Building End Use Solutions

Using Machine Learning Models

Shivas Jayaram

JANUARY 8, 2020

HARVARD UNIVERSITY

Overview

Who I am:

- Harvard Data Science, 2020
- CTO @ BrainCradle (Originated from Harvard i-Lab)
- Data Scientist & Business Solutions Architect @ Sonoco Products Company (Previously)
- https://www.linkedin.com/in/shivasj/

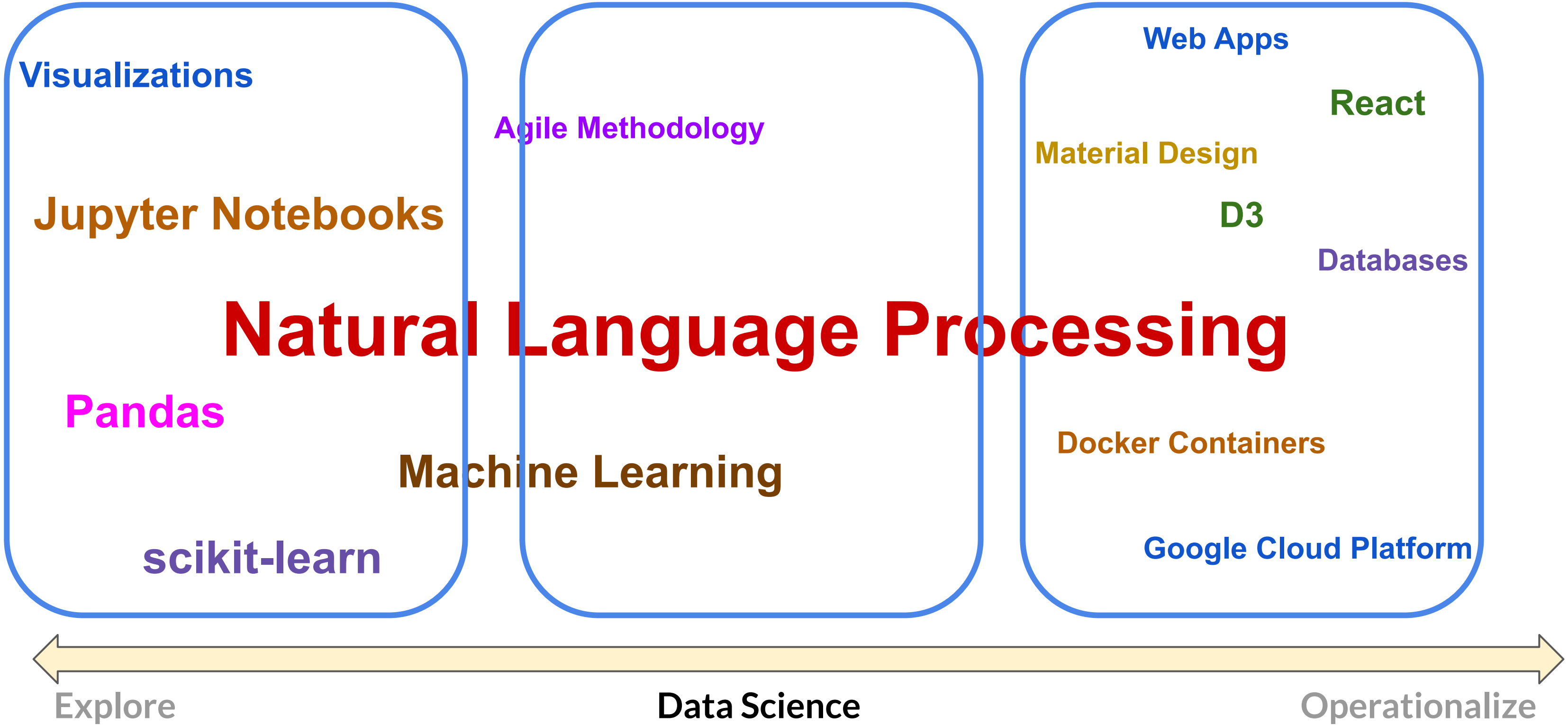
What we are doing today:

- Review workflow to build an end to end application
- Extracting content from the web
- Explore models for NLP
- Using Docker containers
- Setting up a project development environment
- Building an End Use Application

