Collection Management

Presenters: Yufeng Ma & Dong Nan
May 3, 2016

CS5604, Information Storage and Retrieval, Spring 2016
Virginia Polytechnic Institute and State University
Blacksburg VA
Professor: Dr. Edward A. Fox

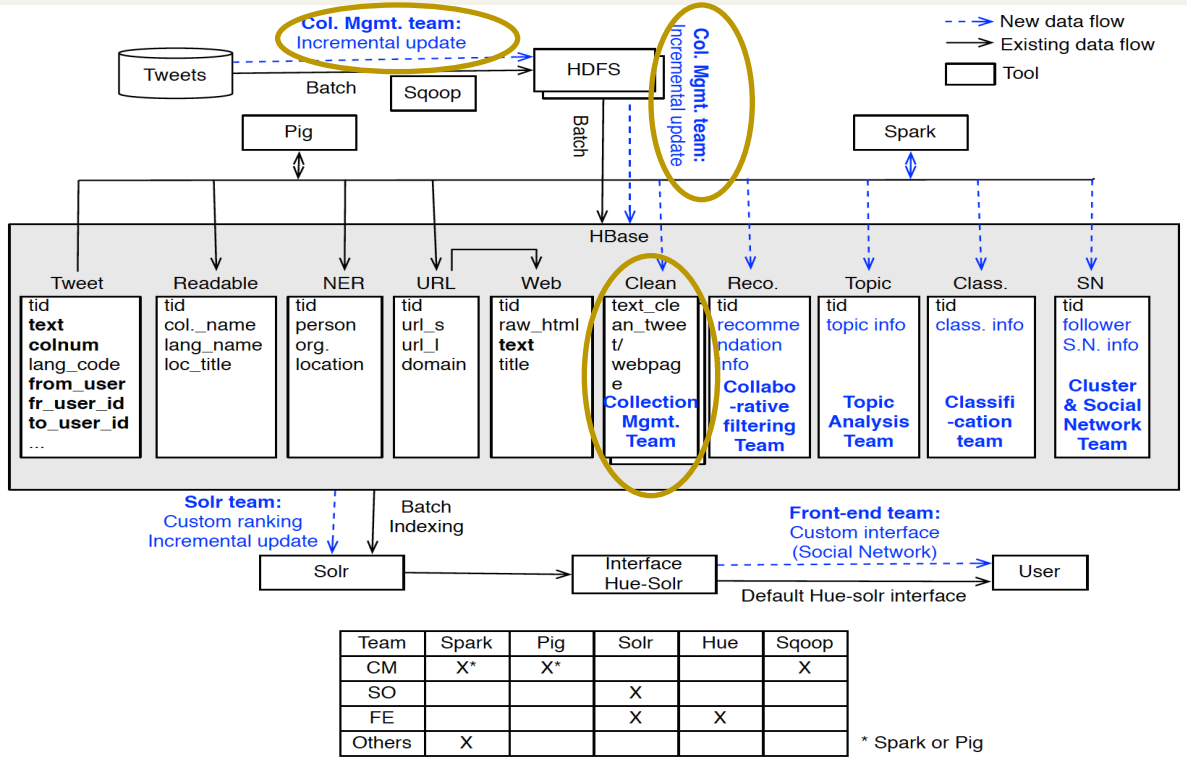
Outline

- Goals & Data Flow
- Incremental update
- Tweet cleaning
- Webpage cleaning

Goals

- Keep data in HBase current
- Providing “quality” data
  - Identify and remove “noisy” data
  - Process and clean “sound” data
  - Extract and organize data

Data Flow Diagram



Team	Spark	Pig	Solr	Hue	Sqoop
CM	X*	X*			X
SO			X		
FE			X	X	
Others	X				

* Spark or Pig

Image credit: Sunshin Lee

Incremental Update

Incremental Update: MySQL → HDFS

- Previous bash script importing 700+ tables, without incremental feature.
- Incremental import new rows in the relational database (MySQL) to HDFS.
- Use *incremental append* mode of Sqoop to import data incrementally.

```
sqoop import \
  --connect jdbc:mysql://mysql.example.com/sqoop \
  --username sqoop \
  --password sqoop \
  --table visits \
  --incremental append \
  --check-column id \
  --last-value 1
```

Code credit: Kathleen Ting and Jarek Jarcec Cecho. *Apache Sqoop Cookbook*. Sebastopol: O'Reilly Media, Inc., 2013. Print.

