ApacheTikaHtmlEncodingStudy

Apache Tika's Html Encoding Study

In support of TIKA-2038, we gathered a new subset of html pages from CC-MAIN-2017-04.

This page offers a first rough draft of the process. Some of the code is available on a personal github site. This code relies heavily on Dominik Stadler's CommonCrawlDocumentDownload code, and the author of SimpleCommonCrawlExtractor is extremely grateful to Dominik.

1. Determined which top level domains (TLDs) were of interest 2. Downloaded the 300 index files from Common Crawl via Groovy (217 GB of data):

```
def cc = "CC-MAIN-2017-04"
def url1 = "https://commoncrawl.s3.amazonaws.com/cc-index/collections/"
def url2 = "/indexes/cdx-"

(0..299).each{ i ->
    def u = url1+cc+url2+"$i".padLeft(5, '0')+".gz"
    def p = "wget -q $u".execute()
    p.waitForProcessOutput(System.out, System.err);
}
```

3.#3 Counted the number of pages per TLD that had "html/text" in the http Content-Type header
Map:

```
java -cp cc-extractor-0.0.1.jar org.tallison.cc.index.CCIndexBatchReader
    10 /data1/commoncrawl_indices/CC-MAIN-2017-04/ CountMimesByTopLevelDomains
    mime_tld_counts
```

Reduce:

```
java -cp cc-extractor-0.0.1.jar org.tallison.cc.index.reducers.DoubleKeyReducer
    mime_tld_counts mime_tld_total.txt
```

4.#4 Created sampling frequencies per TLD, with a target of 50k per TLD, with the exception of 100k for ".com" – this was done by loading mime_tld_total.txt into a database and doing some `group by` queries. See tld_mimes.txt.
5. Randomly sampled according to the sampling frequencies per TLD from the 300 index files
Map:

```
java -cp cc-extractor-0.0.1.jar org.tallison.cc.index.CCIndexBatchReader
    10 /data1/commoncrawl_indices/CC-MAIN-2017-04/ DownSample
    tld_mimes.txt tld_mimes_down_sampled
```

Reduce:

```
java -cp cc-extractor-0.0.1.jar org.tallison.cc.index.reducers.ConcatReducer
    tld_mimes_down_sampled tld_mimes_down_sampled_index
```

6.#6 Pulled the data from Common Crawl

```
java -cp cc-extractor-0.0.1.jar org.tallison.cc.CCGetter
    tld_mimes_down_sampled_index /data4/docs/commoncrawl_html_study
    cc_html_study_crawl_status.txt
```