

XML Technologies

Hannu Sirén G2458

Jarmo Puttonen F0899

Summary

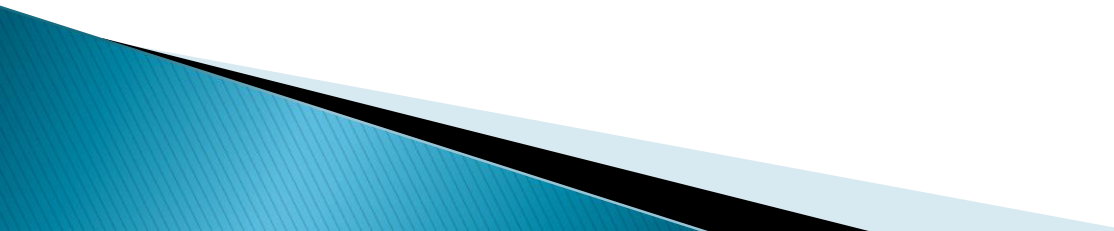
- ▶ Software that learns to search specific news from reddit
- ▶ User scores pages either useful or useless
 - Each vote calibrates keywords
- ▶ Software calculates scores for pages
 - Score is calculated from keywords

Important Libraries

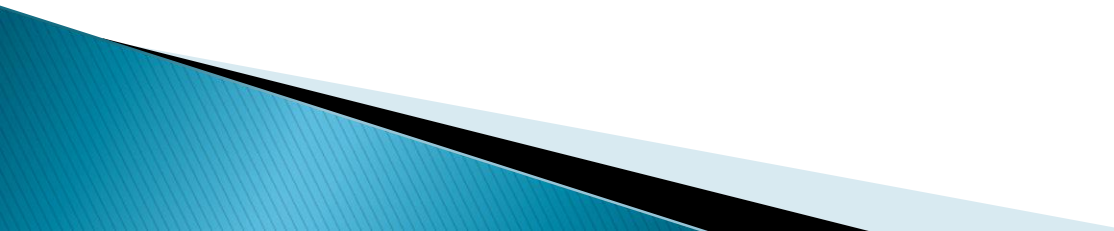
- ▶ Multiprocessing
- ▶ Threading
- ▶ Httpplib2
- ▶ BeautifulSoup
- ▶ `xml.etree.ElementTree`

Conf.xml

```
<data>
  <sites>
    <site>http://www.reddit.com/r/programming/
  </site>
  </sites>
  <params>
    <search_param>c++</search_param>
  </params>
  <pages>1</pages>
  <mixml>results/data.xml</mixml>
</data>
```



Elements

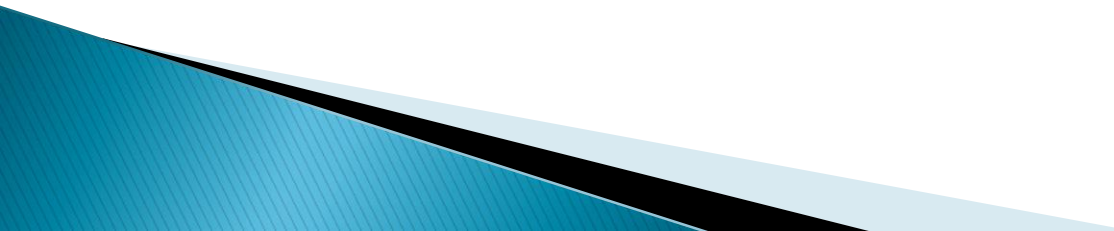
- ▶ Parent
 - ▶ Child
 - ▶ UI
 - ▶ Updater
- 
- A decorative graphic element in the bottom-left corner of the slide, consisting of overlapping blue and black geometric shapes.