Incremental Update: HDFS → HBase

- Keep HBase in sync with imported data on Hadoop.
- Write Pig script to import new data from HDFS to HBase.
- Use job scheduler *Cron* on Linux (by creating *crontab* file), periodically run the Pig script.

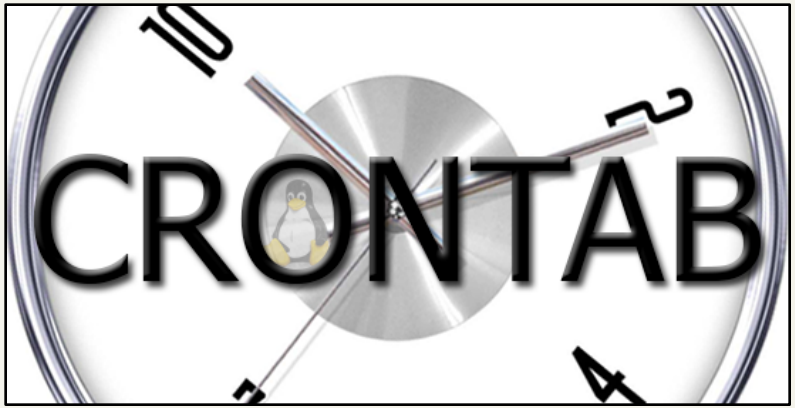


Image credit: http://itekblog.com/wp-content/uploads/2013/03/crontab.png

Tweet Cleaning

Tweet: Text Cleaning and Info Extraction

- Remove URLs, profanities, and non-characters from raw tweets.
- Extract short URLs from raw tweets, expand, and map to corresponding web pages.
- Extract hash tags (#) and mentions (@) out from raw tweets.
- Store cleaned text, extracted hash tags and mentions from HDFS node to HBase.
- All the cleaning, extracting and storing process done by Pig Latin.

rowkey	**clean_tweet**							
collection # - tweet id	clean_text	urls	hashtags	mentions	mappings	empty	collection	doctype

Tweet: Text Cleaning and Info Extraction (Example)

```
541-552940627672174595  FBI looks for motive in explosion near Colorado @NAACP office:
 http://t.co/qiA1hAiTll #NAACPBombing http://t.co/4OxJDhAENa
```

Raw tweet on HDFS

```
hbase(main):001:0> get 'ideal-cs5604s16', '541-552940627672174595', {COLUMNS => 'clean_tweet'}
COLUMN                          CELL
 clean_tweet:clean_text         timestamp=1460396472264, value=FBI looks for motive in explosion near Colorado NAACP office  NAACPBombing
 clean_tweet:collection         timestamp=1461350976814, value=#NAACPBombing
 clean_tweet:empty              timestamp=1461348659084, value=0
 clean_tweet:hashtags           timestamp=1461179962849, value=#NAACPBombing
 clean_tweet:lurls              timestamp=1461862933472, value=https://twitter.com/WMCActionNews5/status/552940627672174595/photo/1
 clean_tweet:mappings           timestamp=1461180450804, value=https://twitter.com/WMCActionNews5/status/552940627672174595/photo/1
 clean_tweet:mentions           timestamp=1462201316377, value=@NAACP
 clean_tweet:urls               timestamp=1461183196044, value=http://t.co/4OxJDhAENa
8 row(s) in 0.1300 seconds
```

