Common Crawl Mining

Team:	Brian Clarke, Tommy Dean, Ali Pasha, Casey Butenhoff
Manager:	Don Sanderson (Eastman Chemical Company)
Client:	Ken Denmark (Eastman Chemical Company)
Instructor:	Edward Fox

Multimedia, Hypertext, and Information Access (CS 4624)
Virginia Tech, Blacksburg, VA 24061
Spring 2017

Project Goals

- Improve Eastman Chemical Company's ability to inform key business decisions
  - Assist product marketing and strategy development
- Generate information for trend analysis of keywords
  - Publication date extraction
  - Common Crawl keyword searches

History

- Established client contact
- Experimented with MapReduce searching
- Transitioned to Elasticsearch[1]
  - Reduced dataset
  - Configured prototype
- Tested Elasticsearch prototype
- Refined Elasticsearch prototype
  - Parallelization
  - Custom relevance scoring
  - Deduplication

Problems

- EMR cluster crashing (node size and default timeout)
  - Change configurations
- Webhose printing to stdout
  - Dup2()
- JavaScript in HTTP responses
  - Beautiful Soup
- Alternatives to Webhose date extraction (11,031/43,748)
  - IBM AlchemyLanguage
- Scalability
  - Multiprocessing

Demo

- Querying[2]
- Visualization[2]

Consider all problems you may face and construct solutions ahead of time.

Acknowledgements

- Ken Denmark (Eastman Chemical Company)
- Dr. Edward Fox (Virginia Tech)
- Liuqing Li (Virginia Tech)
- Don Sanderson (Eastman Chemical Company)

References

[1] "Elastic MapReduce Using Python and MRJob." Elastic MapReduce Using Python and MRJob - Cs4980s15. MoinMoin, 2015-02-12 Web. https://weblog.cs.uiowa.edu/cs4980s15/Elastic%20MapReduce%20using%20Python%20and% 20MRJob 19 Feb. 2017.

[2] “Kibana: Explore, Visualize, Discover Data”. Elastic. 10 Apr. 2017. Web. https://www.elastic.co/products/kibana. 10 Apr. 2017.

[3]: Common Crawl. Common Crawl, 2017 Web. http://commoncrawl.org/the-data/get-started/ 19 Feb. 2017. http://commoncrawl.org/the-data/get-started/

Thanks!

Questions?

[3]