

Collection Management Webpages

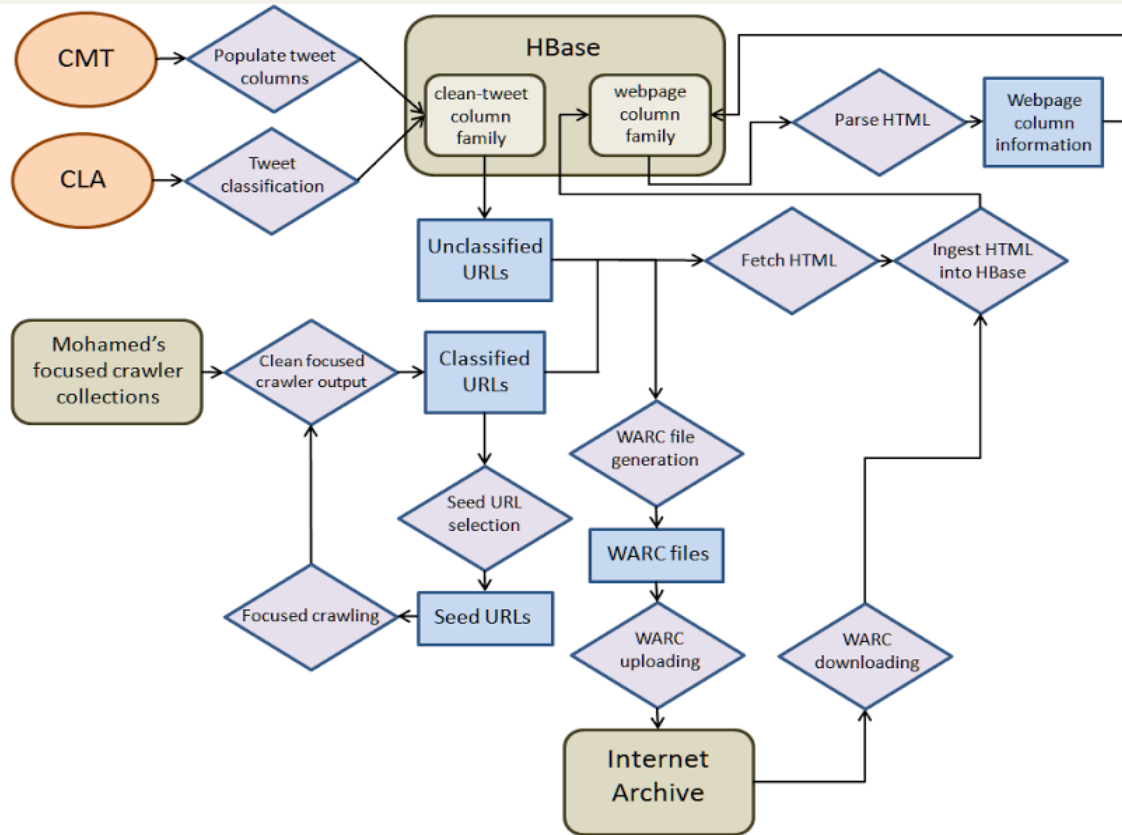
Final Presentation

Tung Dao
Weigang Liu
Christopher Wakeley

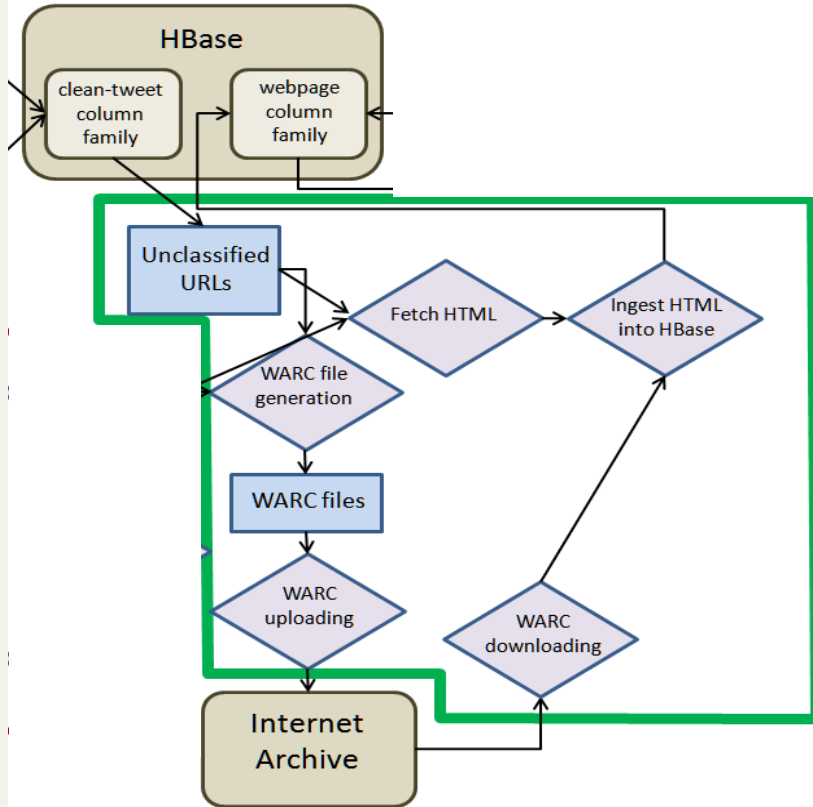
CS5604 – Information Storage and Retrieval
Fall 2016
Virginia Polytechnic Institute and State University
Blacksburg, VA
Professor Edward FoxA

