Web scraping and social media scraping – assessment criteria

Jacek Lewkowicz, Dorota Celińska-Kopczyńska

University of Warsaw

March 6, 2019

Tutors

- PhD Jacek Lewkowicz
  - email: jlewkowicz@wne.uw.edu.pl
  - office hours: please contact via email
- mgr Dorota Celińska-Kopczyńska
  - email: dcelinska@wne.uw.edu.pl
  - site: coin.wne.uw.edu.pl/dcelinska
  - office hours: Wednesday 14:00-15:00 – please contact via email

Mails

- It is easier to look for your emails if the topic of the mail contains some information related to the course

  WS_Surname_Summary

- Please, do not start a new thread in a thread related to a completely different issue (e.g., do not send us topic proposals in the thread related to exercises from the classes)

Type of course

- Computer laboratory
- 7 classes in the first half of the spring semester
- Prior registration in USOS is mandatory
- You may come to any of the groups (the limitation is the number of seats)
- ... however try not to change the group on the test days

Elements of assessment

- Short tests during the classes (20%)
- Activity during the classes (20%)
- Group projects (60%)
- **One has to score at least 50% of points for a positive final grade!**

Short tests

- There will be **two** short tests
- Pen and paper, without using the computers, no open book.
- We will inform you about the dates and which part of the material will be covered
- You cannot improve the grade from the tests
- **Absence during the test does not allow for writing the test later**

Activity during the classes

- Starting from the next week (nearly) each class will have similar form
- We will start with a "lecture" part given by a tutor
- Then you will be given the exercises which you **have to solve by yourselves** (you may work in pairs)
- Probably, those exercises will be of various level of difficulty
- In the end of the classes, you will be asked to send the results **via email** to the tutor

Activity during the classes

- During the grading session, the points from the activity part will be assigned upon the completeness and difficulty of the tasks you solved in the classes
- **Note: this is your "classroom activity" so you cannot send the results later, e.g., after a few days or so**. This also means that its score should not be improved
- ... however, you can discuss the tasks with us if you failed to solve them (or had any other concerns)
- We will try to send you the feedback as soon as possible

Group projects

- We would like you to form a few groups and collectively solve the problem with scraping a website and preparing a basic analysis of the gathered output
- **One group consists of 2-3 participants**
- In a few weeks we will provide the list of websites and topics (with perceived difficulty of the tasks)
- You are welcome to propose your own topics
- The deadline for sending solutions to the project: **May 19, 2019**

What will we be working on?

- **Introduction**: what is scraping, the controversies and the good practices
- **Basic scraping**: selectors, scraping links, lists, tables, downloading ready-made files
- **Crawling the web**: scraping multiple files, multiple pages, finding rules for crawling
- **Simulating the user**: HTML forms, authentication, cookies
- **Alternative scraping**: APIs, scraping twitter and/or reddit, web data

Why scraping the web?

- At the end of the course, you should be ready to acquire the tools for extracting the data from the web. This workshop will introduce techniques for automated extraction of content. In particular, you will be able to:
  - gather and mine the information from the web/social media using the most suitable software,
  - create custom web/social media scrapers,
  - use Python/R packages for the data analysis.

Software

- R (especially twitteR, self-explanatory)
- Beautiful Soup (Python)
- Scrapy (Python)
- Additional libraries for special tasks, e.g., Selenium, Splash...
- *Anti-scraping*: wget and core utils (cat, grep, sed, awk...)

Suggested reading

- Papers, blog articles, manuals...
- R. Mitchell (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media.
- R. Lawson (2015). Web Scraping with Python. Packt Publishing.
- S. Munzert, Ch. Rubba, P. Meissner, D. Nyhuis (2015). Web Scraping with Python: Collecting Data from the Modern Web. Wiley
- S. Munzert et al. (2015). Automated Data Collection with R. Wiley.