Separating Bots from Humans

Ryan Mitchell
@kludgist

DEF CON 23 August 8th, 2015
Who am I?

- **Software Engineer**
- **Author of two books:**
  - Web Scraping with Python (O’Reilly, 2015)
  - Instant Web Scraping with Java (Packt, 2013)
- **Engineering grad from Olin College**
- **Masters student at Harvard University School of Extension Studies, 2016**
A history of this talk
The O’Reilly Hacking Book:
Separating Bots from Humans
Pro-tips to get what you want:

- Include some market research
- Write it in Python, because it’s really popular
What are Web Scrapers, Bots, etc?

- They can use browsers
- They can take their sweet time
- They can be surprisingly smart
- They can be stunningly idiotic
Why They’re Important

In 2014, bots account for 56% of all website traffic. Good bots account for 27% of all website visits, while bad bots account for 29%.

- **Bad Bots**
  - 22% Impersonators
  - 3.5% Hacking Tools
  - 3% Scrapers
  - 0.5% Spammers

- **Good Bots**
  - 44% Human Traffic

Source: https://www.incapsula.com/blog/bot-traffic-report-2014.html
On the Defense Side of Things

(For better or worse)
robots.txt?

- “No Trespassing, please?”
Terms of Service

● “Hey! You said you wouldn’t trespass!”
Headers

- “I’m totally not a bot. Promise”
JavaScript

- Make your site un-indexable for anyone but the bad guys
Embedding Text in Images

● Oh come on.
● You’re the type of person who writes email addresses like “m e (at sign) domain . com”
  ○ And you have duct tape on your laptop’s web cam, mostly because you never use it.
CAPTCHAs

Annoying
Breakable
Honepots

- Can be effective, if implemented correctly
- Please don’t block the Google bots
Example time!

http://ryanemitchell.com/honeypots.html
Behavioral Patterns

- Now we’re getting somewhere!
- Again, please don’t block the Google bots
IP Address Blocking

- It’s sort of effective… If they didn’t really care in the first place
- Lists are a pain to maintain
- You can easily block the good guys
On the Attack Side of Things...
Targeted vs. Non-Targeted Attacks

- Non-targeted: Also known as, “look for /phpMyAdmin”
- Targeted, usually to get proprietary data
OCR

- Works best on relatively normal text
- Can be used to solve CAPTCHAs
  - Time consuming to create training data. Have a series or two of a TV show ready
OCR Training Tool

- Everything you need to solve a CAPTCHA!
  https://github.com/REMitche01/tesseract-trainer
JavaScript Execution

- Selenium
- PhantomJS
Honeypot Avoidance

- Better than you might expect -- it’s biggest weakness is color

https://github.com/REMitchell/python-scraping/blob/master/chapter12/3-honeypotDetection.py
Stop Caring!

- Bot-proofing sites is way too much work, and often impedes accessibility
- Is your data really that valuable?
  - Consider API costs, ease of use -- make it more attractive to pay for data
- If your application is vulnerable to automated attacks, it’s vulnerable, period.
Question time!