

# 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

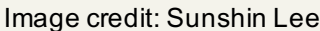
---

- 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



\* Spark or Pig

---

# 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
```

# 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/qiA1hAiTl1 #NAACPBombing http://t.co/40xJDhAENa
```

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/40xJDhAENa  
8 row(s) in 0.1300 seconds
```

Cleaned tweet and extracted info in HBase

---

# Webpage Cleaning

[illegible]

## Banner(Noise)

BLOG VIDEO RESEARCH ISSUES MYTHOEDIA TAKE ACTION

[DONATE](#)

---

**NRA News Tries To Shut Down The Debate: Calls For New Gun Laws Show "A Lack of Shared Humanity"**

Blog » Aug 27, 2015 4:30 PM EDT » TIMOTHY JOHNSON

[f Like](#) {5.3K} 
 [t Tweet](#)
[G+1](#) {1} 
 [Print](#)
[1030](#)

REPLAY

**Picture(Noise)**

Alison Parker      Adam Ward

**BREAKING NEWS**

**WDBJ: REPORTER, PHOTOGRAPHER SHOT TO DEATH ON AIR**

CNN 900 AM ET NEWSROOM

The host of the National Rifle Association's radio show reacted to the fatal shooting of two journalists in Virginia by attacking "anti-gun politicians" and "anti-gun activists" for using the tragedy to call for stronger gun laws, claiming they "politicized" it and demonstrated "a lack of shared humanity."

But not only is the NRA hypocritical for saying gun policy debates should be off-limits after a shooting -- it has used mass shootings to **call for looser gun laws** -- it's also self-serving, because its political agenda benefits when potential new laws that it opposes are not debated and discussed.

The NRA's declaration that this is not the time to discuss gun policy also stands in stark contrast to comments made just hours after the shooting by the father of one of the victims, who said publicly that he will make it his life's work to convince politicians to close loopholes in gun laws.

**Actual Text to be Processed (Sound)**

During the morning of August 26, reporters Alison Parker and cameraman Adam Ward, of Norfolk, Virginia's ABC affiliate station WDBJ, were gunned down while doing a live report from a recreation area. The shooter, who later that day committed suicide, was a local road construction worker. The tragedy quickly made national headlines and prompted calls for stronger gun laws and action by President Obama, Hillary Clinton, and Virginia Gov. Terry McAuliffe (D).

