Sprint 1 Review

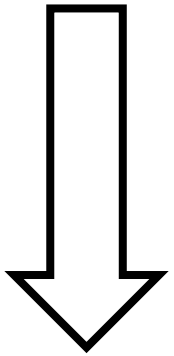
Reading Web News Sources

November 16, 2017

Jon Honda

devleague

Sentiment moves market prices



Identify article sentiment

- Is it good news? Bad news? (bias)
- Are different sources covering different aspects of story?
- Are different sources tending toward different biases?

devleague

Step 1: How do web sources encode their articles?

AP

AP NEWS Log in | Sign up

AP Top News Sports Entertainment Explore

Following Trump visit, China sending envoy to North Korea

16 minutes ago

https://apn

RELATED TOPICS

Beijing

North Korea

North America

China

BEIJING (AP) — Following President Donald Trump's visit to Beijing, China says it's sending a high-level special envoy to North Korea amid an extended chill in relations between the neighbors over Pyongyang's nuclear weapons and missile programs.

The official Xinhua News Agency said Wednesday that director of the ruling Communist Party's International Liaison Department, Song Tao, would travel to Pyongyang on Friday to report on the party's national congress held last month.

Xinhua made no mention of Trump's visit or the North's defense programs, although Trump has repeatedly called on Beijing to do more to use its influence to pressure Pyongyang into altering its behavior.

Song would be the first ministerial-level Chinese official to visit North Korea since October 2015 when Politburo Standing

devleague

Step 1: How do web sources encode their articles?

```
<!DOCTYPE html>
<html lang="en"> ev
 <head>…</head>
 <body class="ng-scope" ng-app="articleContent" style="background-color: white"> ev
   <script type="text/javascript" src="../../dist/assets/js/socialMedia.js"></script>
   <!--Modal-->
  <div id="imageModal" class="modal fade" role="dialog">…</div>
  <div class="header ng-isolate-scope" ng-class="{'mainFeedHeader': hc.isMainFeed()}" page-header="">…</div>
  <div class="articleView">
   <div class="articleContentContainer">
    <div class="sideRail leftRail">…</div>
    <div id="articleContent" class="articleContent">
     <div class="dtTitle">…</div>
     <div class="gradientContainer blueGradient">…</div>
     <script>…</script>
     <article id="contentArea" class=" noPrimaryImage ">
      <div class="tabletTitle">…</div>
      <div class="mobile">…</div>
      <div class="mobile mobileShareTemplate">…</div>
      <div class="articleBody" mark-urls="">
       <p>
          BEIJING (AP) — Following President Donald Trump's visit to Beijing, China says it's sending a high-level special envoy to North Korea amid an extended chill in
          relations between the neighbors over Pyongyang's nuclear weapons and missile programs.
        </p>
       <p>
```

Answer: HTML code (a form of XML)

devleague

Step 1: How do web sources encode their articles?

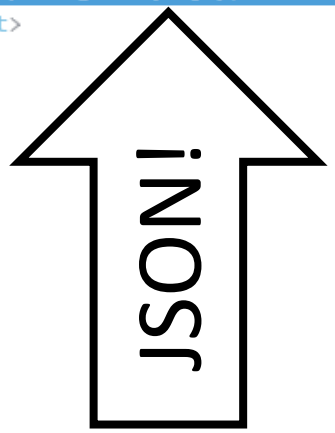
Answer: HTML code (a form of XML)

…but java script made

devleague

Step 2: Find HTML coding patterns

```
<script type="application/ld+json">
(function (i, s, o, g, r, a, m) { i['GoogleAnalyticsObject'] = r; i[r] = i[r] || function () { (i[r].q = i[r].q || []).push(arguments) }, i[r].l = 1 * new Date(); a =
s.createElement(o), m = s.getElementsByTagName(o)[0]; a.async = 1; a.src = g; m.parentNode.insertBefore(a, m) })(window, document, 'script', 'https://www.google-analytics.com
/analytics.js', 'ga'); ga('create', 'UA-19104461-33', 'auto'); ga('set', { 'dimension1': '04d7dcc8c52b474cac498314240f891a', 'dimension4': "Following Trump visit, China
sending envoy to North Korea", 'dimension8': '', 'dimension11' : 'NO' }); ga('send', 'pageview');
{ "@context": "http://schema.org", "@type": "NewsArticle", "mainEntityOfPage": { "@type": "WebPage", "@id": "https://apnews.com/04d7dcc8c52b474cac498314240f891a" },
"headline": "Following Trump visit, China sending envoy to North Korea", "image":{ "@type": "ImageObject", "url": "", "height": "0", "width": "None" }, "datePublished":
"2017-11-15 04:08:26", "dateModified": "2017-11-15 04:08:26", "author": { "@type": "Person", "name": "AP Staff" }, "publisher": { "@type": "Organization", "name": "Associated
Press", "logo": { "@type": "ImageObject", "url": "http://apcontextual.appspot.com/Support/AMPLogo.png", "width": 600, "height": 60 } }, "description": "" }
</script>
```

