INF 558 Assignment 1

Student: Cheng-Lin Li USC ID: 9799716705

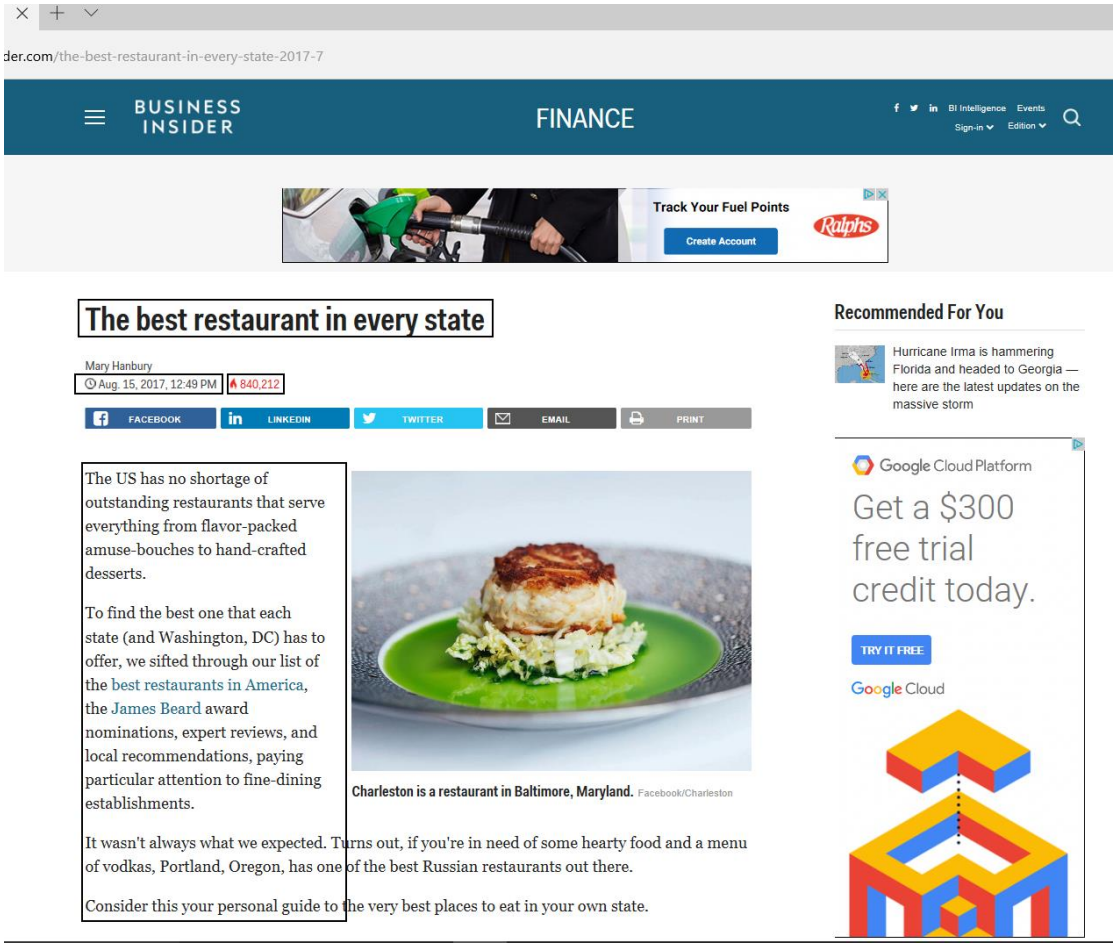
Task1:

- Website URL: https://www.businessinsider.com
- Description: the website provides trending, tech insider, finance, politics, strategy, life, sports, and video reports for readers.

Task2:

- Sample webpage: Figure 1
- Interesting information:
  - o Article Title
  - o Publish date, time
  - o Engagement / View numbers.
  - o Content

Figure 1:



Task3:

- Describe the crawler you used to get the results.
    - ACHE Focused Crawler is selected for this assignment.
- What is the seed URL(s)?
    - Seed URLs list below:

