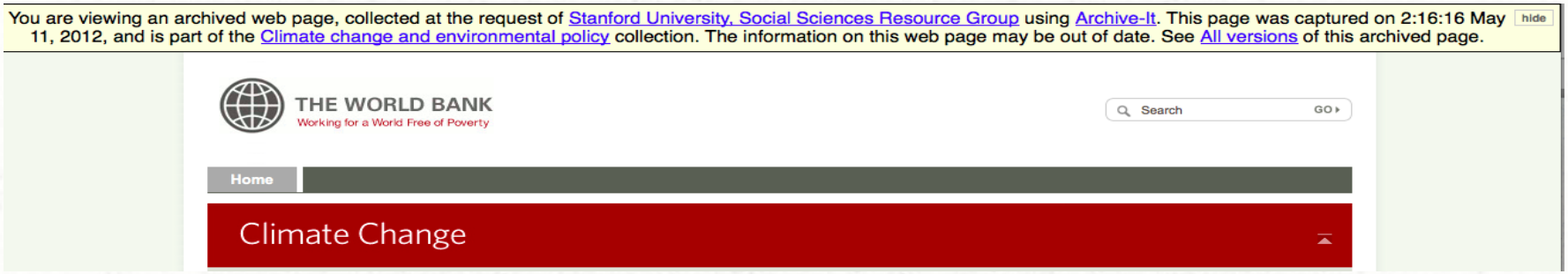
Web Archive Analysis

Vinay Goel
Senior Data Engineer
Internet Archive
vinay@archive.org

Access Archive-It Data

Wayback Machine



Archive-It Search

Download Archive-It Data

- Download crawl data
  - Use command-line tools / browser
- Download collection, seed and document level metadata
  - Use Archive-It Web Application / OAI-PMH XML feed
- About 25% of all partners download their collection data
- The rest are encouraged to do the same
- Reasons to download crawl data:
  - Satisfy specific preservation requirements
  - Provide access to restricted collections
  - Custom reporting, wayback and search
  - Custom research & analysis


ACT! for America
Brigitte Gabriel
ACT! for America
Welcomes You...
CONTINUE TO THE WEBSITE
ADC
ADL
100
Enable Research & Analysis

Data: WARC

- Data written by web crawlers
- Web ARchive Container File Format (ISO standard)
- Each file contains a series of concatenated records
  - Full HTTP request/response records
  - Metadata records (links, crawler path, encoding etc.)
  - Records to store duplicate detection events
  - Records to support segmentation and conversion
- Usually 1 Gigabyte in size

Derived Data: CDX

- Index for Wayback Machine
- Space delimited text file
- Contains only essential fields needed by Wayback
  - URL, Timestamp, Content Digest
  - MIME type, HTTP Status Code
  - WARC filename and file offset of record
  - Optional: redirect URL, meta tags, size

Wayback Machine

Clinical and Translational Science Awards Web Archive (National Institutes of Health)