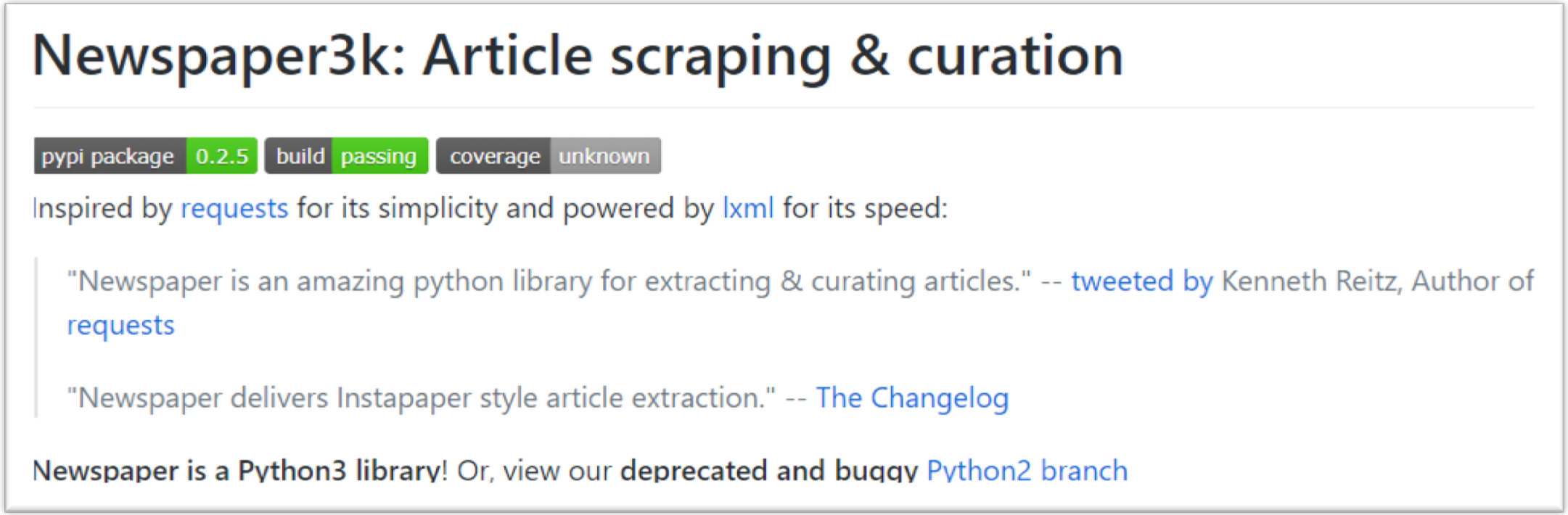


devleague

Step 2: Some JSON News Encodements


```
{ "@context": "http://schema.org", "@type": "NewsArticle", "mainEntityOfPage": { "@type": "WebPage", "@id": "https
"headline": "Following Trump visit, China sending envoy to North Korea", "image":{ "@type": "ImageObject", "url":
"2017-11-15 04:08:26", "dateModified": "2017-11-15 04:08:26", "author": { "@type": "Person", "name": "AP Staff" }
Press", "logo": { "@type": "ImageObject", "url": "http://apcontextual.appspot.com/Support/AMPLogo.png", "width": 6
```


Step 3: Programmatically Extract Data (webscraping)



devleague

Step 3: Get web news into Newspaper Library

Code Description	Python Code
Get Article “class” from Library	>>> from newspaper import Article
Define news article url	>>> url = 'https://apnews.com/04d7dcc8c52b474cac498314240f891a/Following-Trump-visit,-China-sending-envoy-to-North-Korea’
Make instance of Article “class”	>>> myArticle = Article(url)
Download web article	>>> myArticle.download()
Extract useful data	>>> myArticle.parse()