GitHub Repo

- **Main repo:** https://github.com/shivasj/nlp-textanalyzer-application
- **Backend Container:** https://github.com/shivasj/nlp-textanalyzer-application/tree/master/platform-container
- **Frontend React App:** https://github.com/shivasj/nlp-textanalyzer-application/tree/master/platform-client
- **Jupyter Notebooks:** https://github.com/shivasj/nlp-textanalyzer-application/tree/master/platform-container/notebooks

Approach



Web Apps

Visualizations

React

Agile Methodology

Material Design

Jupyter Notebooks

D3

Databases

Natural Language Processing

Pandas

Docker Containers

Machine Learning

Google Cloud Platform

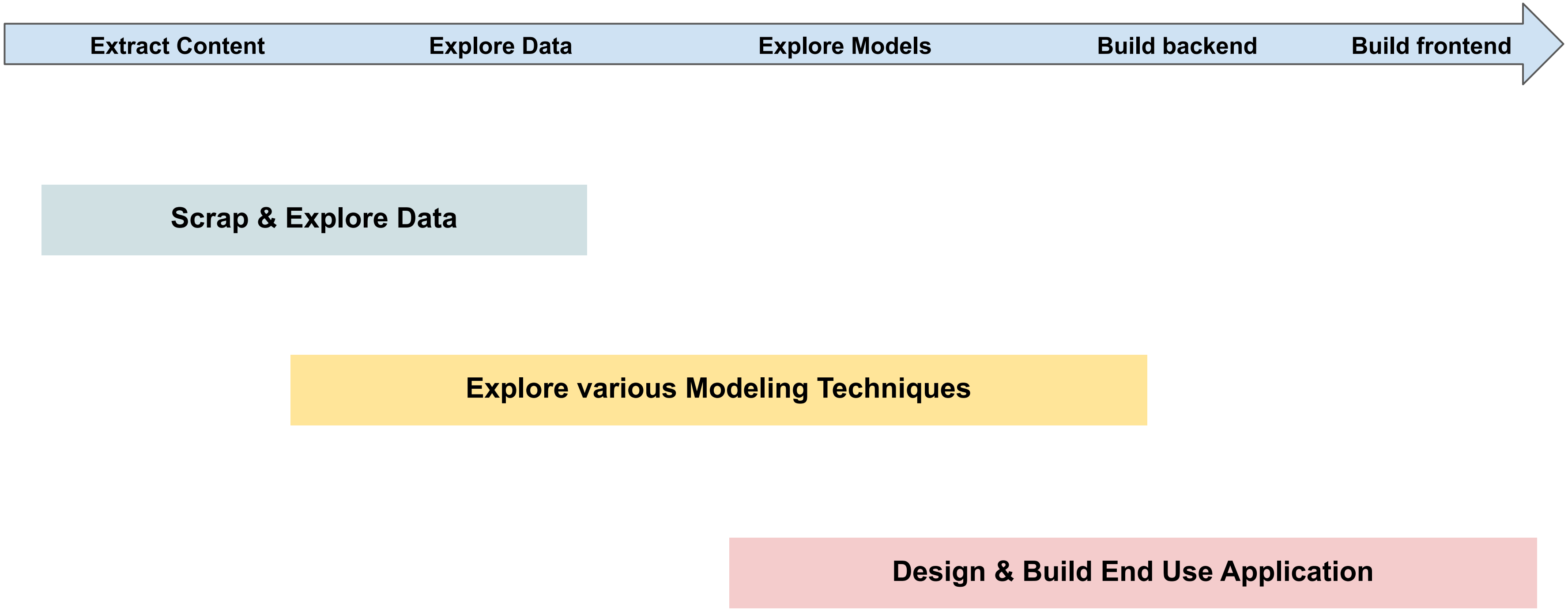
scikit-learn

Explore

Data Science

Operationalize

Workflow



Extract Content

Explore Data

Explore Models

Build backend

Build frontend

Scrap & Explore Data

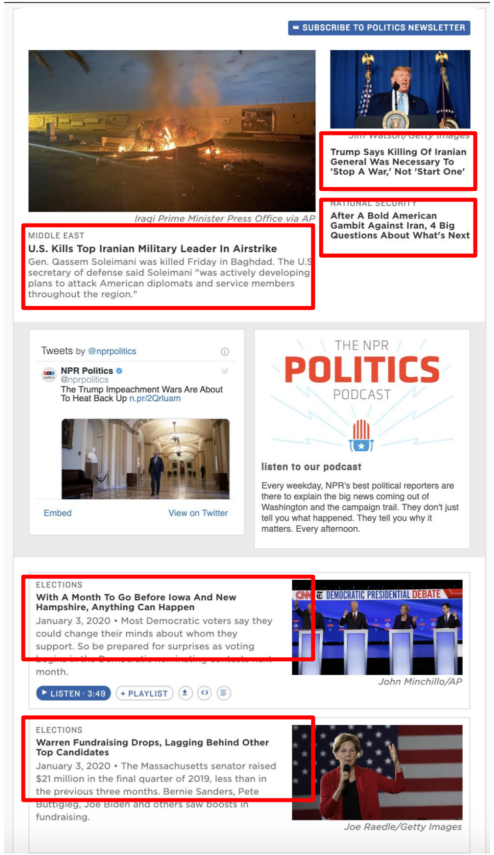
Explore various Modeling Techniques

Design & Build End Use Application

Extracting Content

- **Identify news sources**
  - https://www.npr.org/sections/politics/
  - http://nymag.com/intelligencer/
  - https://edition.cnn.com/specials/last-50-stories
  - https://www.politico.com/news/2020-elections
  - ...
- **Determine navigation hierarchy and content structure**
- **Scrap web page and extract content**
- **Save content**
  - source
  - article_link + article link hash as unique id
  - article_date + date time stamp
  - article_title
  - article_content

Extracting Content



▪ SUBSCRIBE TO POLITICS NEWSLETTER

Iraqi Prime Minister Press Office via AP

MIDDLE EAST

