Automating website QA with Quotemine

Open Source Web Crawler

Michael Nolan, PhD
Cerium Software LLC

Today's Topics

- What is Quotemine?
- QA with QuoteMine
  - Setting up your tests
  - Site Health: Broken Links/Site Structure
  - Regex searches
  - Caching/SEO
  - Load Testing / Continuous monitoring
- Future work and additional use cases

Cerium

What is Quotemine?

- Python Web Crawler
  - Uses Requests, LXML, BeautifulSoup
  - Systematically crawls all pages in a domain
- Main crawler is pluggable
  - Add/remove/config plugins as needed
- Basic report generation
  - CSV reporting tool to perform basic categorization and statistics

Cerium

Quotemine's history

- Pre-Quotemine: Work at ReachLocal
  - QA scripting for company websites & forms
  - International SEO
- First development and use
  - Content replacement on erc-assoc.org
  - “Which 70 of 2,300 pages have this name?”
- Open Source Project
  - Launched Dec. 29, 2015
  - Media coverage & sentiment study

Cerium

Quotemine hook system

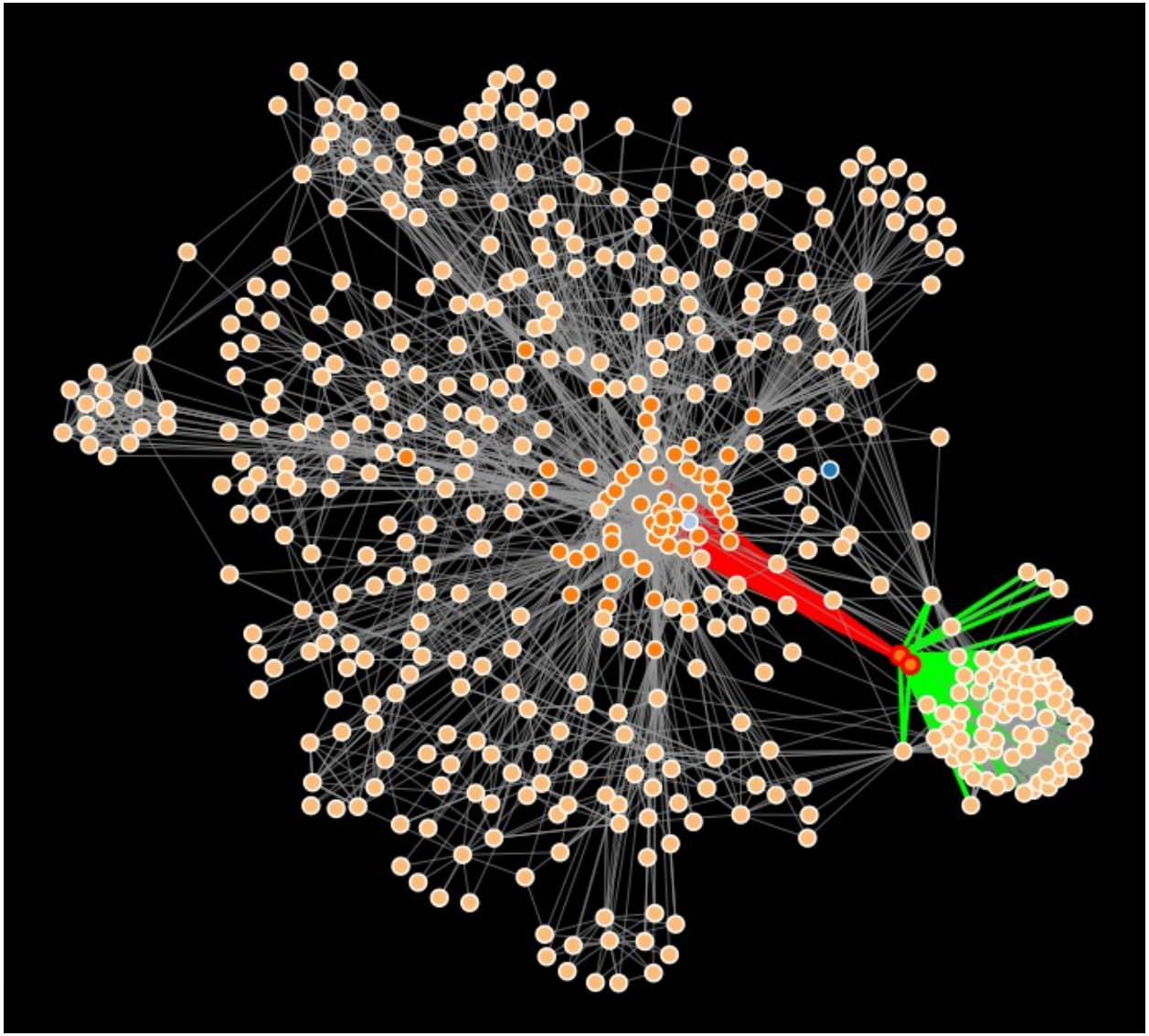
Miner step	Hooks
Import/Load	Process
	Postprocess
Mine	Preprocess
	Process
	Postprocess
Process URL	Parse
	Postprocess
Process Data	Process
Generate Reports	Preprocess
	Process
	Postprocess
Export/Save data	Preprocess
	Process
	Postprocess
Reset Data	