Parent

- ▶ Loads xml confs
- ▶ Starts new thread
 - Thread starts new child processes
 - Thread communicates with child processes
- ▶ Combines child processes results as one xml
- ▶ Starts browser

```
hannu@hannu-X501A: ~/codes/python/httpContentTracker
hannu@hannu-X501A: ~/codes/python/httpContentTracker 83x39
hannu@hannu-X501A:~/codes/python/httpContentTracker$ python3 parent.py
Searching: I finished writing my free book on game programming!
Score: -100
Searching: GCC 4.9.0 Released
Score: -5
Searching: LibreSSL: OpenBSD's fork from OpenSSL
Score: -3
Searching: Project Naptha: a browser extension that enables text selection on any
image
Score: -104
Searching: Game Mechanic Explorer
Score: -7
Searching: My Quest to Build the Ultimate Music Player
Score: -310
Searching: Link-time optimization in GCC 4.9: A brief history
Score: 0
Searching: How to be a great software developer
Score: -1
Searching: Origins of LibreSSL (the OpenBSD fork of OpenSSL)
Score: -155
0 Searching: The case for formal verification of software
Score: -100
3 Searching: Functional Programming in JavaScript == Garbage
Score: 0
4 Searching: Maze Solver II
Score: 1
5 Searching: Worst common denominator programming [where OpenSSL portability code
oes bad]
Score: 0
6 Searching: A Quick Look at Haskell
Score: 0
7 Searching: Fixing old bugs, without the source.
Score: -150
8 Searching: PMP Preparation
Score: 0
9 Searching: Embed a chess board on your site with chessboard.js
Score: 88
0 Searching: wreq: a capable new HTTP client library for Haskell
Score: 140
```

Child

- ▶ Loads reddit page
 - Get links
 - ▶ Check if link is unique
 - ▶ Follow links
 - Get page data
 - ▶ Save page data
 - ▶ Calculate score
 - ▶ Create xml file containing results
- 

UI

- ▶ HTML Page
- ▶ Javascript loads xml data
- ▶ Presents data dynamically
- ▶ Inputs are send to server
 - Server calls for Updater
- ▶ Repeat

Getting Lazy with C++ - Score: 702

[Getting Lazy with C++](#)

wreq: a capable new HTTP client library for Haskell - Score: 140

[wreq: a capable new HTTP client library for Haskell](#)

Embed a chess board on your site with chessboard.js - Score: 88

[Embed a chess board on your site with chessboard.js](#)

Maze Solver II - Score: 1

[Maze Solver II](#)

Link-time optimization in GCC 4.9: A brief history - Score: 0

[Link-time optimization in GCC 4.9: A brief history](#)

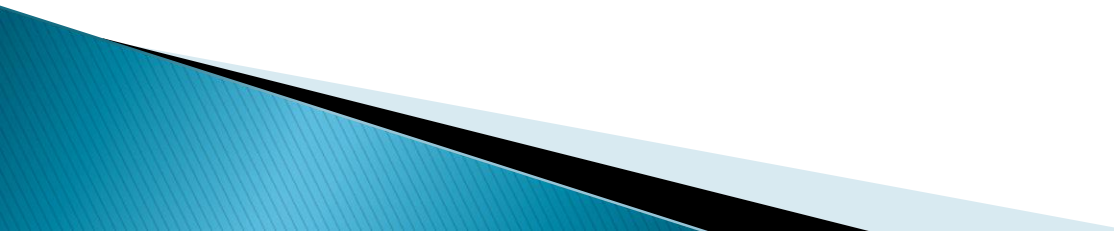
Functional Programming in JavaScript === Garbage - Score: 0

