Webscraping mit Scrapy

Florian Preinstorfer

https://nblock.org

VALUG

13.05.2016

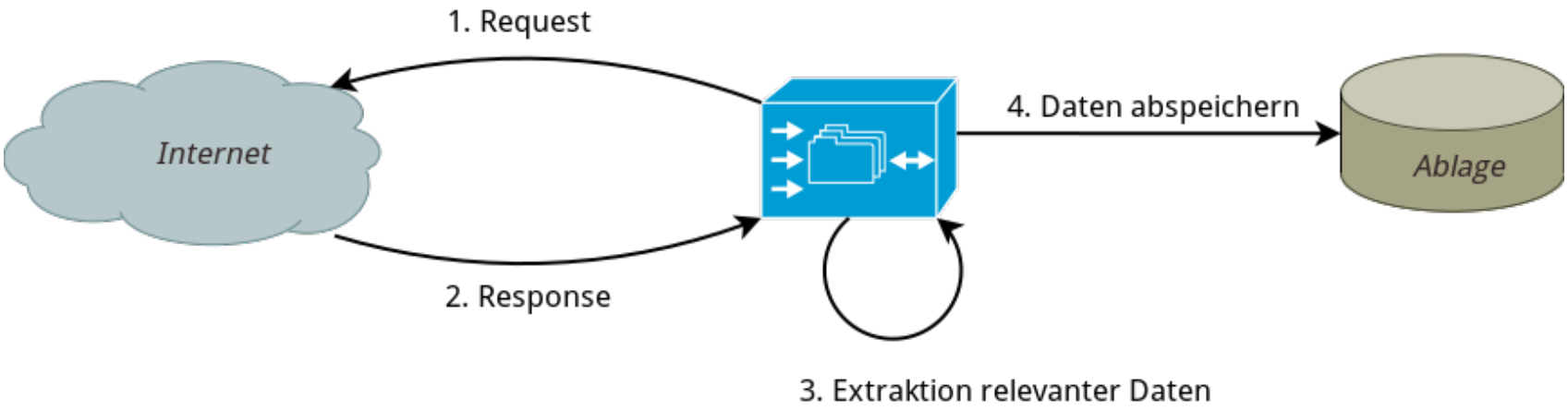
Was ist Webscraping?

- Daten von Webseiten extrahieren
- Gewinnung von relevanten Informationen
- Umwandlung in strukturierte Daten

Einsatzgebiete

- Suchmaschinen
- Web Services (z.B. als API-Ersatz)
- Monitoring von Änderungen (Webseiten, Preise, . . .)
- Automatisierung

Funktionsweise



Scrapy

- Ein Webscraping Framework
- Python 2 & Python 3
- Basierend auf Twisted
- Lizenz: BSD

