600.466.FINAL PROJECT

Member:
Xiaoxiao Liu xliu91@jhu.edu
Da Chen dachen78@jhu.edu

Domain:
Our project focuses on building an intelligent web news crawler whose goal is to find the latest, hottest technology news. This crawler can crawl through the main news authorities (such as New York Times, CNN, BBC) and group similar news together base on their similarity.

To view our results:
We have submitted our hard code on websubmit.pl.
In case of updates on the submission, we suggest view our code on:
Github link: https://github.com/Firmamenter/News-Cluster

- **Project Summary:**

  In modern days, with the explosion of information, it is extremely time consuming to read all news from different news websites. Especially many of them report news with different titles but almost the same contents. To save people's time and energy, we may try to collect news from different websites and compare their similarities so as to generate a report for certain date's news with less redundancy and provide links in case people want to read deeply.

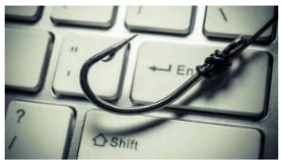
- **Sample Result:**

  Attached to this README is our sample pdf (whose screen shot is as follow). This output file contains the crawled (latest) news title, link, picture and clustered similar news.

Tech News Cluster

Cyber-attack glossary: What are malware, patches and worms? - BBC News

Read More. **Related News 1.** **Related News 2.**



DocuSign users sent phishing emails after data breach - BBC News

Read More. **Related News 1.** **Related News 2.**

Google wants to help you to get a job - May. 17, 2017

Read More. **Related News 1.** **Related News 2.**

How to Protect Yourself From Ransomware Attacks

Read More. **Related News 1.**

- **Run our code**

  We have made a bash script that have connected all the scripts in our project in order. Just run this script by `$ bash run.sh`. Easy and efficient!

```
#!/bin/bash

cd ./final/spiders/
rm -f news.json

python3 auto.py

cp news.json ../../PDF_Generator/

cd ../../crawler2/

rm -f crawler.db

python run.py

cp res.json ../PDF_Generator/

cd ../token/

python3 token.py
python3 tokenize.py articles.raw
perl make_hist.prl < articles.tokenized > articles.tokenized.hist

cd ../similarity/

python3 compute.py

cd ../from_list_to_dict/

python3 toDict.py

cd ../PDF_Generator/

python3 entry.py
```

- **Areas of specialization (success)**

  1. Filtering
  In our crawler we have made the following rules to make sure the news fetched back is the most accurate, latest, hottest and related news in the technology field from different major news authorities on the internet.
     1) only select news posted within 7 days from today
     2) only within certain depth of crawling
     3) only within a certain news authority domain
     4) only related to Technology topics

2. Cached data
We used database (sqlite3) to store crawled urls and data to avoid having to requesting the same web page again and again (performing politeness). Each time a new url is to be crawled, it is first checked in the cache database to determine new or old.

3. DATA VISUALIZATION
We render out result in a well-formatted pdf file with clickable links, images and text.

- **Complexity**

1. Our crawler can automatically extract and uniform certain data from complex html structures from different websites.
2. As the second major part of our project, our program can calculate the similarity between different news retrieved and cluster together similar ones. We believe this feature of our program will benefit user by saving their time on browsing news in real life.