U.S. Kills Top Iranian Military Leader In Airstrike

Gen. Qassem Soleimani was killed Friday in Baghdad. The U.S secretary of defense said Soleimani "was actively developing plans to attack American diplomats and service members throughout the region."

Jim Watson/Getty Images

Trump Says Killing Of Iranian General Was Necessary To 'Stop A War,' Not 'Start One'

NATIONAL SECURITY

After A Bold American Gambit Against Iran, 4 Big Questions About What's Next

Tweets by @nprpolitics

NPR Politics
@nprpolitics
The Trump Impeachment Wars Are About To Heat Back Up n.pr/2Qrluam

Embed   View on Twitter

THE NPR POLITICS PODCAST

listen to our podcast

Every weekday, NPR's best political reporters are there to explain the big news coming out of Washington and the campaign trail. They don't just tell you what happened. They tell you why it matters. Every afternoon.

ELECTIONS

With A Month To Go Before Iowa And New Hampshire, Anything Can Happen

January 3, 2020 • Most Democratic voters say they could change their minds about whom they support. So be prepared for surprises as voting begins in the Democratic nominating contests next month.

▶ LISTEN · 3:49   + PLAYLIST

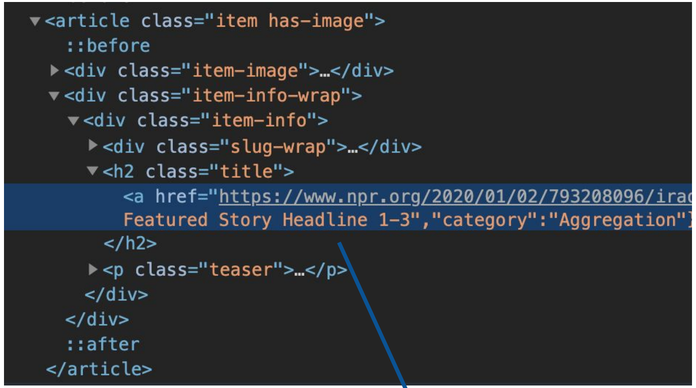
John Minchillo/AP

ELECTIONS

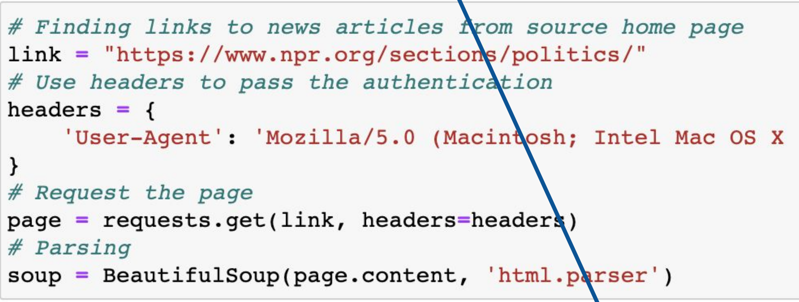
Warren Fundraising Drops, Lagging Behind Other Top Candidates

January 3, 2020 • The Massachusetts senator raised $21 million in the final quarter of 2019, less than in the previous three months. Bernie Sanders, Pete Buttigieg, Joe Biden and others saw boosts in fundraising.

