HW4: Crawler & Main Content Extraction

403410034 資工四 黃鈺程

Outline

- Environment
- Main Content Extraction
  - Data
  - Block Segmentation
  - Reconstruction Error
- Hacker News Crawler
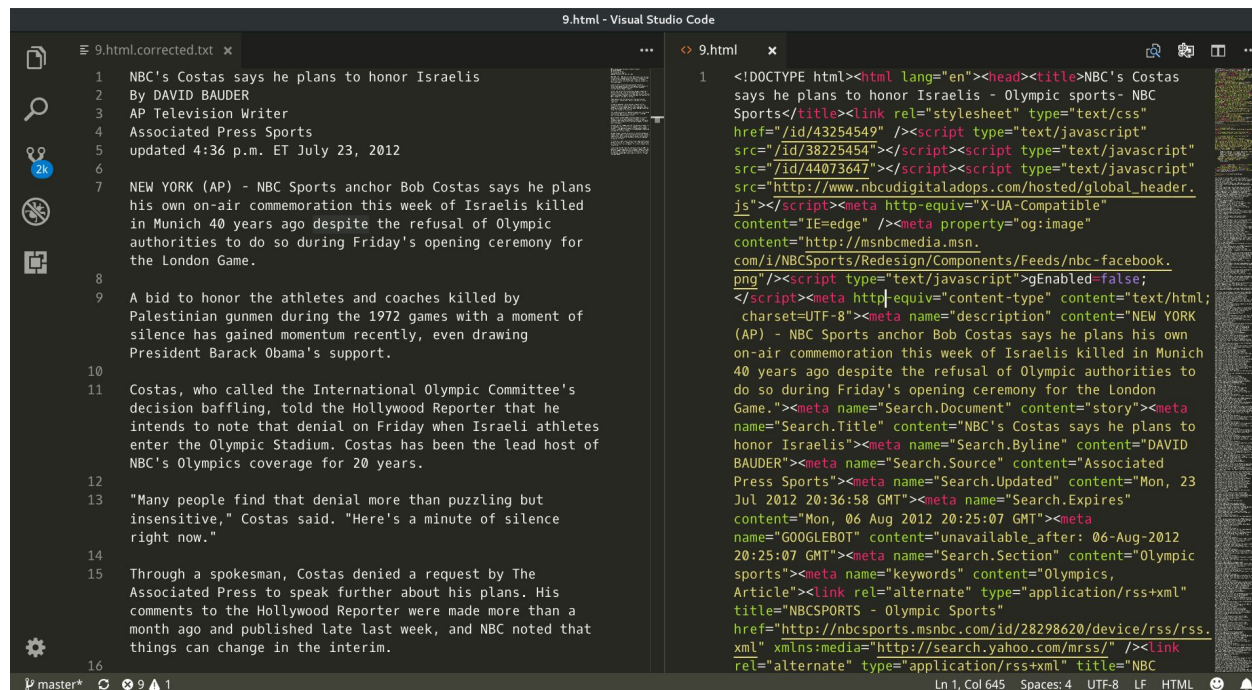 - URLs
  - Result

Environment

- Fedora 27
  - Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz
  - 4 Cores
  - 4 GB
- Python 3
  - Jupyter, Numpy, Pandas, Matplotlib,
  - Requests, Selenium, Beautifulsoup4

Main Content Extraction

- Plan
  - Perform classification using Random Forest on DOM elements.
  - Features:
    - Tag type
    - Depth of element
    - Content length
    - etc
  - We need to have a label on each elements

Data

- From Dragnet
  - https://github.com/seomoz/dragnet_data
  - HTML & Corrected Text
  - 1000 Training, 300 Testing

Block Segmentation

- Split DOM into **disjoint** blocks
- I use reversed DFS to iterate on DOM
  - Deeper elements first
  - Emit a block if content of subtree > 15
    - And extracted from DOM
    - And label is TRUE if
      - LCS(content, truth) / max(len(content), len(truth) > 0.2
  - Pass, otherwise

Reconstruction Error

- My segmentation is good in some cases, but bad in other cases

- To quantitatively measure, I set up a metric for reconstruction error.

