

INF 558 Information Graph – Assignment 2

Student: Cheng-Lin Li

USC ID: 9799716705

Task 1:

1. Interesting fields from business insider webpage:

- Answer:

○ Content:

- Category
- Subject
- Content
- Publish Date
- Views

○ Author/Contributor:

- Name

The screenshot shows a web browser window displaying a Business Insider article. The browser's address bar shows the URL: localhost:3333/project/3/cluster/4/page/275/html/cached. The Business Insider logo is visible in the top left, and the 'LIFE' logo is in the top right. A banner at the top of the article area reads: 'There will be free admission. Sunday, September 17' for the 'Pacific Standard Time: LA/LA' art festival. The article title is 'World's Richest Woman Tells Jealous People To Drink Less And Work More' by Joshua Berlinger, dated Aug 30 2012 10:54 AM, with 25,090 views. Below the title are social media sharing buttons for Facebook, LinkedIn, and Twitter. The article text begins with 'Gina Rinehart, the world's richest woman, doesn't have much pity for those who are jealous of her wealth. In a piece written for Australian Resources and Investment Magazine, Rinehart told readers to stop complaining and spend less'. To the right of the text is a photo of Gina Rinehart smiling and being interviewed. Further right is a 'Recommended For You' section featuring a post from a retired millionaire. At the bottom right is a large advertisement for the 'Pacific Standard Time: LA/LA' art festival, including the festival's logo and presenting sponsors: The Getty and Bank of America.

BUSINESS INSIDER **LIFE** f t in BI Intelligence Events

There will be free admission. **Pacific Standard Time: LA/LA**
Latin American & Latino Art in LA
Presenting Sponsors: The Getty, Bank of America. [Learn more](#)

World's Richest Woman Tells Jealous People To Drink Less And Work More

Joshua Berlinger [Email](#) [Twitter](#)
Aug 30 2012 10:54 AM **25,090**

[Facebook](#) [LinkedIn](#) [Twitter](#) [Email](#) [Print](#)

Gina Rinehart, the world's richest woman, doesn't have much pity for those who are jealous of her wealth. In a piece written for Australian Resources and Investment Magazine, Rinehart told readers to stop complaining and spend less

Recommended For You

I retired a millionaire at 43 — here are 6 of my favorite spending and saving hacks that helped me get there

Pacific Standard Time: LA/LA
Latin American & Latino Art in LA
Presenting Sponsors: The Getty, Bank of America

INF 558 Information Graph – Assignment 2

Task 2:

Page	845	965	976	982	1034
page57	Enterprise	50 enterprise startups you've probably never heard ...	Jan. 15, 2016, 2:16 PM	254,364	static3.businessinsider.com/image/56846caae6183e59 ...
page58	Tech Insider	Way To Go, Tim! (Are You Paying Attention, Yahoo?)	Apr. 9, 2012, 9:03 AM	2,137	static2.businessinsider.com/image/4b50e7800000000 ...
page61	Life	World's Richest Woman Tells Jealous People To Drift ...	Aug. 30, 2012, 10:54 AM	25,090	static3.businessinsider.com/image/4be594a0b637795 ...
page59	Your Money	A 27-year-old who saves 65% of his income shares h ...	May 4, 2017, 1:56 PM	80,230	static4.businessinsider.com/image/590b4e57cd8eb1b ...
page56	AFP	Harvey hits Louisiana as Texas rescuers race again ...	Aug. 30, 2017, 11:10 AM	109	static5.businessinsider.com/image/59a6d7a45124c985 ...
page60	Finance	The 6 best books to pick up if you just moved to N ...	Sep. 2, 2015, 2:24 PM	16,157	static1.businessinsider.com/image/55e5d0099dd7c016 ...

1. Can Inferlink tool extract your highlighted fields?

- Answer: It can extract most of fields except “author/contributor”.

2. If not, list up to 3 fields that cannot be extracted and explain why the tool cannot extract these fields. Hint: using lectures about wrapper learning and Inferlink extraction tool.

- Answer: During the template learning phase, Inferlink finds those n-grams that occur on all pages but exactly once. The “author/contributor” attributes are NOT exactly once in all training pages, i.e. Some of pages contain “ks-author-byline” in class of tag but some other pages contain “single-author” in class of tag for author(s) in training pages.

3. Choose one extracted rule and explain how the rules can be used to extract field from webpages.

- Answer:

o Select rule 2:

```
{  
  "begin_regex":  
  "\\>\\s+\\</div\\>\\s+\\</div\\>\\s+\\</div\\>\\s+\\<div\\s+class=\\\"content\\s+post\\.\\?\\\"\\>\\s+\\</div\\s+class=\\\"sl\\-layout\\.\\?\\-post\\\"\\>\\s+\\<\\.\\?h1\\>\\s+\\</h1\\>\\s+\\<div\\s+id=\\\"content\\\"\\s+class=\\\"content\\\"\\>\\s+\\</div\\s+class=\\\"content\\\"\\s+id=\\\"af701a31-8f37-4b4b-8aae-ef7f5a24fc4\\\"\\s+include_end_regex=\\\"true\\\"\\s+name=\\\"965\\\""
```