December 1, 2016

System Overview



HTML Fetching and WARC Files



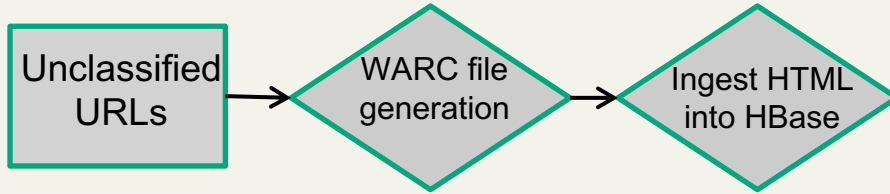
- **Fetch HTML**

- **Generate WARC files**

- **Ingest WARC files from IA**

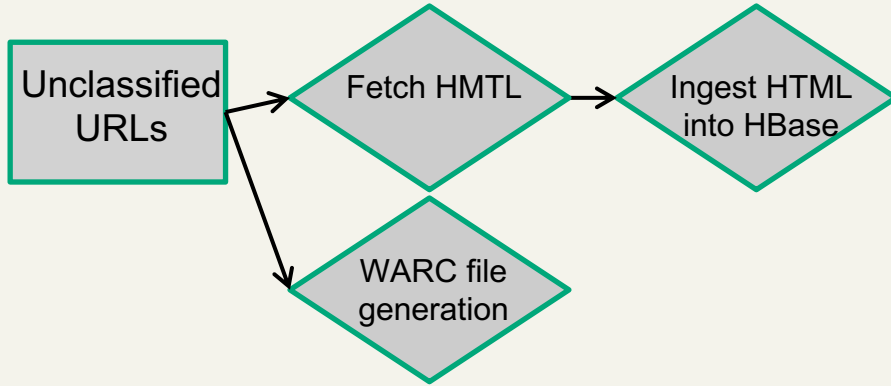
Fetching HTML

Original Pipeline



- Only hit server once
- Performance
- Politeness

Revised Pipeline



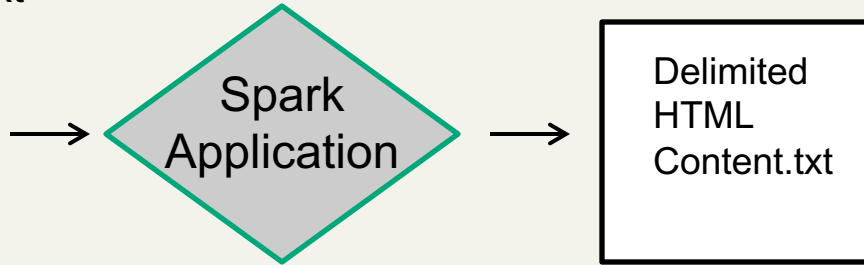
Problem:

- Minutes to generate WARC

Fetching HTML Implementation

Line Delimited URLs.txt

```
1 http://abcnews.go.com/International/charlie
id=28105639
2 http://america.aljazeera.com/blogs/scrutine
utm_content=main&utm_campaign=ajam&utm_sour
3 http://blogs.mediapart.fr/blog/olivier-tonn
4 http://blogs.wsj.com/chinarealtime/2015/01/
in-beijing/
5 http://imgur.com/a/zd5rL/
6 http://indiatoday.intoday.in/story/charlie-
france-paris-killing/1/412358.html
7 http://linkis.com/org/RBHKy
8 http://mashable.com/2015/01/12/simpsons-cha
9 http://news.blogs.nytimes.com/2015/01/08/up
10 http://nymag.com/daily/intelligencer/2015/0
victims.html?mid=twitter_nymag
11 http://stream.wsj.com/story/charlie-hebdo-o
12 http://time.com/3657919/paris-vigil-charlie
13 http://www.canewse.com/2015/01/charlie-hebd
```

