

Web scrapping

Presented By:

www.seminarlab.com

contents

- History.
- What is web scraping.
- Methodology.
- Evolution.
- Future.
- Application.

What is web scrapping

- Web scrapers are conventionally used to extract data from web documents.
- Bulk of web data are in html by a layer presentation.
- These data updated with additional content. So working with them are difficult.
- Site scraper tool solve these problem.

Site scraper

- It can deal changing web content / structure.
- It is written in python. So it is platform independent.

Previous work

- Chicken foot
- Piggy Bank
- Sifter
- Scrubyt
- WWW Mechanize
- Template Maker

METHODOLOGY

- Parse
- Search
- Generalise
- Attributes

EVALUATION

To estimate the performance of Site Scraper over different scraper needs, we identified three defining parameters for the text chunk types:

- Whether there was a single or multiple text chunks to extract;
- Whether the number of text chunks was static or would vary and require abstraction;
- Whether the text chunk was simple, such as a single number, or complex, such as a news article.

FUTURE WORK

- SiteScraper's model needs to be less brittle and able to handle greater variation
- SiteScraper does not interpret the effect of JavaScript events. This could be addressed by embedding SiteScraper in the browser as a plug-in, or using a library such as Watir16 to interface with the browser
- A final feature that needs implementing is determining whether the structures of two WebPages match.

Thank You

By

www.seminarlab.com

www.seminarlab.com