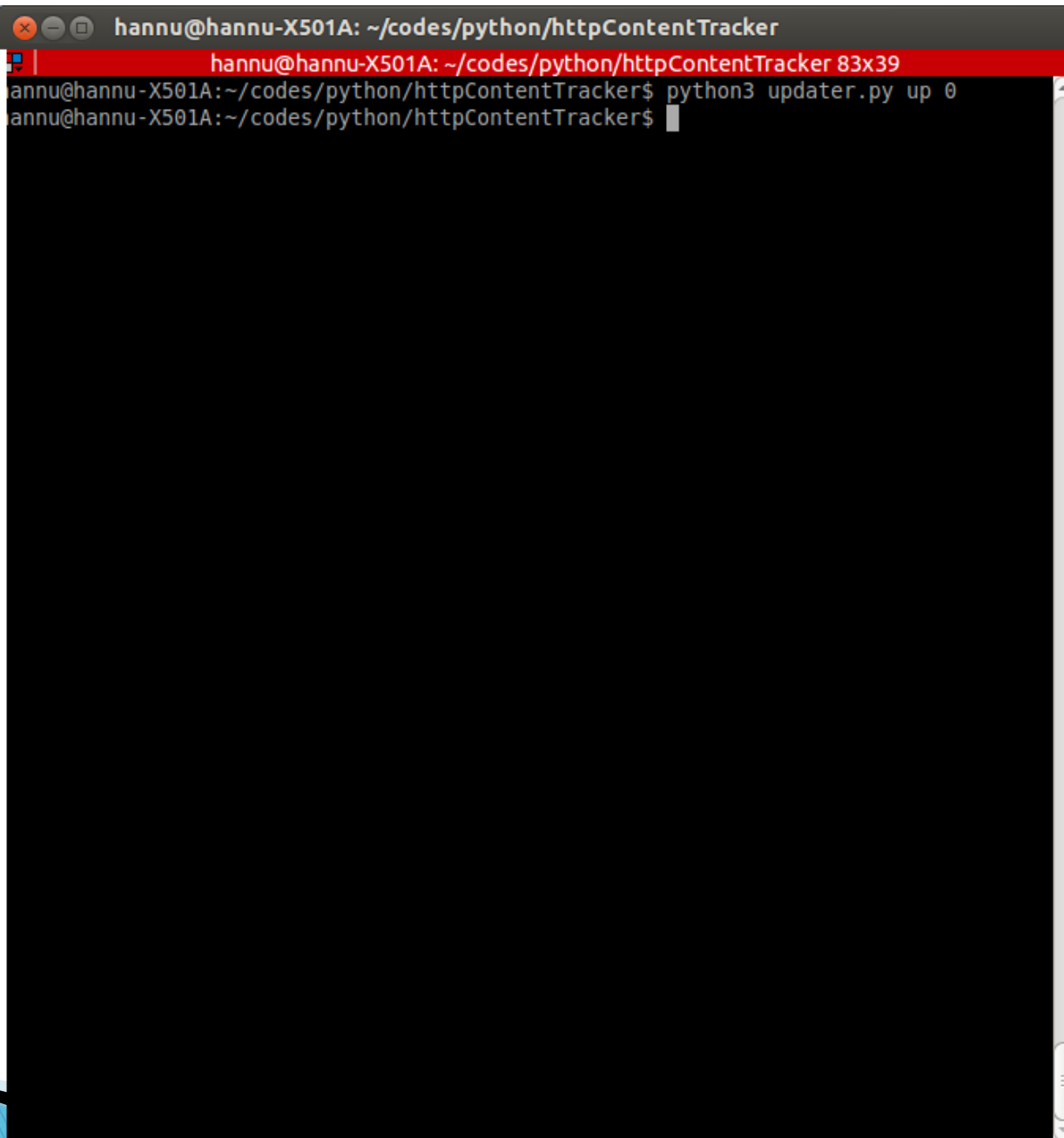
[Functional Programming in JavaScript === Garbage](#)

Worst common denominator programming [where OpenSSL portability code goes bad] - Score: 0

[Worst common denominator programming \[where OpenSSL portability code goes bad\]](#)

Updater

- ▶ Is called by server with vote and ID
 - ▶ Load page data corresponding to ID
 - ▶ Store link, title, headers and chapters to words.xml
 - ▶ Recalculate keywords
 - ▶ Remove loaded page data, and xml element from results
 - ▶ Return to UI
- 

A terminal window with a dark background. The title bar at the top shows window control icons and the text "hannu@hannu-X501A: ~/codes/python/httpContentTracker". A red status bar below the title bar displays "hannu@hannu-X501A: ~/codes/python/httpContentTracker 83x39". The main area of the terminal shows a command prompt "hannu@hannu-X501A:~/codes/python/httpContentTracker\$" followed by the command "python3 updater.py up 0". The command has been executed, and the prompt has moved to the next line, showing "hannu@hannu-X501A:~/codes/python/httpContentTracker\$".

```
hannu@hannu-X501A: ~/codes/python/httpContentTracker
hannu@hannu-X501A: ~/codes/python/httpContentTracker 83x39
hannu@hannu-X501A:~/codes/python/httpContentTracker$ python3 updater.py up 0
hannu@hannu-X501A:~/codes/python/httpContentTracker$
```