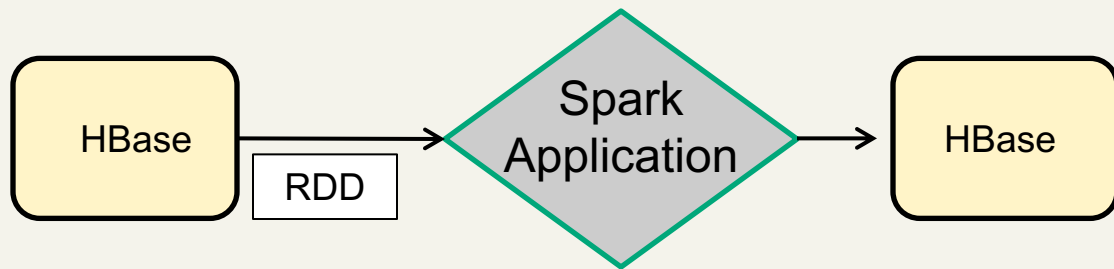


Performance (local mode):

Measure speedup in future

URLs	Runtime (s)
64	23.031
128	10.752
256	16.876
512	38.756

Fetching HTML Future Work



■ Avoid Coalesce

- Don't store all results on one partition
- Scalability

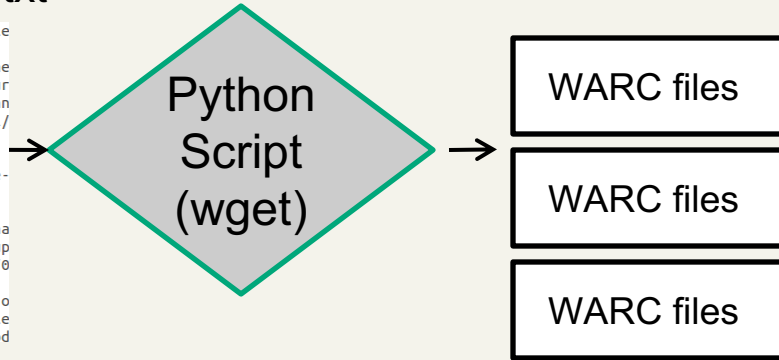
■ Incremental Update

- Add “fetched” column
- Compare with timestamp (Freshness)

WARC Generation

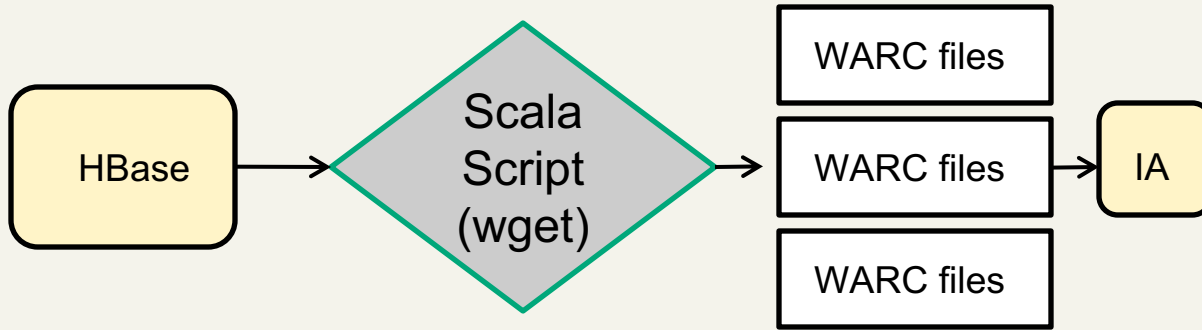
Line Delimited URLs.txt

```
1 http://abcnews.go.com/International/charlie
id=28105639
2 http://america.aljazeera.com/blogs/scrutine
utm_content=main&utm_campaign=ajam&utm_sour
3 http://blogs.mediapart.fr/blog/olivier-tonn
4 http://blogs.wsj.com/chinarealtime/2015/01/
in-beijing/
5 http://imgur.com/a/zd5r1/
6 http://indiatoday.intoday.in/story/charlie-
france-paris-killing/1/412358.html
7 http://linkis.com/org/RBHKy
8 http://mashable.com/2015/01/12/simpsons-cha
9 http://news.blogs.nytimes.com/2015/01/08/up
10 http://nymag.com/daily/intelligencer/2015/0
victims.html?mid=twitter_nymag
11 http://stream.wsj.com/story/charlie-hebdo-o
12 http://time.com/3657919/paris-vigil-charlie
13 http://www.canewse.com/2015/01/charlie-hebdo
```