Later that same day during an afternoon broadcast, Cam Edwards, host of the NRA radio show, *Cam & Company*, lashed out at people who consider this latest incident of shocking public gun violence as more evidence the nation needs stronger gun laws.

Edwards complained, "Before we know any of the details, we are seeing anti-gun politicians, anti-gun activists trying to turn this tragedy into some sort of political advantage," and went on to characterize calls for new gun laws as "the wrong response to take here. I think it shows a lack of shared humanity."

search | 🔍 🌐

---

**ABOUT THE BLOG**

Our blog section features rapid response fact-checks of conservative misinformation, links to media criticism from around the web, commentary, analysis and breaking news from Media Matters' senior fellows, investigative team, researchers and other staff.

---

**FOLLOW US** »

[t Follow @mmfa](#) {208K followers}
 [f Like](#) {504K} Like on Facebook
 [G+1](#) {41k} Recommend on Google
 [tumblr](#) Follow on Tumblr
 [y YouTube](#) Subscribe
 [Pinterest](#) New Gopher Boards
 [Email](#) RSS Feeds and mailing lists

---

**TIMOTHY JOHNSON** »

Timothy Johnson is the guns and public safety program director at Media Matters, having previously spent time at the Brady Center to Prevent Gun Violence Legal Action Project and the Coalition to Stop Gun Violence. He is a graduate of The George Washington University.

[All posts](#) » 
 [Twitter](#) »

---

**LATEST** »

**Fox Contributor: Rubio Is "Embarrassing Himself" The Longer He Stays in Presidential Race**

43 minutes ago [Video](#)

# Webpage Cleaning Rules

---

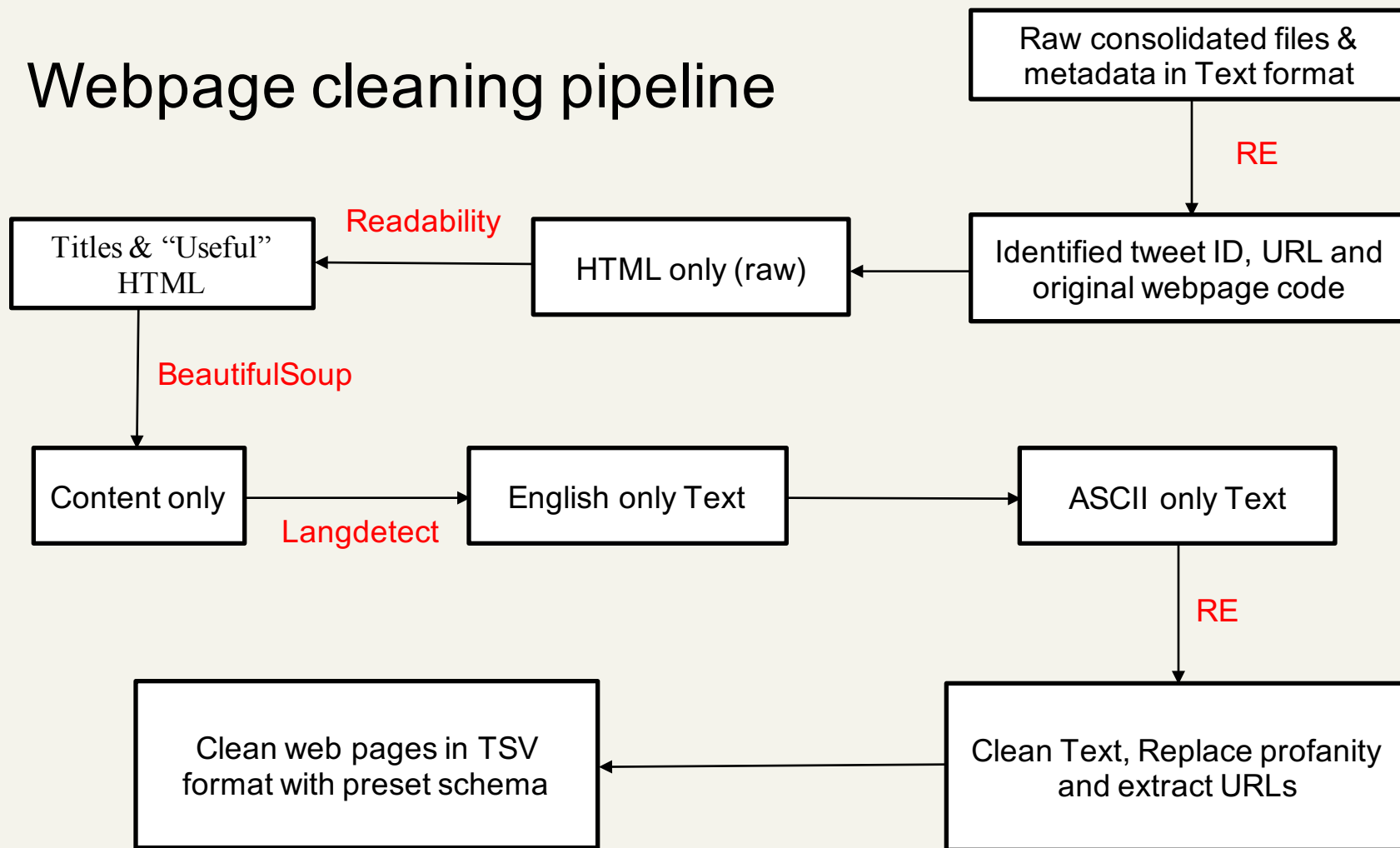
- Remove Non-ASCII characters
- Keep English text only
- Extract URLs
- Remove profane words

# 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 a 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 a 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 a 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 a statement. It is especially significant that the 16 state governments are supporting HBU and ETBU at the Supreme Court. The case directly challenges the 11th 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't 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!