Enter Web Address: http:// All Take Me Back

Searched for http://casemed.case.edu/ctsc/ 11 Results RSS Metadata

Look up URL in general Internet Archive web collection Proxy Mode Help

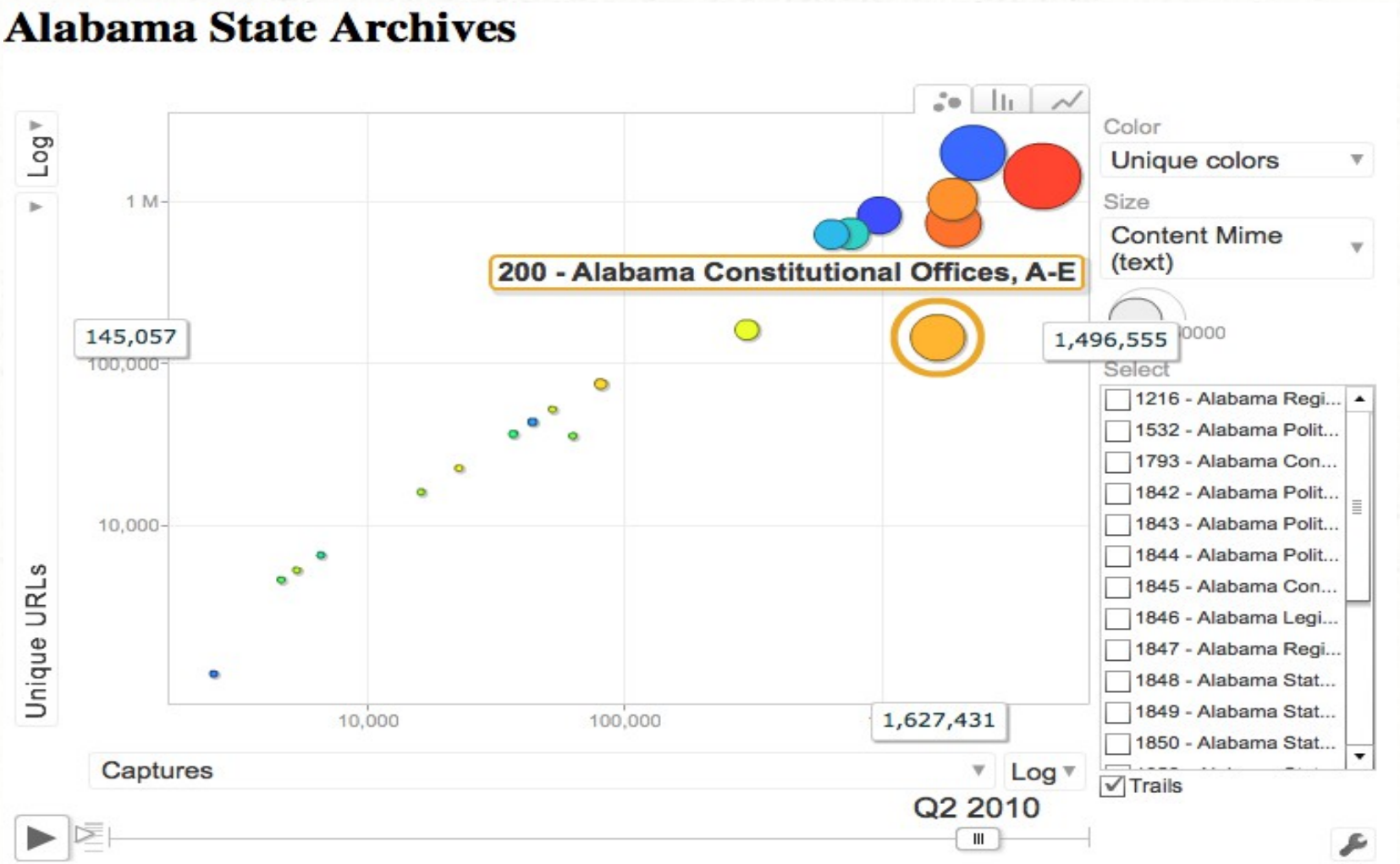
* denotes when page was updated

Found 11 Captures between Oct 14, 2010 - Oct 22, 2012

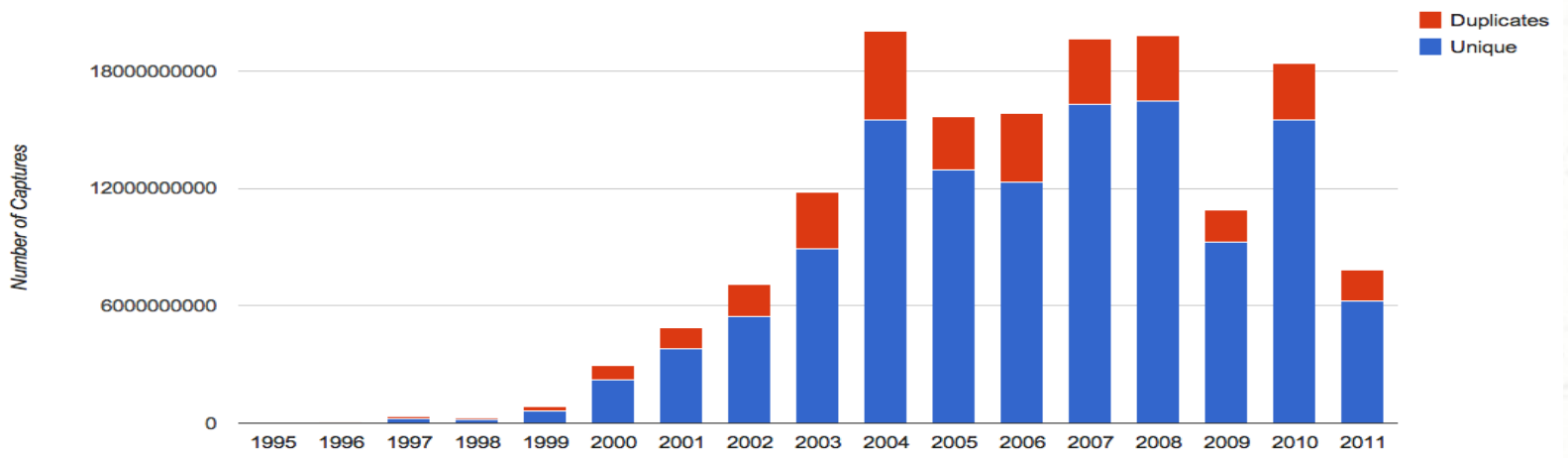
2010	2011	2012
3 pages	4 pages	4 pages
Oct 14, 2010 *	Jan 22, 2011 *	Jan 22, 2012 *
Oct 14, 2010 *	Apr 22, 2011 *	Apr 22, 2012 *
Oct 22, 2010 *	Jul 22, 2011 *	Jul 22, 2012 *
	Oct 22, 2011 *	Oct 22, 2012 *