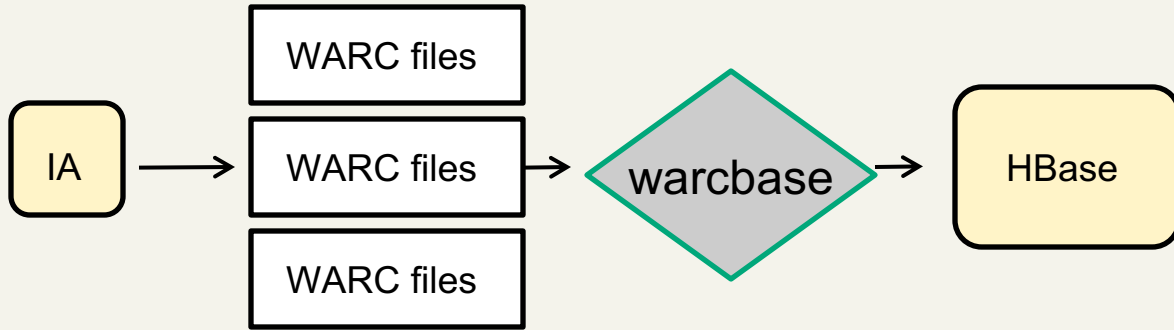
- Existing Tools
 - NOT distributed
 - All implement crawling functionality
 - We already have a crawler (Focused Crawler)

WARC Generation Future Work



- Read from HBase
- Upload to IA

WARC Ingestion (All Future Work)

