



Advanced crawling techniques

Patrik Hudák, 2017

PV211 – Introduction to Information Retrieval



Dealing with forms

- Logins – set correct session cookie
- Text boxes
 - Build dictionary of terms from previous pages
 - Submit combination of terms (ngrams)
 - CAPTCHA – tesseract + ~10 lines of Python
- GitHub projects
 - Formasaurus
 - Autologin



Dealing with client-side rendering

- MVC/MVVM JavaScript frameworks are popular nowadays
 - Vue.js, Angular, Ember, ...
 - ReactJS as a View in MVC
- **Solution:** Headless browsers (not only Javascript engine!)
 - PhantomJS (not maintained since April 2017)
 - Google Chrome will include headless mode from version 59



Adaptive rate limiting

- ▶ Check delay of HTTP responses
- ▶ Check ratio of HTTP status codes (200 vs. 403, 500, ...)
- ▶ Calibrate after – slow down or pause and wait
- ▶ Rate limiting is not that prevalent in scraping, REST API calls often have these limitations
- ▶ DNS caching – some crawlers resolve each domain every time



Attacking crawlers

- Serving different content based on identification
 - "User-Agent" header
- Banning IP after exceeding number of requests
- Check behavior
 - Does agent handle cookies?
 - Does agent download JS / CSS files?
- Create canaries in robots.txt, ban IP when accessed
- Bots accept gzip compression – give them ZIP bombs
- Something similar possible with XML Bomb (*Billion laughs*)



Scraping tools / frameworks

- **Scrapy (Python)** – sufficient for most usecases
 - AutoThrottle extension for rate limiting
 - Can be extended with headless browser
- Nokogiri (Ruby)
- Build your own stack
 - Requests + BeautifulSoup (Python)
 - Selenium + Driver for specific browser
- Searx – self-hosted (meta)search engine
- scrapinghub.com – Scraping as a Service




Keyword monitoring



- **Pastebin.com** and other “paste” sites often contain interesting data
 - Password dumps
 - Links to hidden content (GDrive documents, Dropbox files, ...)
 - CC data / Bitcoin wallets
 - Private keys (SSL, PGP)
- Scrape; Apply pattern matching; Index
- **Commercial:** Recorded Future, Digital Shadows



Common Crawl

- <https://commoncrawl.org/>
 - WARC file format
 - Many example projects and papers
 - Latest release (March 2017) has ~60TB compressed
 - Hosted on AWS
- 



MEMEX

- Advanced (dark) web crawler by DARPA
- *“Memex seeks to develop software that advances online search capabilities far beyond the current state of the art.”*
- Whole project is collection of open-source tools from various authors
 - <https://opencatalog.darpa.mil/MEMEX.html>
- Deployment itself is a secret sauce, along with configs



Contact

hudak@mail.muni.cz