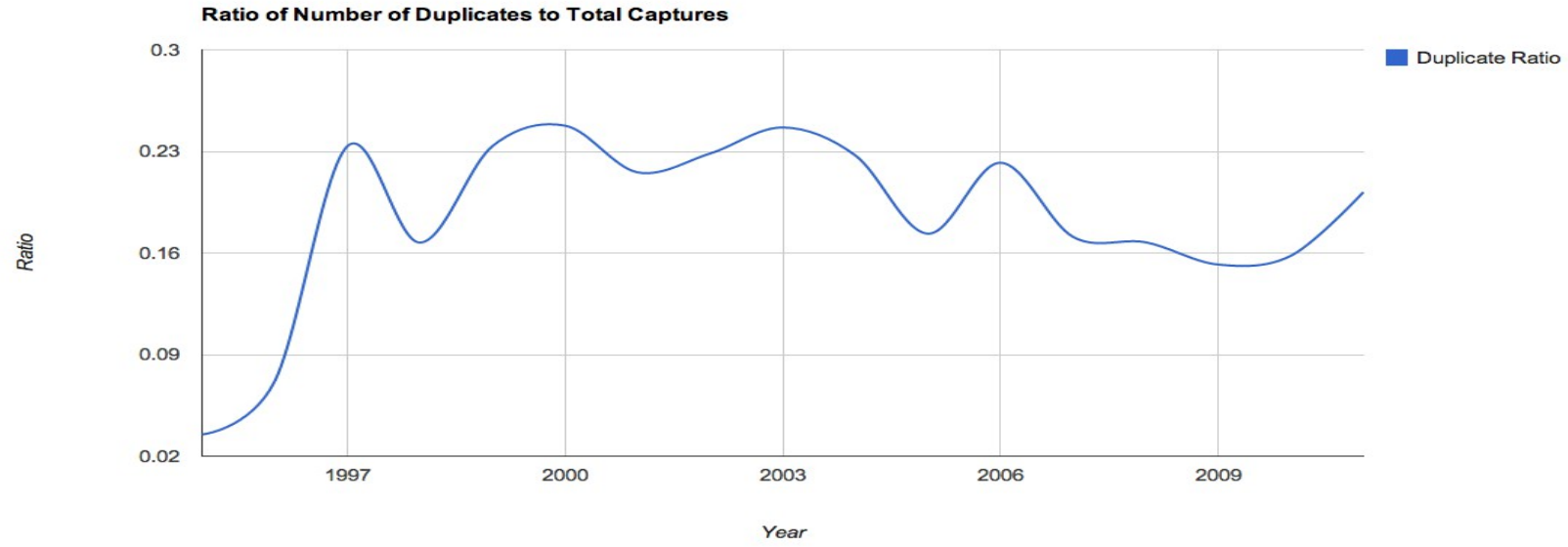
Home | Internet Archive

Track growth of content



Find rate of duplication





Derived Data: Parsed Text

- Input to build text indexes for Search
- Text is extracted from WARC files
- HTML boilerplate is stripped out
- Also contains metadata for each record
  - URL, Timestamp, Content Digest
  - MIME type, HTTP status code
  - Record Length
- Stored in Hadoop Sequence Files

Derived Data: WAT

- Metadata format
- Essential metadata for many types of analyses
- Avoids barriers to data exchange: copyright, privacy
- Less data than WARC, more than CDX
- Extensible
- WAT records are WARC metadata records