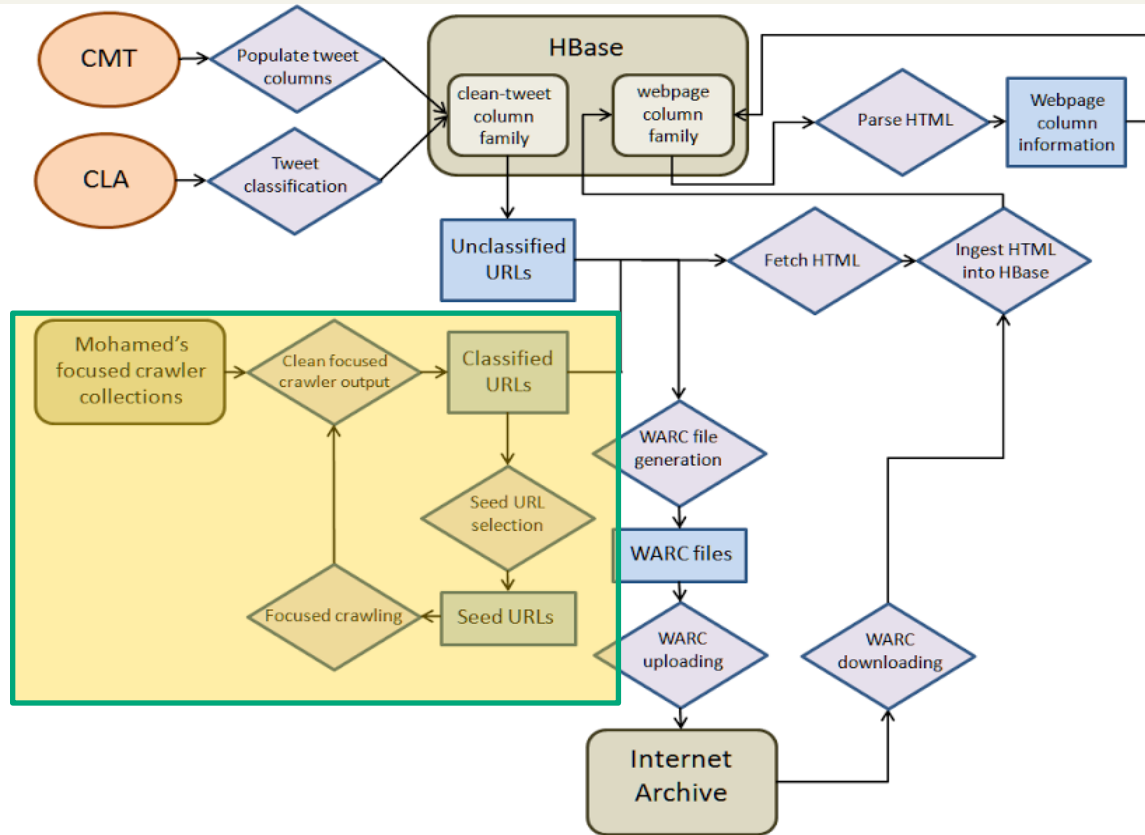


- **Modify HBase insertion**
 - Input Schema
- **Implement IA downloads**

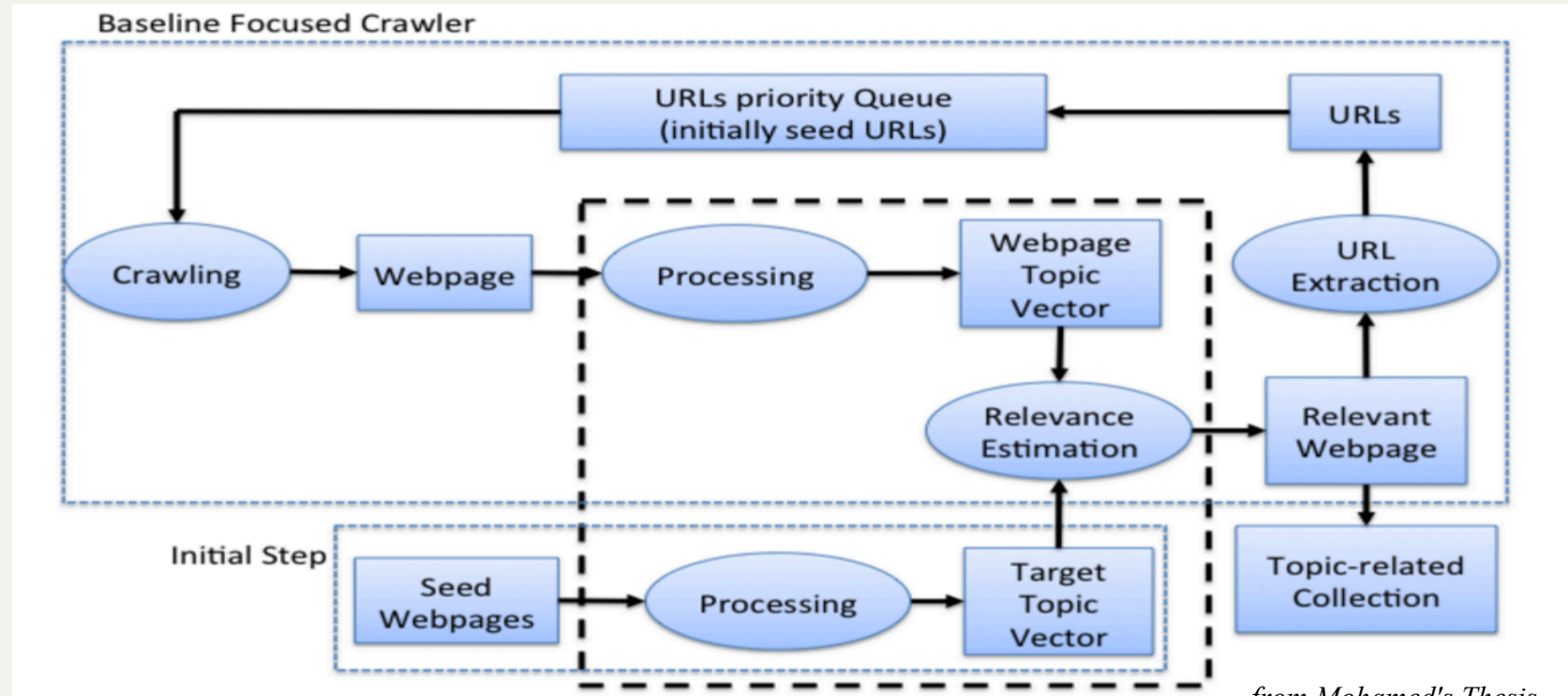
Focused Crawler

- Focused Crawler:
 - Introduction
 - Role in CMW
 - Outline
 - Implementation
 - Original Design
 - Extensions
 - Experiments & Results
 - Effectiveness: Relevance and Correctness
 - Efficiency: Running Time & Space (Memory)
 - Future Ideas

