Downloading the Internet with a Snake

a.k.a Web Scraping in Python

Mihail Gavrilita

FAF Depfafrtment
FAF Fafculty
Fafical Fafersity of FAF

Chisifaf, 20faf19

The Internet

A near-infinite source of data

- Your data is there but how to collect it?
- Use it in a Hackathon presentation, ML project or a data-processing startup idea.

Web Scraping

Mechanical copy-pasting monkey

- Automatic data collection (scripts);
- For search engines or business market research.

Web Scraping

Traditional approach

- Crawling
  - Visit a link;
  - Find all links on that page;
  - Save the links;
  - Repeat.
- Parsing
  - Visit a link;
  - Detect you are on a page you wanna scrape;
  - Collect data you are interested in;
  - Save the link.

Web Scraping

Scrapy – a Python framework

Scrapy allows it's users to crawl websites using spiders and parse pages containing data, saving it into items. On the fly processing of the data we collect happens in itemloaders. The processing of the items themselves is handled by pipelines.

Web Scraping

Topics not covered here

- JavaScript rendering (headless browsers);
- Techniques used by websites to slow down / stop scraping;
- Legality of web scraping.

Web Scraping

Links 2 3 4

For the live part:

- https://help.ubuntu.com/community/VirtualBox
- http://releases.ubuntu.com/
- https://askubuntu.com/questions/298290/smbus-bios-error-while-booting-ubuntu-in-virtualbox
- https://www.psychocats.net/ubuntu/virtualbox
- https://docs.scrapy.org/en/latest/intro/install.html
- https://docs.conda.io/projects/conda/en/latest/user-guide/install/
- https://www.sublimetext.com/docs/3/linux_repositories.html
- https://alternativeto.net/software/scrapy/