#!/usr/bin/python3

# Lines for CSS-based parse for each and all news sources.

# Reference link

# For Google Docs optimization, use [Tools] > [Preferences...] menu, and disable all autocorrect options.

## https://www.nytimes.com/section/sports

## https://www.wsj.com/news/life-arts/sports

## https://www.washingtonpost.com/sports/

## https://nypost.com/sports/

## https://bleacherreport.com/

## https://www.nbcsports.com/

## https://abcnews.go.com/Sports

## https://www.foxnews.com/sports

## https://www.npr.org/sections/sports/

“”

# Usage

### get all items matching tag, in array form.

# response.css('tag here').extract()

### get only first item matching tag.

# response.css('tag here').extract\_first()

# response.css('tag here')[0].extract()

### get all wrapping items, with all sub-items, in array form.

# response.css('tag here').css('tag there').extract()

### get only first wrapping tag item, and all matching sub-items, in array form.

# response.css('tag here')[0].css('tag there').extract()

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **Content** | **Tag** | **Note** |
| NYT | Article | - |  |
| Title | - | Too complicated |
| Author | - |  |
| Date | - |  |
|  |  |  |
| WP | Article |  |  |
| Title |  |  |
| Author |  |  |
| Date |  |  |
|  |  |  |
| WSJ | Article | div.wsj-snippet-body p::text | Does not let access the full text without subscibing |
| Title | div.wsj-article-headline-wrap h1::text |  |
| Author | div.info-name::text |  |
| Date |  |  |
|  |  | Not let us do it |
| **NY Post**  **confirmed.** | Article | div.entry-content p::text  div.entry-content p \*::text | Array of paragraphs  <p> includes <a> tags with texts. |
| Title | h1 a::text |  |
| Author | p.byline a::text |  |
| Date | p.byline-date::text | May contain newline character  Date and time split into two. |
| Link | h5.trending-story\_\_headline a::attr(href) |  |
| CNN Sports/  Bleacher Report | Article | div.zn-body\_\_paragraph::text |  |
| Title | h1.pg-headline::text |  |
| Author |  | Not available |
| Date | p.update-time::text |  |
| Link |  | No link |
| **NBC news**  **All others**  **confirmed?** | Article | div.entry-content p::text  div.entry-content p \*::text  div.entry-content[0] p \*::text | Page contains multiple articles.  Only select first div.entry-content.  Contains "[ MORE: ]" possibly with <b> tag. |
| Title | h1.entry-title::text |  |
| Author | span.author a::text  span.author a::text[0]  span.author[0] a::text | Only select first element. |
| Date | span.entry-date::text  span.entry-date::text[0] | Only select first element. |
| Link | a.story-explorer-item::attr(href) |  |
| **NBC news**  **profootballtalk**  **confirmed.** | Article | div.entry-content p::text  div.entry-content p \*::text | <p> includes <a> tags with texts. |
| Title | h1.entry-title::text | Title contains \xA0 or &nbsp; |
| Author | span.byline::text | String contains "Posted by " |
| Date | span.entry-date::text |  |
| Link |  |  |
| **ABC News**  **confirmed.** | Article | div.article-copy p::text | Array of paragraphs |
| Title | h1::text | Array contains empty elements |
| Author | div.author::text | Array contains empty elements  Trim spaces. |
| Date | span.timestamp::text | Date may come with location. |
| Link | article.news-feed-item h1 a  article.news-feed-item h1 a::attr(href) |  |
| **Fox news**  **confirmed.** | Article | div.article-body p::text |  |
| Title | ~~div.caption p::text~~  h1.headline::text |  |
| Author | div.author-byline span span a::text |  |
| Date | div.article-date time::text |  |
| Link | div.m a::attr(href) |  |
| **NPR news**  **confirmed.** | Article | div.storytext p::text |  |
| Title | div.storytitle h1::text |  |
| Author | div.byline-container--block p a::text  div.byline p \*::text |  |
| Date | div.dateblock span::text |  |
| link |  | No sport links |