Focused Crawler: Role in CMW

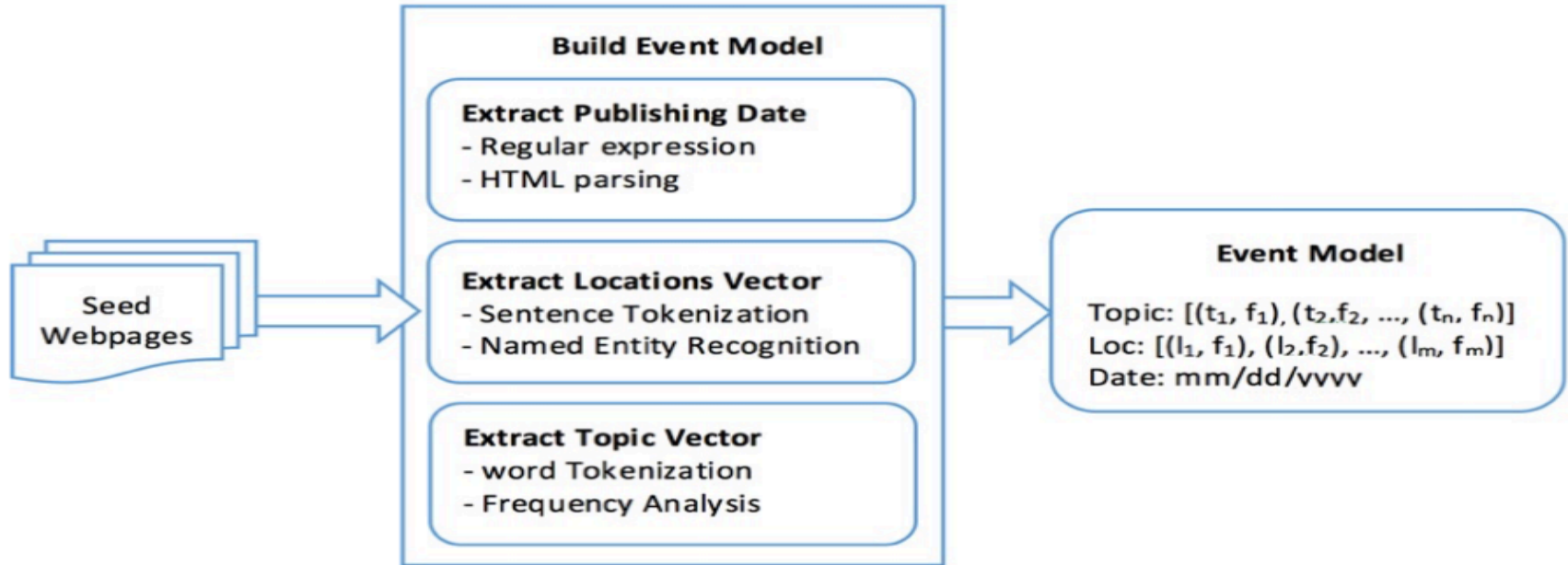


Focused Crawler: Architecture



from Mohamed's Thesis

Focused Crawler: Event Model



Event Focused Crawler: Implementation

- Three main components:
 - **Crawler** → Baseline Focused Crawler (Topic only) → Event Focused Crawler (Topic, Location, Date)
 - **Feature Extractor:**
 - Topic
 - Location
 - Date
 - Using Stanford NER
 - **Event Model**
 - Represent an event
 - Calculate similarity/relevance score (webpage and event)
 - Using TFIDF/Cosine model
- Implemented in Python (~ 1K LOC)

Event Focused Crawler: Extensions (1/3)

■ Output Format

- “Column-based” format (title, URL, topic, locations, dates)– like JSON, instead of “flatted text”.
- Standardized WARC file, instead of text file (using WARC Python APIs).

■ Accuracy

- Distinguish (dates & locations) in the title and (dates & locations) in the content.
 - Using BeautifulSoup & Stanford NER, respectively
 - Weighting them differently (more for the first one)

Event Focused Crawler: Extensions (2/3)

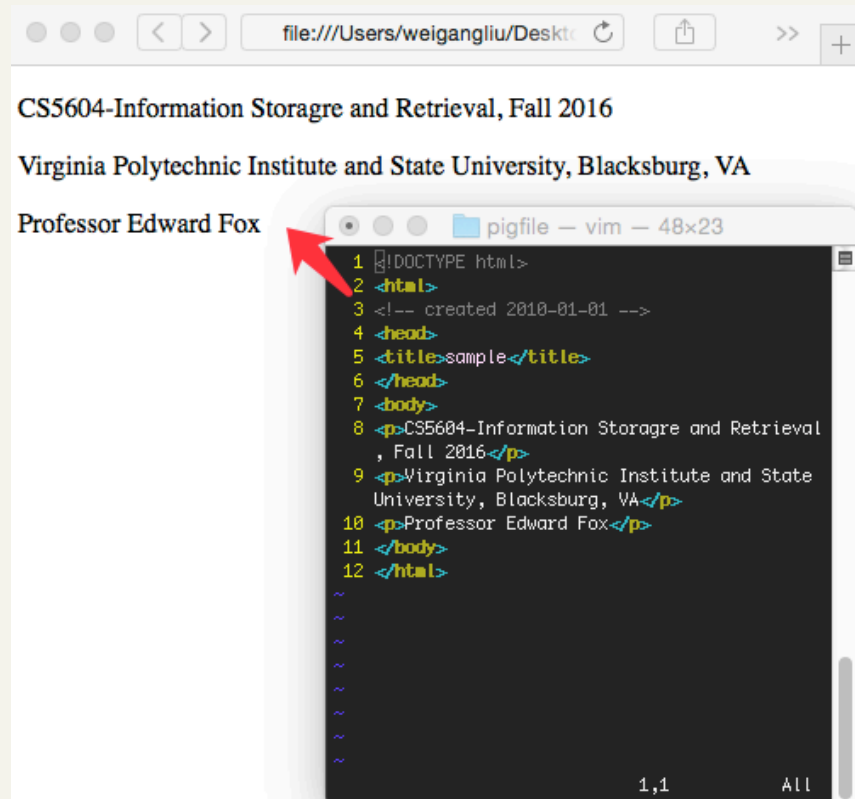
- Evaluation & Comparison
 - Evaluation
 - Crawl three events:
 - “South China Sea Dispute”
 - “USA President Election 2016”
 - “South Korean President Protest”
 - Numbers of seeds: 25
 - PageThreshold: 0.5
 - Top-K: 10
 - Page Limits: 100; 10,000; 100,000 (couldn't terminate in a time manner)
 - Comparison
 - Event Focused Crawler vs. Heritrix (not yet completed)

