WEB SCRAPING

BY EMIL KJER
EMIL@KJER.INFO
RMIT UNIVERSITY
MARCH 2012

March 28, 2012

WHAT IS SCRAPING?

March 28, 2012

From http://youtu.be/--THSIi85I8

WHAT IS SCRAPING?

March 28, 2012

From http://youtu.be/--THSIi85I8

WHAT IS SCRAPING?

March 28, 2012

From http://youtu.be/--THSIi85I8

WHAT SCRAPING IS

Web scraping / web harvesting / web data extraction

1. **Collect content from a source**
   - Firefox, Chrome, Safari
   - wget, curl, urllib2[python], URL[java], Net::HTTP[ruby]
2. **Do something to the content**
   - Parser with regular expression
   - BeautifulSoup [python]
   - Jsoup [java]

WHAT TO USE IT FOR?

- Research
- News mashup
- Weather app
- Movie comparison
- Webcrawler

REMARKS

Politeness

- Heavy load
- Amount of data
- Private sources
- Bad code

Robustness

- Data changes form
- Broken links

TIMETABLES

In a browser go to:

http://www.metlinkmelbourne.com.au/timetables

Select transport mode

http://www.metlinkmelbourne.com.au/timetables/metropolitan-trains

Select a route e.g. Alamain Line

http://www.metlinkmelbourne.com.au/route/view/1

Select a direction e.g. to city

http://tt.metlinkmelbourne.com.au/tt/XSLT_TTB_REQUEST?command=direct&language=en&outputFormat=0&net=vic&line=02ALM&project=ttb&itdLPxx_selLineDir=H&sup=B

Download the page via. the browser e.g. in Safari File > Save As… > Format "Page Source".

GET AND EXTRACT 1

Simple and “solution” using Python and BeautifulSoup

- From file
- From url
- HTML parsed to objects

BeautifulSoup

http://www.crummy.com/software/BeautifulSoup/

GET AND EXTRACT 2

Using Java and Jsoup

- From file
- From url
- HTML parsed to objects

Jsoup

http://jsoup.org

Alternative from javax for parsing html as documents
javax.swing.text.html.HTMLEditorKit

SUMMARY

Many applications

1. **Collect data**
2. **Do something with the data**

Politeness and robustness

Source code