- LCS(re, gt) / max(len(re), len(gt)) where
  - re: reconstruction text
  - gt: ground truth

- Result is bad
  - min: 0.06
  - max: 1.00
 - avg: 0.6x

Hacker News Crawler

- Found no solution so far, give up.
- Build a simiple crawler instead
- Crawl on Hacker News' Stories
  - Lots of different websites
- Top 30 stories of each day.
  - 90 days
- 2700 urls, 1415 distinct sites

Hacker News Crawler

Y | Hacker News new | comments | show | ask | jobs | submit

1. ▲ Gojay – Highly performant JSON encoder/decoder for Golang (github.com)
   98 points by francoisllm 5 hours ago | hide | 42 comments

2. ▲ Getting 1Password 7 ready for the Mac App Store (agilebits.com)
   224 points by okket 5 hours ago | hide | 220 comments

3. ▲ Oceangoing robots that could revolutionize fishing, drilling, and science (bloomberg.com)
   144 points by walrus01 7 hours ago | hide | 55 comments

4. ▲ CSCI E-23a: Introduction to Game Development (cs50.github.io)
   113 points by sharjeelsayed 3 hours ago | hide | 5 comments

5. ▲ The Stunning, Sacred Retreats of California (2013) (nytimes.com)
   11 points by davidjnelson 1 hour ago | hide | 1 comment

6. ▲ History of Symbolics lisp machines (2007) (danluu.com)
   49 points by kristianp 6 hours ago | hide | 3 comments

7. ▲ The Markov Property, Chain, Reward Process and Decision Process (xaviergeerinck.com)
   96 points by thebillkidy 7 hours ago | hide | 16 comments

8. ▲ Understanding Image Sharpness (normankoren.com)
   85 points by brudgers 7 hours ago | hide | 13 comments

9. ▲ AsteroidOS – An open-source operating system for smartwatches (asteroidos.org)
   167 points by skbohra123 13 hours ago | hide | 36 comments

10. ▲ Abusing Proxies for DSLs (mcluck90.github.io)
    42 points by fagnerbrack 6 hours ago | hide | 5 comments

11. ▲ Python as a declarative programming language (2017) (benfrederickson.com)
    63 points by akashtndn 6 hours ago | hide | 17 comments

12. ▲ The Physics and Physicality of Extreme Juggling (wired.com)
    41 points by ColinWright 5 hours ago | hide | 20 comments