INF 558 Information Graph – Assignment 2

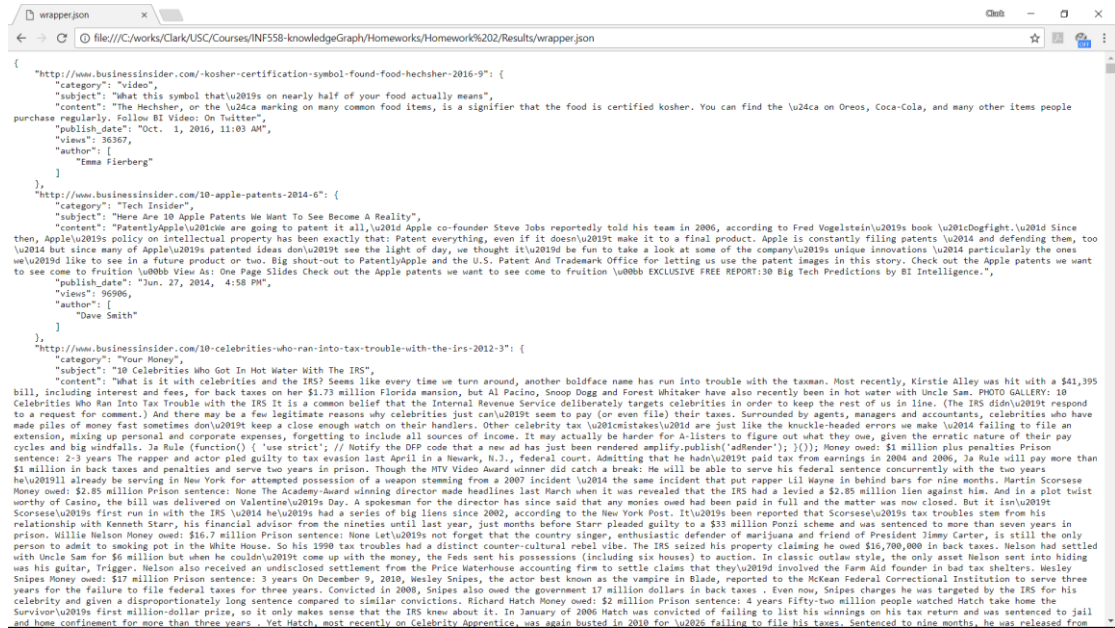
```
"removehtml": false,  
"rule_type": "ItemRule",  
"strip_end_regex":  
"\\</h1\\>\\s+\\<div\\s+id=\"content\"\\s+class=\"content\"\\>\\s+\\<div\\s+class\"  
},
```

○ Explanation:

- String Begin Pattern:
“>” + whitespace (include carriage return) + 3 continuous
 (“</div>” + whitespace) + “<div class=“content post” + zero to
multiple character + “”>” + whitespace + “<div class=“s1-
layout” + zero to multiple character + “-post”>” + “<” + zero
to multiple character + “h1>”
- String End Pattern:
“</h1>” + whitespace (include carriage return) + “<div
id=“content” class=“content”>” + whitespace + “<div class=“
- Strip End Pattern: The same as String End Pattern
“</h1>” + whitespace (include carriage return) + “<div
id=“content” class=“content”>” + whitespace + “<div class=“
- All characters between pattern begin and end will be treated
as article subject.

INF 558 Information Graph – Assignment 2

Task 3: Manual developed wrapper in Python 3.6 with BeautifulSoup 4.6



Usage: python wrapper.py <JSON Lines input file name> </output/path>

The program requires Python 3.6 to execute.

- JSON Lines input file path = input file path of JSON Lines file
- output_path = Output file path of JSON file

Output JSON format:

```
{
  "url": {
    "category": "Category Text",
    "subject": "Subject Text",
    "content": "Article content",
    "publish_date": "Article publish date",
    "views": "How many views on this article",
    "author": ["author1", "author 2"...]
  },
  ...
}
```

How to wrap:

1. Category:
 - a. Locate <h2> and its class contains string "vert-name" then get the text field.
2. Subject: Locate <h1> to get the text field for the subject of the article.
3. Content:

INF 558 Information Graph – Assignment 2

- a. locate <div> and its class contains string "KonaBody post-content" or
 - b. locate <div> and its class contains string "intro-content"then get the text field and remove \n, \t characters.
4. Publish Date: Locate and its class contains string "svg sprites date-icon", then get the text of the field from its next, next sibling tag.
5. Views: Locate and its class contains string "Engagement" then get its text and convert to integer number.
6. Author:
 - a. Locate and its class contains string "ks-author-byline" then get its content, replace 'and' to ',' if any.Or
 - b. Locate and its class contains string "single-author" then get its content.If the string includes ',' ; then split the string by ',' else put author into list.
7. During above process, if the field cannot be located, put null into the data field.
8. All above information was packaged into a dictionary object as a value which will be assigned into an external dictionary data structure with url as key.
9. Dump the external dictionary as JSON object to file.
10. Check the JSON file and revise rules to reduce the null fields until all those null fields are real no contains cases.