 import webdriver
  - Browser = webdriver.Chrome()
  - Browser.get(website)
  - Source = Browser.page_source

- URLLIB
  - From urllib improt urlopen
  - url = urlopen(website)
  - Source = url.read()

Legality


LEGALITY
“U.S. courts have acknowledged that users of "scrapers" or "robots" may be held liable for committing trespass to chattels” -Wikipedia

Issues

- Real life is difficult
  - You have to see which all domains called
    - Crawling “Youtube” ?
  - Bad HTML Structure
  - Websites changing pages

References

- 1. Web Scraping with Python
- 2. Beautiful Soup Documentation
- 3. Other softwares:
  - Scrapy
  - RapidMiner