Cleaned tweet and extracted info in HBase

Webpage Cleaning

Raw Data

700-636800318748106752 http://www.theguardian.com/us-news/2015/aug/27/virginia-shooting-what-we-know-so-far-alison-parker-adam-ward-vester-flanagan-wdbj7:URL:<!DOCTYPE html> <html id="js-context" class="js-off is-not-modern id--signed-out" data-page-path="/us-news/2015/aug/27/virginia-shooting-what-we-know-so-far-alison-parker-adam-ward-vester-flanagan-wdbj7"> <!-- ##::::: ##: #######::::::: ###:::: ########:: ########:::: ##:::: ##: ####: ########:: ####: ##::: ##:: ######:: ##: ##: ##: ##.....:::::::: ## ##::: ##.... ##: ##.....::: :: ##:::: ##:. ##:: ##.... ##:. ##:: ###:: ##: ##... ##: ##: ##: ##: ##::::::::::: ##:. ##:: ##:::: ##: ##:::::::::: ##:::: ##:: ##:: ##:::: ##:: ##:: ####: ##: ##:::..:: ##: ##: ##: ######:::::: ##:::. ##: ########:: ######:::::: #########:: ##:: ########::: ##:: ## ## ##: ##:: ####: ##: ##: ##...:::::: #########: ##.. ##::: ##...:::::: ##.... ##:: ##:: ##.. ##::::: ##:: ##. ####: ##::: ##: ##: ##: ##:: ##:::::::::: ##.... ##: ##::. ##:: ##:: ##:::::::::: ##::::: ##:: ##:: ##::. ##::: ##:: ##:: ##:. ###: ##::: ##: ###. ###:: ########:: ##:: ##::. ##: ########:::: ##:::: ##: ####: ##:: ##: ####: ##::. ##:. ######:: Ever thought about joining us? http://developers.theguardian.com/join-the-team.html --> <head> <meta charset="utf-8"/> <title>Virginia TV shooting: what we know so far | US news | The Guardian</title> <meta http-equiv="X-UA-Compatible" content="IE=Edge"/> <meta name="format-detection" content="telephone=no"/> <meta name="HandheldFriendly" content="True"/> <meta name="viewport" content="width=device-width,minimum-scale=1,initial-scale=1"> <link rel="dns-prefetch" href="https://assets.guim.co.uk/"/> <link rel="dns-prefetch" href="https://i.guim.co.uk"/> <link rel="dns-prefetch" href="https://api.nextgen.guardianapps.co.uk"/> <link rel="dns-prefetch" href="https://hits-secure.theguardian.com"/> <link rel="dns-prefetch" href="//j.ophan.co.uk"/> <link rel="dns-prefetch" href="//ophan.theguardian.com"/> <link rel="dns-prefetch" href="//oas.theguardian.com"/> <link rel="dns-prefetch" href="//beacon.guim.co.uk"/> <link rel="apple-touch-icon" sizes="152x152" href="https://assets.guim.co.uk/images/favicons/451963ac2e2368[illegible]f44e2856d3f04/152x152.png"/> <link rel="apple-touch-icon" sizes="144x144" href="https://assets.guim.co.uk/images/favicons/[illegible]0491[illegible]785da86b/144x144.png"/> <link rel="apple-touch-icon" sizes="120x120" href="https://assets.guim.co.uk/images/favicons/c12308060002e5015869c72f5a904bd/120x120.png"/> <link rel="apple-touch-icon" sizes="114x114" href="https://assets.guim.co.uk/images/favicons/f438f6041a4c1d0289e6debd112880c2/114x114.png"/> <link rel="apple-touch-icon" sizes="72x72" href="https://assets.guim.co.uk/images/favicons/b5050517955e7cf1e493ccc53e64ca05/72x72.png"/> <link rel="apple-touch-icon-precomposed" href="https://assets.guim.co.uk/images/favicons/4fd650035a2cebafea4e210990874c64/57x57.png"/> <link rel="manifest" href="/2015-06-24-manifest.json" crossorigin="use-credentials"> <link rel="amphtml" href="https://www.theguardian.com/us-news/2015/aug/27/virginia-shooting-what-we-know-so-far-alison-parker-adam-ward-vester-flanagan-wdbj7/amp"> <link rel="shortcut icon" type="image/png" href="https://assets.guim.co.uk/images/favicons/79d7ab5a729562cebca9c6a13c324f0e/32x32.ico"/> <link rel="canonical" href="http://www.theguardian.com/us-news/2015/aug/27/virginia-shooting-what-we-know-so-far-alison-parker-adam-ward-vester-flanagan-wdbj7"/> <meta name="apple-mobile-web-app-title" content="Guardian"/> <meta name="application-name" content="The Guardian"/> <meta name="msapplication-TileColor" content="#005689"/> <meta name="theme-color" content="#005689"> <meta name="msapplication-TileImage" content="https://assets.guim.co.uk/images/favicons/f06f6996e193d1ddcd614ea852322d25/windows_tile_144_b.png"/> <link rel="publisher" href="https://plus.google.com/113000071431138202574"/> <meta name="author" content="Claire Phipps"/> <meta name="description" content="Vigils held for Alison Parker and Adam Ward, two WDBJ7 news journalists killed when former colleague Vester Flanagan opened fire on them live on air"/> <meta name="thumbnail"