Quotemine PageParser hooks

Parser Step	Hooks
Parse document	Parse Tag
	Parse data in tag

Cerium

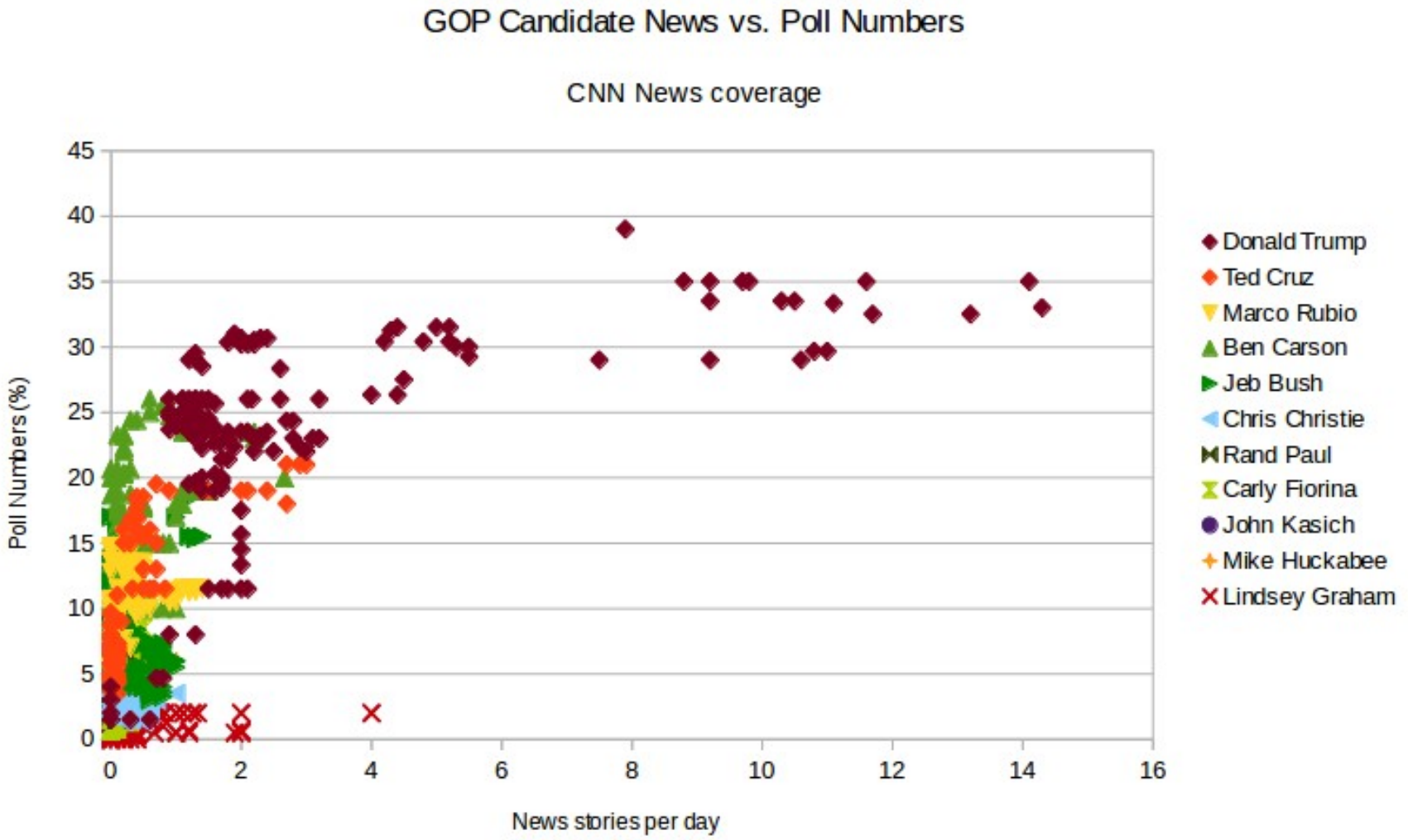
Quotemine Data Examples



EPA.gov linkmap

Cerium

Quotemine Data Examples



Quotemine Data examples

	A	B	C	D
1	author	title	Donald Trump	Donald Trump
2	Michael Tanner	Greece and America -- Too Similar for Comfo	negative	-0.640691
3	Kevin D. Williamson	Donald Trump's 2016 Debate Lies: He Did G	negative	-0.539474
4	David French	Ted Cruz Defeats Donald Trump in GOP Deb	negative	-0.43995
5	Veronique Rugy Brenda	The Corner | National Review Online	negative	-0.434278
6	Andrew C. Mc Carthy	Hillary Clinton's E-mail -- Felony Violations	negative	-0.432525
7	Andrew C. Mc Carthy	Hillary Clinton's E-mail -- Felony Violations	negative	-0.432525

	A	B	C	D
1	author	title	Donald Trump	Donald Trump
2	Jim Geraghty	Donald Trump's Rise Foreshadowed by CPA	positive	0.528398
3	Donald R. Brand David I	Donald Trump	positive	0.322581
4	Donald R. Brand David I	Donald Trump	positive	0.322581
5	Victor Davis Hanson	Donald Trump	positive	0.259788
6	Victor Davis Hanson	Donald Trump	positive	0.259788
7	Elaina Plott	Speaker Ryan Chats with Former Speaker N	positive	0.181751
8	Mona Charen	Donald Trump	positive	0.177872

Cerium

QA with Quotemine

- Quotemine is primarily a **content** analyzer
  - Examines client-side content
  - Cannot see orphaned content, PHP code
    - Combine with New Relic, etc.
- Custom plugins
  - Analyze request
  - Combine with Selenium/Ghost.py to execute JS
  - Send content to 3rd party API for storage/analysis
  - Make additional GET and POST requests

Cerium

Setting up your QA project

What are you looking for?

- Broken Links
- Content/URLs that shouldn't be exposed
- Hard-to-find pages
- Slow/uncached pages
- Background process load

Cerium

Quotemine workflow

- Build URL list
  - Manual feed homepage or load save file/template
- Crawl site
  - Grab metadata, load performance, content of interest
- Generate reports
  - CSV output of all data by URL/revision
  - Category reports with CSVManager
- Export Save file

Cerium

Broken/exposed links

- Miner Core
  - Gets HTTP status of links
