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
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\textsuperscript{[2]}
- Visualization\textsuperscript{[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


Thanks!

Questions?