Historie

2008 Initiales Release

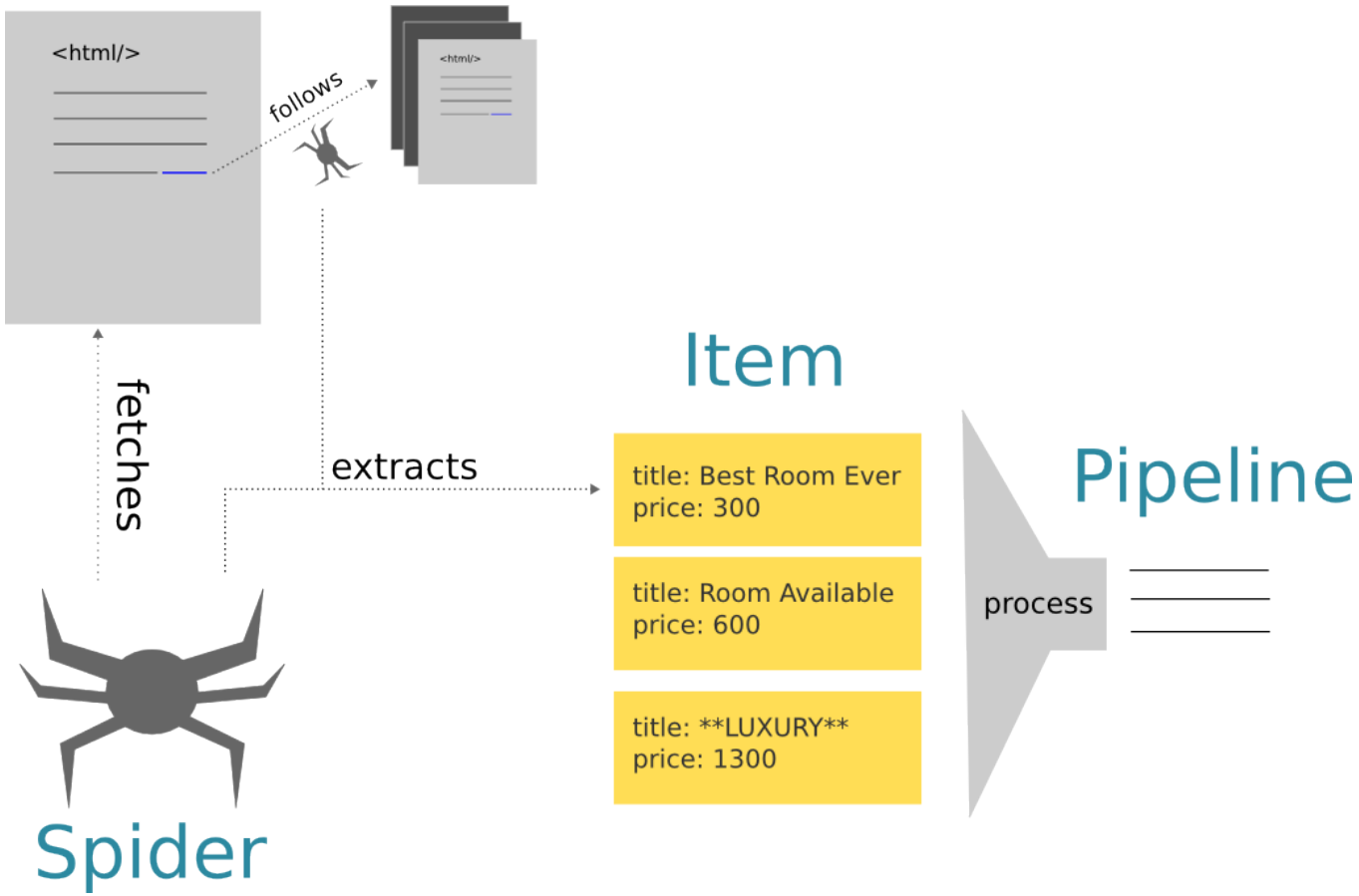
2015 Scrapy 1.0

2016 Scrapy 1.1 (Python 3 Unterstützung)

Projektaktivität

- Scrapy (Version: 1.1)
  - LOC: ∼24k
  - 329 Commits von 35 Entwicklern im letzten halben Jahr.
  - 4 Entwickler mit mehr als 10 Commits im letzten halben Jahr.

Funktionsweise



Quelle: `https://gabrielelanaro.github.io/blog/2015/04/24/scraping-data.html`

Vergangene Veranstaltungen

- Ziel: Eine Liste vergangener Veranstaltungen erstellen
- Extraktion von: Titel, Datum, Anzahl der Teilnehmer
- Folgen von Links
- Speichern für die weitere Verarbeitung

LUGs

- Ziel: Eine Liste von LUGs erstellen
- Extraktion von Name, Region, Land, Gründungsdatum, Letztes Update, Webseite, Kontakt
- Folgen von Links
- CrawlSpider
- Einfachere und robusterer Spider mittels ItemLoader
- Speichern für die weitere Verarbeitung

Projekte

- Demos sind schön, aber lässt sich Scrapy produktiv nutzen?
- Ja: https://github.com/nblock/feeds
- Atom Feeds für Webseiten erstellen, die selbst keine anbieten

Diskussion

- Rechtliche Bewertung von Webscraping?
- Weiterverarbeitung von gescrapten Daten?
- Veröffentlichung von gescrapten Daten?

Fazit

- Stabil und ausgereift
- In allen gängigen Distributionen verfügbar
- Steile Lernkurve und viel Vorwissen nötig
- Robuste Spider möglich

Code

Online verfügbar:
`https://gitlab.com/valug/notizblock-scrapy-demos`

- Getestet mit Scrapy 1.0.5 unter Python 2
- Siehe notizen.txt für Anleitung und Ablaufbeschreibung