Webpage Sample

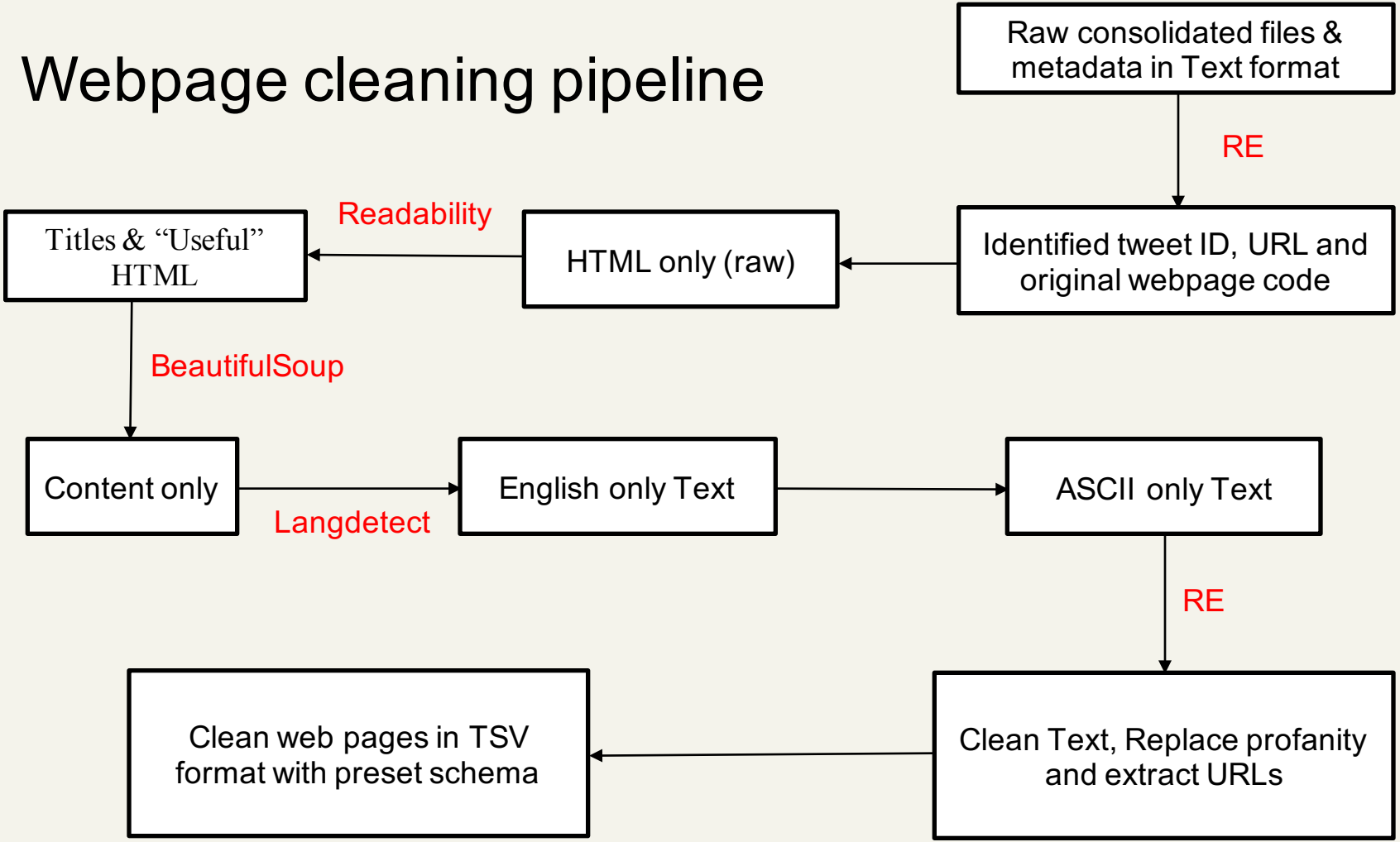
Webpage Cleaning Rules

- Remove Non-ASCII characters
- Keep English text only
- Extract URLs
- Remove profane words

Libraries/Packages

- BeautifulSoup4
  - Parse text out of HTML and XML files
- Readability
  - Pull out the title and main body text from a webpage
- Langdetect
  - Detect language of a text using naive Bayesian filter
- Re
  - Provide regular expression matching operations

Webpage cleaning pipeline



Cleaned Webpage Schema

Rowkey	clean_web						
URL	collection	lang	domain	doc_id	title	text_clean	text_clean_profanity

Rowkey	clean_web				
URL	urls	empty	mappings	doctype	web_original

Cleaned Webpages

```
http://1410wizm.com/index.php/item/26046-walker-obamacare  column=clean_web:text_clean_profanity, timestamp=1460230892402, value="Three Christian universities gained allies Monday in their battle against ObamaCare. Among their supporters: 16 state governments. Those states, along with. handful of other religious rights organizations, filed friend-of-the-court briefs to the Supreme Court supporting Houston Baptist University, East Texas Baptist University, and Westminster Theological Seminary. Those schools have appealed the Supreme Court to overturn. circuit court ruling that forces them to expand contraception options in their health insurance plans. The Becket Fund for Religious Liberty, the schools legal counsel, says the briefs are. major breakthrough. This strong show of support for HBU and ETBU (and Westminster Theological Seminary) demonstrates just how important it is that the Supreme Court address the impact of the HHS mandate, particularly on religious groups, said Diana Verm, Legal Counsel at the Becket Fund, in. statement. It is especially significant that the 16 state governments are supporting HBU and ETBU at the Supreme Court. The case directly challenges the. th Circuit Court. That ruling said that the schools were forced to offer all 14 types of