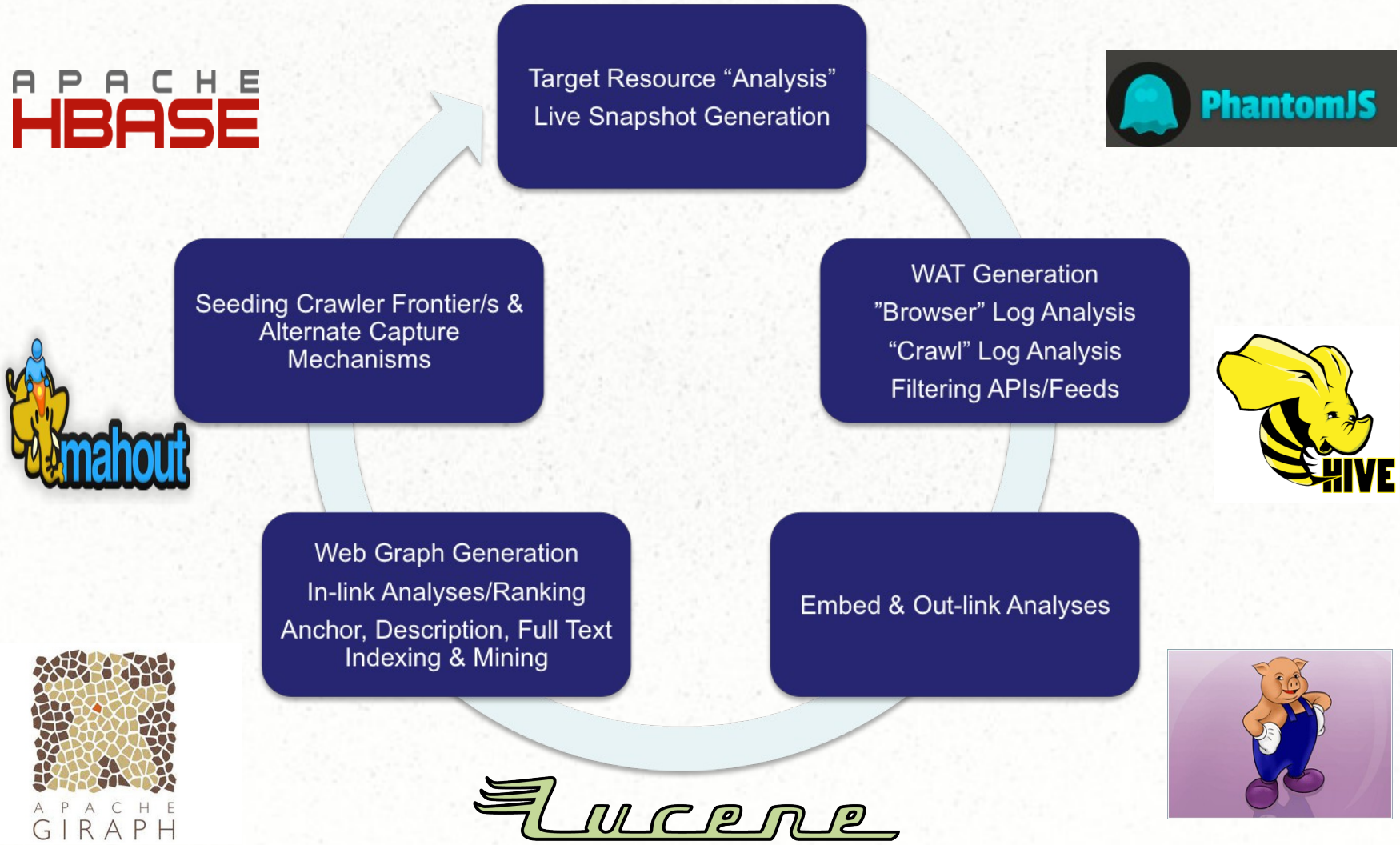
Derived Data: WAT (current version)

- Contains all metadata needed for CDX
- 1:1 mapping to records in the corresponding WARC
- Contains for every HTML page in the WARC,
  - Title text
  - Description text
  - Meta keywords
  - Embeds and outgoing links with alt / anchor text