13. ▲ Why Finland's cities are havens for library lovers (theguardian.com)
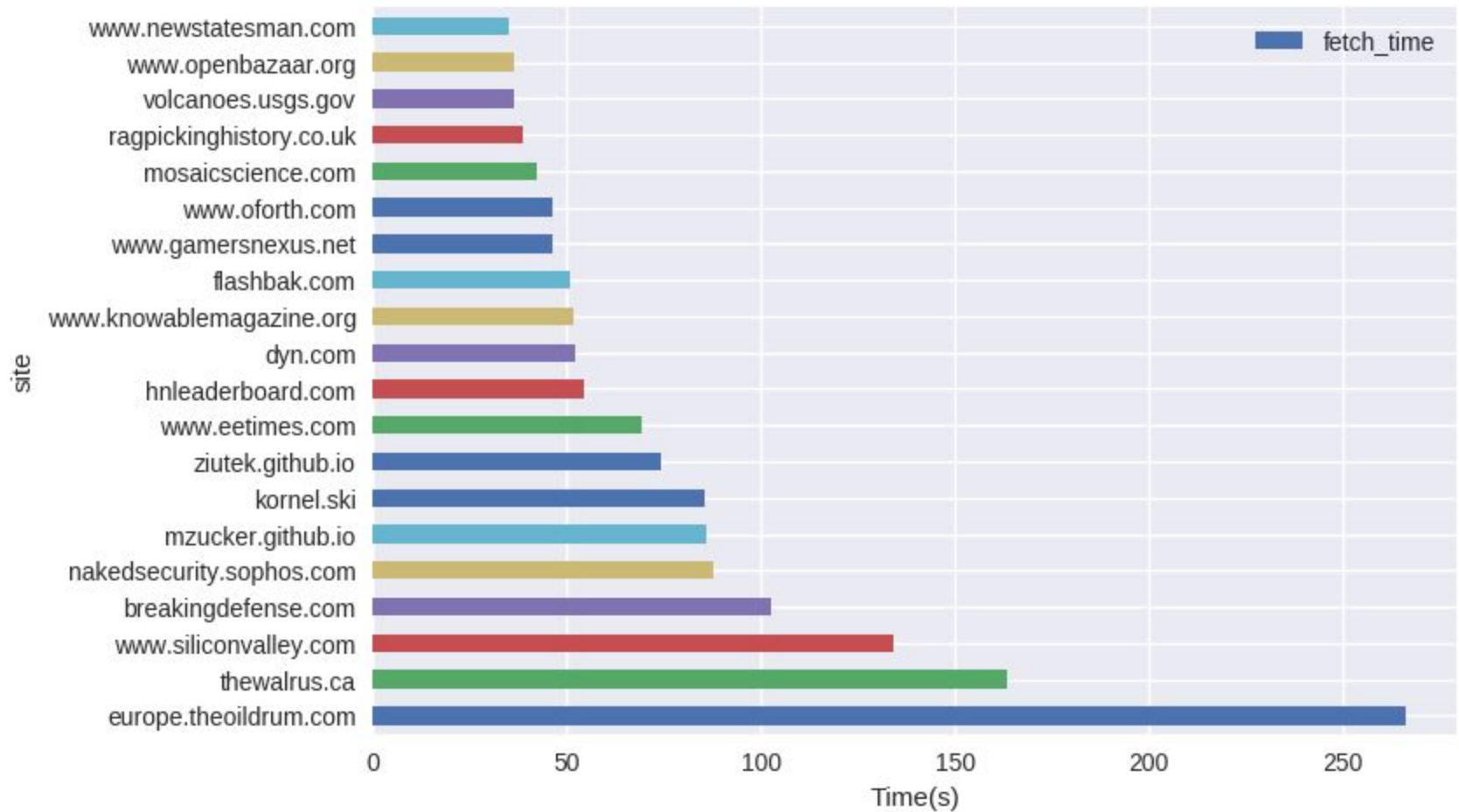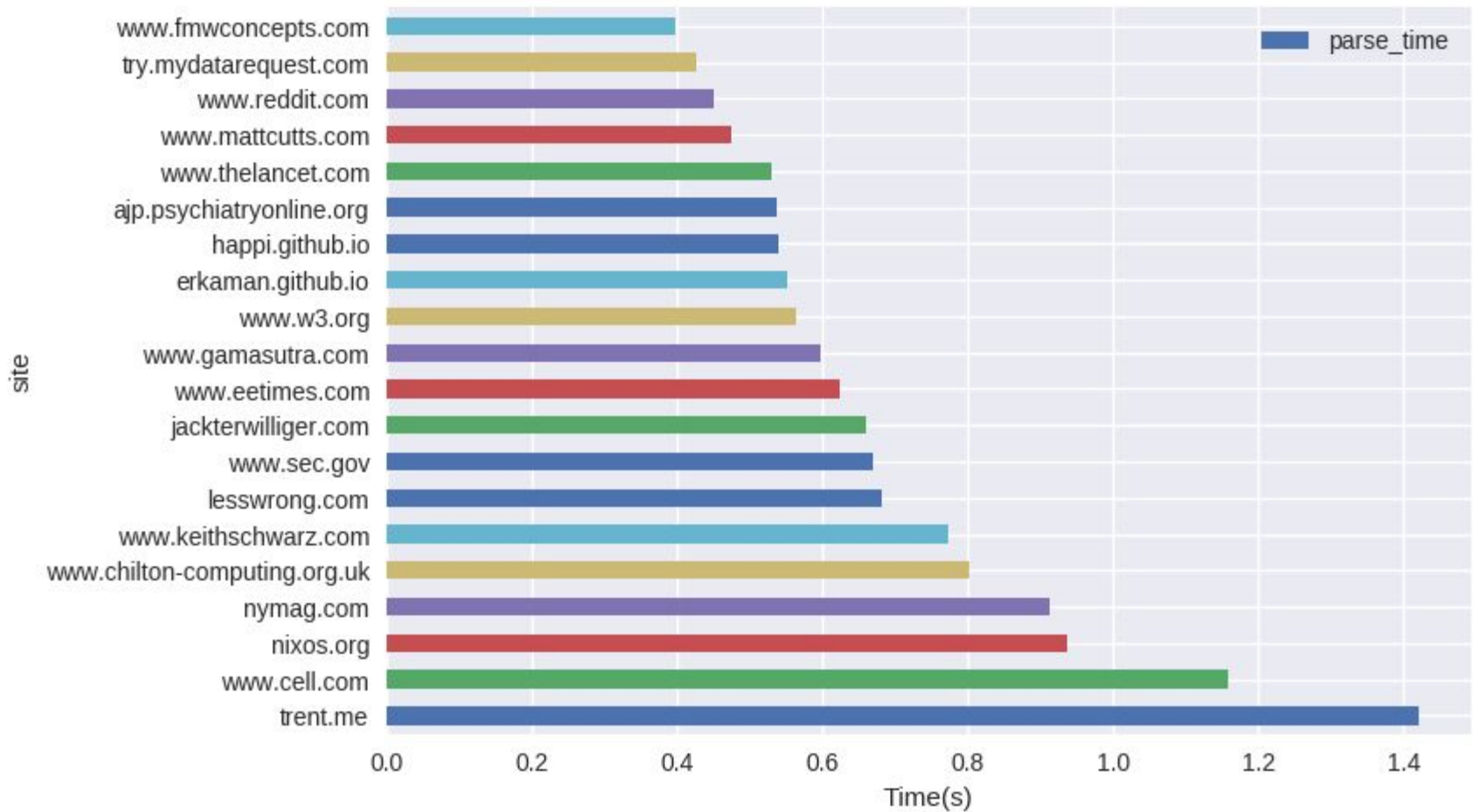   153 points by jogjayr 11 hours ago | hide | 54 comments