1. http://www.businessinsider.com/the-best-restaurant-in-every-state-2017-7
2. http://www.businessinsider.com/100-trips-everyone-should-take-in-their-lifetime-according-to-the-worlds-top-travel-experts-2017-8
3. http://www.businessinsider.com/london-underground-better-than-nyc-subway-2017-8
4. http://www.businessinsider.com/what-its-like-to-fly-virgin-airlines-2017-8/
5. http://www.businessinsider.com/secret-billionaire-menu-delmonicos-wall-street-steakhouse-2017-8
6. http://www.businessinsider.com/how-to-build-better-habits-2017-9
7. http://www.businessinsider.com/num-pang-fast-casual-sandwich-shop-review-2017-9
8. http://www.businessinsider.com/budget-airline-comparison-ryanair-vs-easyjet-2017-9
9. http://www.businessinsider.com/10-hotel-restaurants-with-spectacular-views-2012-8
10. http://www.businessinsider.com/the-twenty-best-bbq-joints-in-america-2012-8
11. http://www.businessinsider.com/how-to-hack-fast-food-menus-2013-7
12. http://www.businessinsider.com/science-backed-things-that-make-you-happier-2013-8
13. http://www.businessinsider.com/the-link-between-narcissicsm-and-wealth-2013-8
14. http://www.businessinsider.com/i-lived-in-a-tiny-house-2014-8
15. http://www.businessinsider.com/what-is-burning-man-like-2014-9
16. http://www.businessinsider.com/st-regis-monarch-beach-for-sale-aig-photos-2011-9
17. http://www.businessinsider.com/richest-people-in-history-2010-8
18. http://www.businessinsider.com/city-with-the-worst-drivers-in-america-2011-9
19. http://www.businessinsider.com/shake-shack-versus-in-n-out-faceoff-2015-9
20. http://www.businessinsider.com/6-books-every-new-new-yorker-should-read-2015-9
21. http://www.businessinsider.com/no-beer-sales-at-next-monday-night-football-game-2010-12
22. http://www.businessinsider.com/a-vc-tells-the-horrible-sexual-harassment-shes-experienced-2017-8
23. http://www.businessinsider.com/early-retirement-increase-income-decrease-spending-2017-9
24. http://www.businessinsider.com/vix-creator-bob-whaley-interview-investors-dont-understand-wall-street-fear-gauge-2017-9
25. http://www.businessinsider.com/stock-market-news-goldman-sachs-2-reasons-safe-from-correction-2017-9
26. http://www.businessinsider.com/hurricane-irma-stock-market-2017-9
27. http://www.businessinsider.com/art-cashin-post-911-commentary-2017-9
28. http://markets.businessinsider.com/news/stocks/apple-stock-price-iphone-x-rising-ahead-of-launch-event-2017-9-1002358671

29. http://www.businessinsider.com/tesla-urban-superchargers-boston-chicago-2017-9
30. http://www.businessinsider.com/r-citigroup-sees-third-quarter-markets-revenue-down-15-percent-versus-year-earlier-2017-9
31. http://markets.businessinsider.com/news/stocks/starbucks-stock-price-rewards-program-coffee-debit-card2017-8-1002358342
32. http://www.businessinsider.com/walmart-will-use-nvidia-and-ai-to-gain-ground-on-amazon-2017-9
33. http://www.businessinsider.com/secretary-of-defense-mattis-warns-north-korea-2017-8
34. http://www.businessinsider.com/sputnik-fbi-russia-investigation-2017-9
35. http://www.businessinsider.com/lush-products-you-should-get-before-discontinued-2017-8
36. http://www.businessinsider.com/facebook-new-ad-format-replace-print-catalogs-2017-9
37. http://www.businessinsider.com/internet-of-things-research-iot-connected-cars-connected-homes-enterprise-government-2016-11
38. http://www.businessinsider.com/ai-ecommerce-report-2017-8
39. http://www.businessinsider.com/nikes-new-tech-creates-custom-sneakers-in-under-2-hours-2017-9
40. http://www.businessinsider.com/the-impact-of-the-feds-tapering-on-mortgages-2017-6
41. http://www.businessinsider.com/getting-help-with-investing-like-learning-to-ride-a-bike-2017-6
42. http://www.businessinsider.com/alan-greenspan-is-wrong-about-a-bond-bubble-2017-8
43. http://www.businessinsider.com/its-time-for-the-fed-to-run-the-economy-hot-2017-8
44. http://www.businessinsider.com/the-benefits-of-living-out-of-a-backpack-2015-11
45. http://www.businessinsider.com/lessons-i-learned-from-while-traveling-the-world-2015-6
46. http://www.businessinsider.com/tiny-home-pay-gas-not-rent-2016-1
47. http://www.businessinsider.com/asia-and-europe-backpacking-photos-2015-12
48. http://www.businessinsider.com/the-brod-trip-2016-2

- How did you manage to only collect the webpages respecting the template(s) in Task 2?
    - Regular expression is adopted to get the content templates.
        - Pageclassifier.yml

```
type: regex
parameters:
    boolean_operator: "AND"
    url:
      boolean_operator: "AND"
      regexes:
        - https?://www.businessinsider.com/.*.+201(7|6|5|4|3|2|1|0)-[1-9]{1,2}$
    content:
      boolean_operator: "AND"
      regexes:
```

- .*\b(ks-header-facebook)\b.*

  - White list to get the content page.
    - Link_whitelist.txt: To avoid the content with multiple pages layout to be included into results. Only requests single page format of content.

      https?:\/\/www\.businessinsider\.com\/.*.+201(7|6|5|4|3|2|1|0)-[1-9]{1,2}$

- How did you discard irrelevant pages?
  - Leverage ACHE configuration to discard irrelevant pages.
    - Ache.yml

      target_storage.store_negative_pages: false
  - Adopt Link filters to discard irrelevant pages.
    - Link_blacklist.txt: To avoid some RSS files include into results.

      https?:\/\/www\.businessinsider\.com\/.*.rss

Task4:

- What is the format of time stamp?
  - DAY Mmm DD HH:Mi:SS YYYY
    "Mon Sep 11 09:29:20 2017"
- Usage of jsonlines package (Python version 3.6)
  - Packing html files into single CDR in JSON lines format.

Usage: python C:\JSONLines\jsonlines.py <d> <html files input path> [output_path] [number of files]

The program requires Python 3.6 to execute.

- d : Only draw the html file size distribution.

- html files input path = input file path of html files

- output_path = Optional, output file path of JSON Lines format

- number of files : Optional, number of html files you want to package into CDR. Default value is all files.

- HTML files size distribution: 2000 files.