- LinkFilter
  - Link discovery engine; finds and categorizes URLs
  - Determines click depth
  - Generates a structure map of URLs by section
- SiteMapper
  - Generates visualizations of site structure
  - Maps click depth and gateway pages

Cerium

Exposed Content

- Miner Core
  - Grabs text from each request for analysis
- TextAnalyzer
  - Can parse all words into word counts/densities
  - Regex-based searches of tag content or full pages
- SentimentAnalyzer
  - Feeds content to IBM Watson API
- AuthorFilter/DateFilter
  - Purpose built for extracting page author/post date

Cerium

Hard-to-find/should-be-hidden pages

- Miner Core
  - Logs redirects
- LinkFilter
  - Lists all found URLs, and pages on which they are located
  - Also reads links in metadata (i.e. short form)

Cerium

Slow pages

- Miner Core
  - Logs page load times
  - Use repeat_mine() to test caching
- HeaderFilter
  - Reads request headers
  - Useful for grabbing Drupal and Varnish Cache Information
- Pair with Charles Proxy for more load info

Cerium

Slow processes

- Pair QuoteMine with New Relic
  - QuoteMine mimics anonymous user behavior
  - Use multiple instances to increase load
- Linkfilter configuration
  - Limit allowed URLs to specified site sections
  - Include/exclude URLs by type
    - Internal, External, and File type URLs

Cerium

Continuous Monitoring

- repeat_mine() function
  - Repeat runs on an interval basis
  - Optionally break reports into time periods
  - Tweak timings to maintain continuous monitoring for extended periods

Cerium

Additional Use Cases

- Content scraping and analysis
  - News topic coverage
  - Mapping topics to authors
  - Sentiment analysis
- Web-based databases
  - Pull down online databases which don't have a download option

Cerium

Future Work

- Ghost.py/Selenium integration
  - Execute JS to load dynamic content
  - Screen capture
  - Interact with content (test forms/AJAX)

- Expanding BeautifulSoup integration
  - Use CSS selectors instead of traversing HTML structure

Cerium

Thank you for your time!

- My email
  - mike@ceriumsoft.com
- QuoteMine repo:
  - http://bit.ly/cerium-quotemine

Cerium