Event Focused Crawler: Extensions (3/3)

- Scale up
- Apply NLP to increase accuracy
 - Synonyms
 - Part-of-Speech taggers
 - Sentiment Analysis
- Multiple related-events focused crawler
 - Focus only on intersection of multiple events
 - Parameterize events' importance

HTML

- Ignore what it does
- Don't display the tags
- Interpret the content!



The image shows a web browser window displaying the rendered HTML content. The browser's address bar shows the file path: `file:///Users/weigangliu/Desktop/...`. The rendered content includes the following text:

CS5604-Information Storage and Retrieval, Fall 2016
Virginia Polytechnic Institute and State University, Blacksburg, VA
Professor Edward Fox

Overlaid on the browser is a vim editor window titled `pigfile -- vim -- 48x23`. A red arrow points from the text "Professor Edward Fox" in the browser to the corresponding line in the vim editor. The vim editor shows the following HTML code:

```
1 |!DOCTYPE html>
2 |<html>
3 |<!-- created 2010-01-01 -->
4 |<head>
5 |<title>sample</title>
6 |</head>
7 |<body>
8 |<p>CS5604-Information Storage and Retrieval
  |, Fall 2016</p>
9 |<p>Virginia Polytechnic Institute and State
  |University, Blacksburg, VA</p>
10|<p>Professor Edward Fox</p>
11|</body>
12|</html>
```

The vim editor also shows a status bar at the bottom right with the text `1,1 All`.

BeautifulSoup

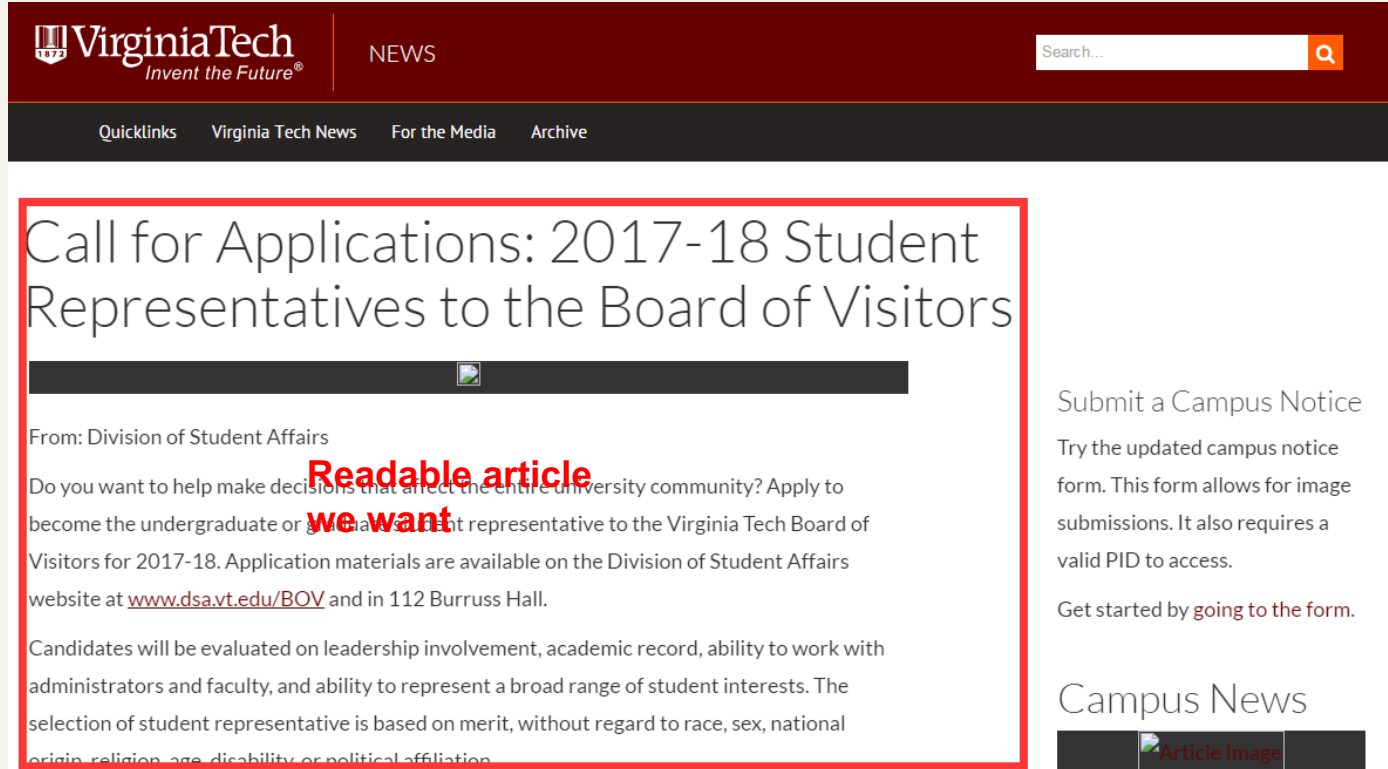
- An HTML or XML parser
- Pythonic idioms for iterating, searching, and modifying the parse tree
- Automatic conversion