Analysis Toolkit

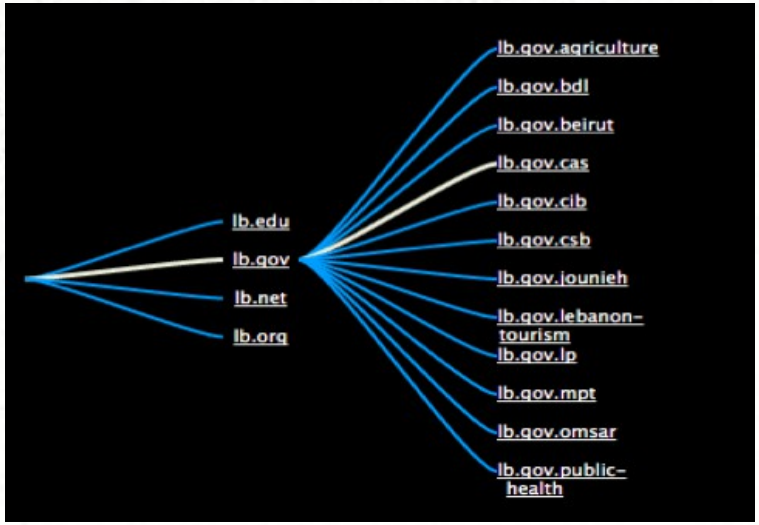
- Enable arbitrary analysis of web archives
- Easy and quick to learn
- Scales up and down
- Tools
  - Hadoop (HDFS, MapReduce)
  - Apache Giraph
  - Apache Hive
  - Apache Mahout
  - Apache Pig

Analysis Workflow



Crawl Log Analysis (Hive/Giraph)

- Crawl Log Warehouse
- Distribution of HTTP status codes, MIME types
- Find timeout errors, duplicate content, crawler traps, robots exclusions
- Trace path of the crawler



Combine with text metadata to fill in missing fields

Youtube Videos

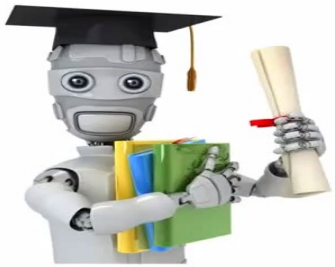
<< < > >> Rows per page: 20 Jump to page: 1

Page 1 of 6 -- displaying rows 1-20 of 109

Title ▲/▼	Size (bytes) ▲/▼	Date ▲/▼	Discovered from Seed ▲/▼	Referring Page (Live Web) ▲/▼	URL of video
Funding for Williston Park Road Improvements - YouTube	24627453	20121221	http://ackerman.house.gov/	www.youtube.com	Play (from Archive) Download (from Archive)
Supporting State Department Authorization Bill - YouTube	5494322	20130101	http://ackerman.house.gov/	www.youtube.com	Play (from Archive) Download (from Archive)
Credit Ratings and the First Amendment - YouTube	16804391	20130101	http://ackerman.house.gov/	www.youtube.com	Play (from Archive) Download (from Archive)
Broader Focus on Afghanistan - YouTube	3807340	20121221	http://ackerman.house.gov/	www.youtube.com	Play (from Archive) Download (from Archive)

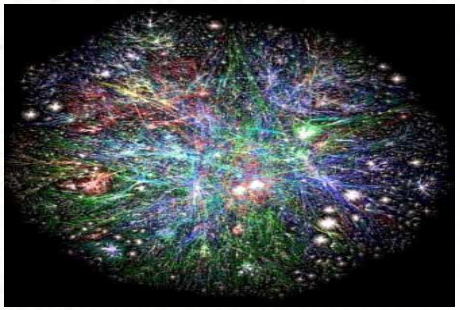
Text Analysis (Pig/Mahout)

- Text extracted from WARC / Parsed Text / WAT files
- Use curated collections to train Classifiers
- Cluster documents in collections
- Topic Modeling
  - Discover topics
  - Study how topics evolve over time
- Compare how a page describes itself (meta text) vs. how other pages linking to it describe it (anchor text)



Link Analysis (Pig/Giraph)

- Links extracted from crawl logs / WARC metadata records / Parsed Text / WAT files
- Indegree and Outdegree information
- Inter-host and Intra-host link information



- Study how linking behavior changes over time
- Rank resources by PageRank
  - Identify important resources
  - Prioritize crawling of missing resources
- Find possible spam pages by running biased PageRank algorithms

Questions?

- E-mail: vinay@archive.org
- Code Repository
  - Https://github.com/internetarchive/
  - Https://github.com/vinaygoel/archive-analysis/