Step 3: Get web news into Newspaper Library

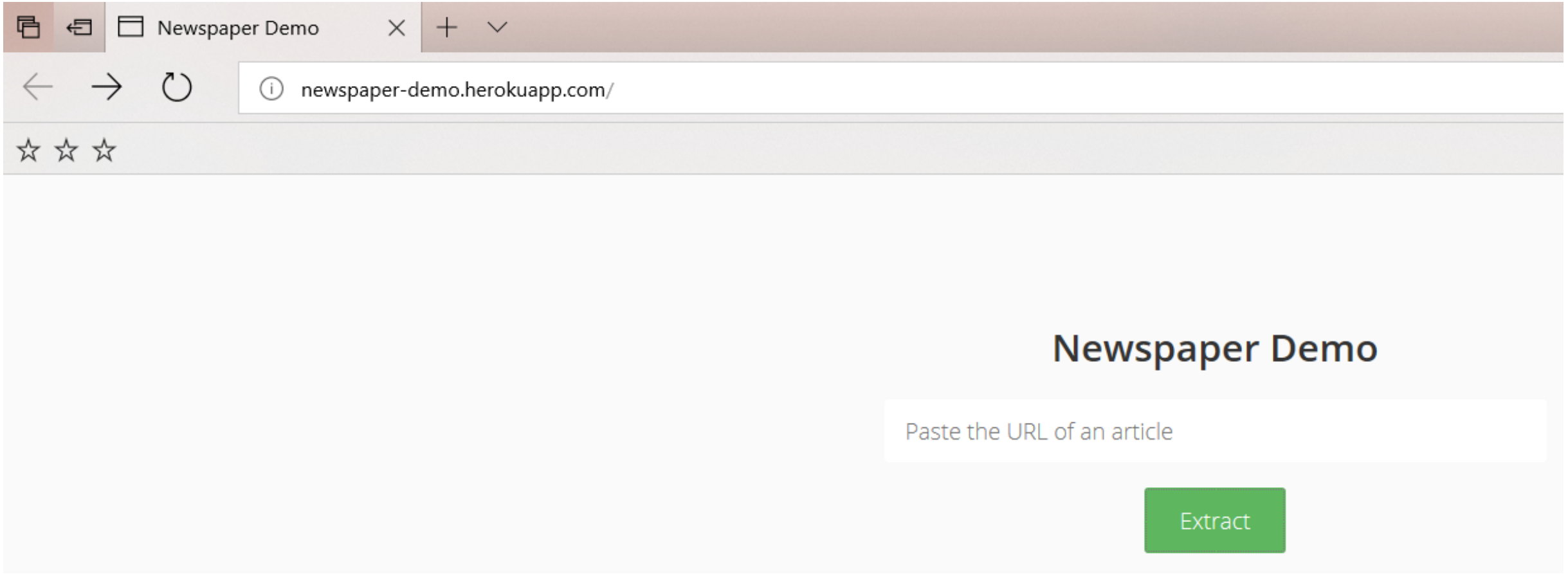
Code Description	Python Code
Get Article “class” from Library	>>> from newspaper import Article
Define news article url	>>> url = 'https://apnews.com/04d7dcc8c52b474cac498314240f891a/Following-Trump-visit,-China-sending-envoy-to-North-Korea’
Make instance of Article “class”	>>> myArticle = Article(url)
Download web article	>>> myArticle.download()
Find data	>>> myArticle.parse()

Step 3: Extract Data w/ Newspaper Library

Code Description	Python Code
Get Authors	`>>> article.authors` `['Leigh Ann Caldwell', 'John Honway']`
Get Publish Date	`>>> article.publish_date` `datetime.datetime(2013, 12, 30, 0, 0)`
Get Article Text	`>>> article.text` `'Washington (CNN) -- Not everyone subscribes to a New Year's resolution...'`
Get 1st Image	`>>> article.top_image` `'http://someCDN.com/blah/blah/blah/file.png'`
Get Movies	`>>> article.movies` `['http://youtube.com/path/to/link.com', ...]`

devleague

Step 3: Newspaper Library Extracted Data:



devleague

Step 3: Newspaper Library Extracted Data:

Extracted Data

http://www.foxnews.com/us/2017/11/16/mystery-buyer-who-purchased-450-3-million-leonardo-da-vinci-painting.html

Title	Mystery buyer: Who purchased the $450.3 million Leonardo da Vinci painting?
Authors	
Text	Who purchased the Leonardo da Vinci painting depicting Jesus Christ for an astonishing $450.3 million at Christie's action house in New York Wednesday night? The buyer's identity, which is a mystery, has intrigued the world, becoming a source of intense speculation among international art dealers and on social media about who holds the painting by the Italian Renaissance master. A Russian oligarch, a Saudi prince or a Japanese billionaire? No one knows. "I think the buyer is unlikely to be an institution. I think it's an individual," said Warren Adelson, who was part of a consortium of art dealers who found and restored the painting in Louisiana in 2005. "When we had the picture, it was my burning desire to sell it to an American museum because I wanted to keep it in this country," Adelson told Fox News Thursday. "I felt very strongly about that." "I offered it to a number of institutions in the U.S. and
Article HTML	Who purchased the Leonardo da Vinci painting depicting Jesus Christ for an astonishing $450.3 million at Christie's action house in New York Wednesday night? The buyer's identity, which is a mystery, has intrigued the world, becoming a source of intense speculation among international art dealers and on social media about who holds the painting by the Italian Renaissance master.

Step 3: Newspaper Library – Thin Documentation

- Hows' it work?
- What are all the commands?
- Any limitations?

Step 3: No Soup for You…

- Hows' it work?
- What are all the commands?
- Any limitations?

Solution ?= **Beautiful Soup**

"A tremendous boon." -- Python411 Podcast

[Download | Documentation | Hall of Fame | Source | Discussion group | Zine]

devleague

Fin
Pau
Done

devleague

Sprint: Reading Web News S