Readability

- Measure the readability of text
- Estimate the grade level of word density
- Can be used for Noise Reduction
- Only works for English

\$ pip install <https://github.com/andreascv/readability/tarball/master>

Readability



The screenshot shows the top navigation bar of the Virginia Tech website. On the left is the Virginia Tech logo with the tagline "Invent the Future". In the center is the word "NEWS". On the right is a search bar with a magnifying glass icon. Below the navigation bar is a dark grey menu with links for "Quicklinks", "Virginia Tech News", "For the Media", and "Archive". The main content area features a large red-bordered box around the article title and introductory text. The article title is "Call for Applications: 2017-18 Student Representatives to the Board of Visitors". Below the title is a small image placeholder. The text below the image reads: "From: Division of Student Affairs", "Do you want to help make decisions that affect the entire university community? Apply to become the undergraduate or graduate student representative to the Virginia Tech Board of Visitors for 2017-18. Application materials are available on the Division of Student Affairs website at www.dsa.vt.edu/BOV and in 112 Burruss Hall.", and "Candidates will be evaluated on leadership involvement, academic record, ability to work with administrators and faculty, and ability to represent a broad range of student interests. The selection of student representative is based on merit, without regard to race, sex, national origin, religion, age, disability, or political affiliation". To the right of the red box is a sidebar with the text "Submit a Campus Notice", "Try the updated campus notice form. This form allows for image submissions. It also requires a valid PID to access.", and "Get started by going to the form.". At the bottom of the sidebar is the heading "Campus News" and a small image placeholder labeled "Article Image".

VirginiaTech
Invent the Future®

NEWS

Search..

Quicklinks Virginia Tech News For the Media Archive

Call for Applications: 2017-18 Student Representatives to the Board of Visitors

From: Division of Student Affairs

Do you want to help make decisions that affect the entire university community? Apply to become the undergraduate or graduate student representative to the Virginia Tech Board of Visitors for 2017-18. Application materials are available on the Division of Student Affairs website at www.dsa.vt.edu/BOV and in 112 Burruss Hall.

Candidates will be evaluated on leadership involvement, academic record, ability to work with administrators and faculty, and ability to represent a broad range of student interests. The selection of student representative is based on merit, without regard to race, sex, national origin, religion, age, disability, or political affiliation.

Submit a Campus Notice

Try the updated campus notice form. This form allows for image submissions. It also requires a valid PID to access.

Get started by going to the form.

Campus News

Article Image

Python Script Results

- Mainly utilize above two packages
- Test results on the static webpage collection of Charlie Hebdo shooting

```
Weigangs-iMac:pigfile weigangliu$ python webclean3.py charlie.txt charlie_web webpage.avsc charlie 001 1
webclean3.py
('charlie.txt', 'charlie_web', 'webpage.avsc')
charlie.txt has been cleaned up
Total webpages: 501
Cleaned webpages: 468
Percentage cleaned: 93.413
Language Statistics: {'fr': 4, 'en': 468}
--- 1608.72004104 seconds ---
```

Further Steps and Improvements

- Final step
 - Load the data into HBase for SOLR, FE
- Future improvement
 - Using AVRO file as the output to avoid text file concatenation
 - Hadoop streaming (parallelization)

Project Summary

- Many working individual components
- Lots of work left to connect everything together
- HBase connection needs implementation in most components

Acknowledgments

- NSF grant IIS-1319578, III: Small: Integrated Digital Event Archiving and Library (IDEAL).
- NSF IIS-1619028: Global Event and Trend Archive Research (GETAR)
- Dr. Fox
- GRAs:
Mohamed Magdy Farag
Sunshin Lee
- Other current/past teams.