Top 20 Sites
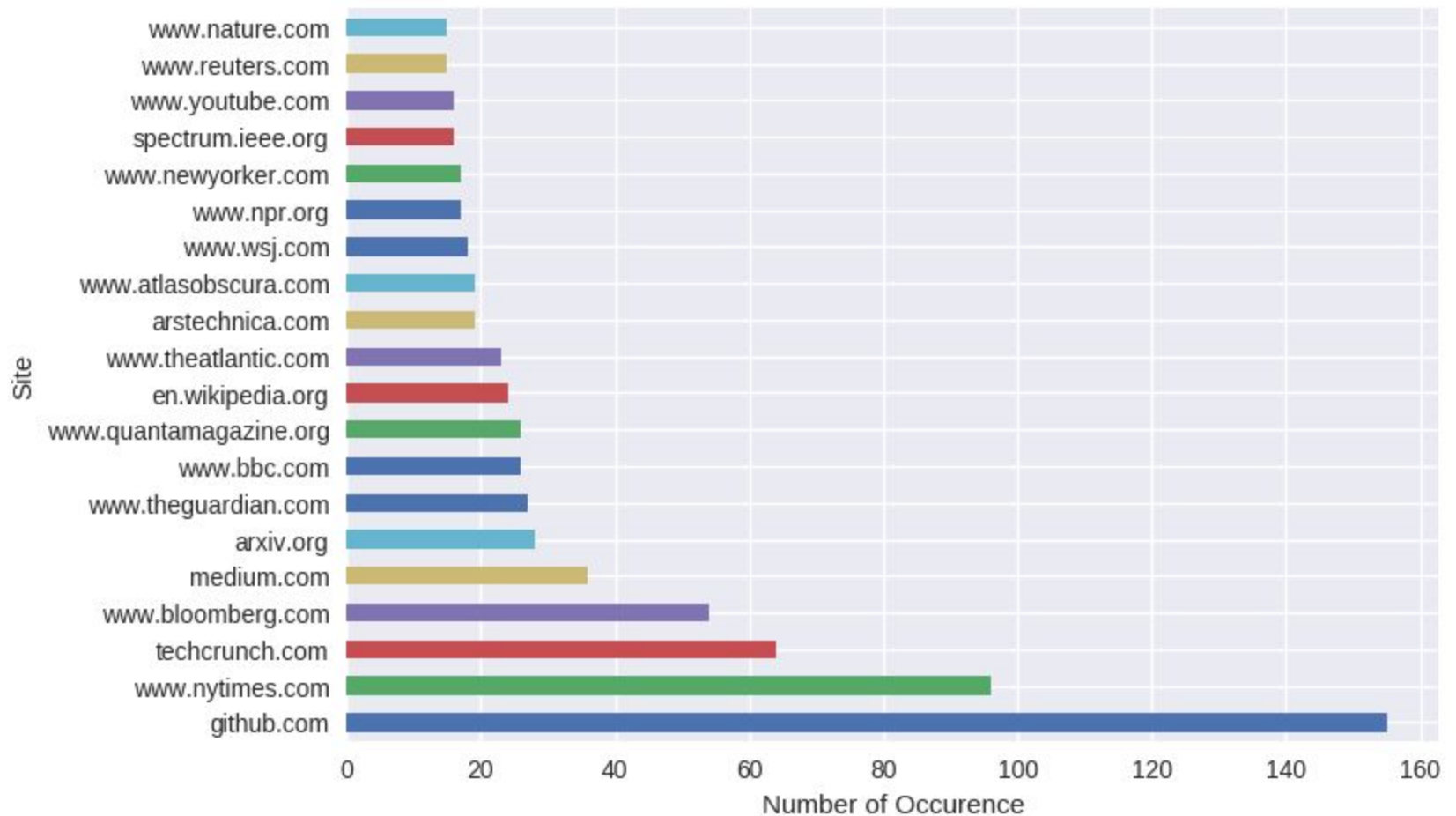
Result

- Take 5 hours

- Use Selenium + Chrome because Chromote has bugs.

Top 20 Average Fetch Time

Top 20 Average HTML Parse Time

Conclusion

- Selenium + Chrome is robust
  - Not Crashing when crawling
  - But difficult to set Timeout
- NLP is so difficult to learn
  - Want something like *"Introduction to NLP and its Application"*
  - Especially, applied to Chinese