Würgeschlange 3 - Lesson 13

Tobias Maschek, Viktor Reusch
https://github.com/jemx/wise1920-python

mit Materialien von Felix Döring, Felix Wittwer https://github.com/fsr/python-lessons

4. Februar 2020

Python-Kurs

Gliederung

1. Vorbereitung

2. Umsetzung

Vorbereitung

Ziel - Was, wie und warum überhaupt?

- Was? - Webcrawler *(just copy the entire internet)*
- Wie? - Natürlich mit Würgeschlange 3!
- Warum? - Weil 's geht und ohne Stützräder

Fragen über Fragen...

— Hier der rechtliche Disclaimer —

Informiert Euch immer zu den Nutzungsbedingungen der abzugrasenden Seite.

Hinweis: Einen *Denieal-Of-Service* auszulösen ist keine gute Idee...

Umsetzung

Grober Fahrplan

- Webserver nach RSS-Feed fragen (HTTP Request)
- XML parsen, filtern
- Gesammelte Daten auswerten

Aufgabe 13-1

Ziel: Ein RSS-/Atom-Feed der Fakultät abgreifen

- `https://tu-dresden.de/ing/informatik/die-fakultaet/news/atom.xml`
- Erhaltenes XML-Dokument in eine Datei schreiben für nächste Aufgabe

requests

Das *pip*-Packet requests ist empfohlen.

http

Ohne extra Packete ist die Standard-Bibliothek zu verwenden.

Aufgabe 13-2

- XML aus der Datei parsen und filtern nach:
  - Titel
  - Zusammenfassung
  - Veröffentlichungsdatum jedes Eintrags
- Überlegt euch, wie ihr die Eigenschaften für alle Einträge in eine Textdatei schreibt

XML

Ihr könnt die Standard-Bibliothek verwenden. Z. B.: `iterfind("{*}entry")` und `find("{*}summary")` sind hilfreich.

Sicherheit geht vor!

XML-Parsen ist gefährlich! defusedxml schützt!

Aufgabe 13-3

Auswertung

- Zählt wie oft welcher Buchstabe vorkommt
- Um welche Zeit (Stunde) wurden die Artikel veröffentlicht? Was fällt auf?