contraception spelled out in the HHS mandate of ObamaCare within their health insurance plans. The schools only offered 10 types. They say that the mandate violates their religious freedom. According to the statement, all three schools would have to pay millions in IRS fines if they aren. allowed exemption. The Becket Fund identified the 16 states to FoxNews.com as: Alabama, Arizona, Florida, Georgia, Kansas, Louisiana, Michigan, Montana, Nevada, Ohio, Oklahoma, South Carolina, South Dakota, Texas, Utah, and West Virginia. Other organizations that pledged support include the Ethics and Religious Liberty Commission of the Southern Baptist Convention, the International Mission Board of the Southern Baptist Convention, the Christian and Missionary Alliance Foundation, and all 181 members of the Council of Christian Colleges and Universities. Today. strong support is an indication that the Court is likely to decide in the upcoming term whether religious ministries, like religious for-profits, will receive protection from the Mandate, the statement said. Verm told FoxNews.com that many businesses have been exempted from the mandate, and that all religious institutions should be afforded the same opportunity. ""The Supreme Court has already issued five preliminary orders in favor of religious organizations facing this choice, and we expect it to protect HBU and ETBU as well,"" she said. FoxNews.com's Matt Fossen contributed to this report."
http://1410wizm.com/index.php/item/26046-walker-obamacare  column=clean_web:title, timestamp=1460230892402, value=Christian institutions garnering support in ObamaCare challenge
```

Future work

- Clean big collection
- Clean documents with multiple languages
- Automating webpage crawling and cleanup

Acknowledgements

- Integrated Digital Event Archiving and Library (IDEAL) NSF IIS – 1319578
- Digital Library Research Laboratory (DLRL)
- Dr. Fox, IDEAL GRA's (Sunshin & Mohamed)
- All of the teams in the class

Thank You!