NATIONAL DIGITAL
INFORMATION INFRASTRUCTURE
AND PRESERVATION PROGRAM

Open Source Tools for Mining and Analysing Web Data @ Scale

Kris Carpenter Negulescu, Internet Archive

Annual Meeting, Washington DC

July 20, 2011

Key Problems to Address & Primary Benefits…

Archived Web Data is often isolated, difficult to link to other related resources by topic, and minimally navigable

Benefits of mining and analysis:

Mapping relationships between links over time
Geo-location maps
Tag clouds
Classification
Facets
Rate of change
Related information; Enhanced keyword search

The Tool Box

- ➔ HDFS
- ➔ Map Reduce
- ➔ Pig Latin
- ➔ Web archive code – metadata extraction jar
- ➔ Other extraction layers: Tika, Jhove(2), etc
- ➔ Google analytics APIs/Drupal modules, Neo4j, etc.

Web Archive Transformation (WAT) - a structured way of storing metadata generated by Web Crawls

- ➔ ARCs and WARCs are “heavy”
- ➔ WAT – Web Archive Transformation file
  - Uses WARC format as a generic meta data container
  - Extract everything you're likely to want from ARCs/WARCs once
- ➔ Store into HDFS; Part of standard ingest process

Web archive code: metadata extractor

➔ The WAT utilities produce structured metadata that is optimized for data analysis, i.e. JavaScript Object Notation (JSON), from compressed (GZIPed) or uncompressed ARC or WARC files.

- Currently just a bit of glue code around an ARC/WARC reader whose function is HTML metadata extraction
- JSON data is written to STDOUT in compressed (GZIP) format. The ARC or WARC file can be a local file, a HTTP accessible file (http://), or an Hadoop File System (HDFS) accessible file (hdfs://).

➔Includes example “UDF” code

➔Will integrate with Jhove(2), Tiki, etc