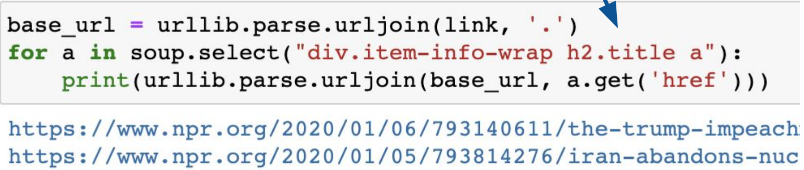
Joe Raedle/Getty Images



```
▼<article class="item has-image">
   ::before
  ▶<div class="item-image">…</div>
  ▼<div class="item-info-wrap">
    ▼<div class="item-info">
      ▶<div class="slug-wrap">…</div>
      ▼<h2 class="title">
          <a href="https://www.npr.org/2020/01/02/793208096/iraq
          Featured Story Headline 1-3","category":"Aggregation"}
        </h2>
      ▶<p class="teaser">…</p>
      </div>
    </div>
   ::after
 </article>
```



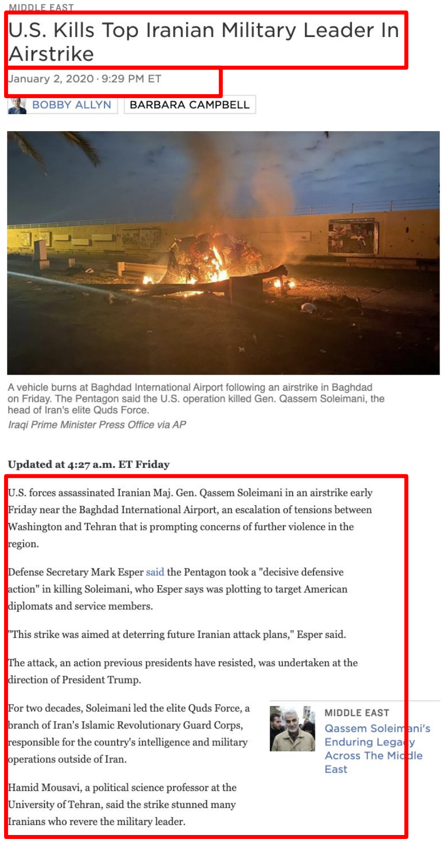
```
# Finding links to news articles from source home page
link = "https://www.npr.org/sections/politics/"
# Use headers to pass the authentication
headers = {
    'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X
}
# Request the page
page = requests.get(link, headers=headers)
# Parsing
soup = BeautifulSoup(page.content, 'html.parser')
```



```
base_url = urllib.parse.urljoin(link, '.')
for a in soup.select("div.item-info-wrap h2.title a"):
    print(urllib.parse.urljoin(base_url, a.get('href')))
```

```
https://www.npr.org/2020/01/06/793140611/the-trump-impeach
https://www.npr.org/2020/01/05/793814276/iran-abandons-nuc
```

Extracting Content



MIDDLE EAST

U.S. Kills Top Iranian Military Leader In Airstrike

January 2, 2020 · 9:29 PM ET

BOBBY ALLYN BARBARA CAMPBELL

A vehicle burns at Baghdad International Airport following an airstrike in Baghdad on Friday. The Pentagon said the U.S. operation killed Gen. Qassem Soleimani, the head of Iran's elite Quds Force.

Iraqi Prime Minister Press Office via AP

Updated at 4:27 a.m. ET Friday

U.S. forces assassinated Iranian Maj. Gen. Qassem Soleimani in an airstrike early Friday near the Baghdad International Airport, an escalation of tensions between Washington and Tehran that is prompting concerns of further violence in the region.

Defense Secretary Mark Esper said the Pentagon took a "decisive defensive action" in killing Soleimani, who Esper says was plotting to target American diplomats and service members.

"This strike was aimed at deterring future Iranian attack plans," Esper said.

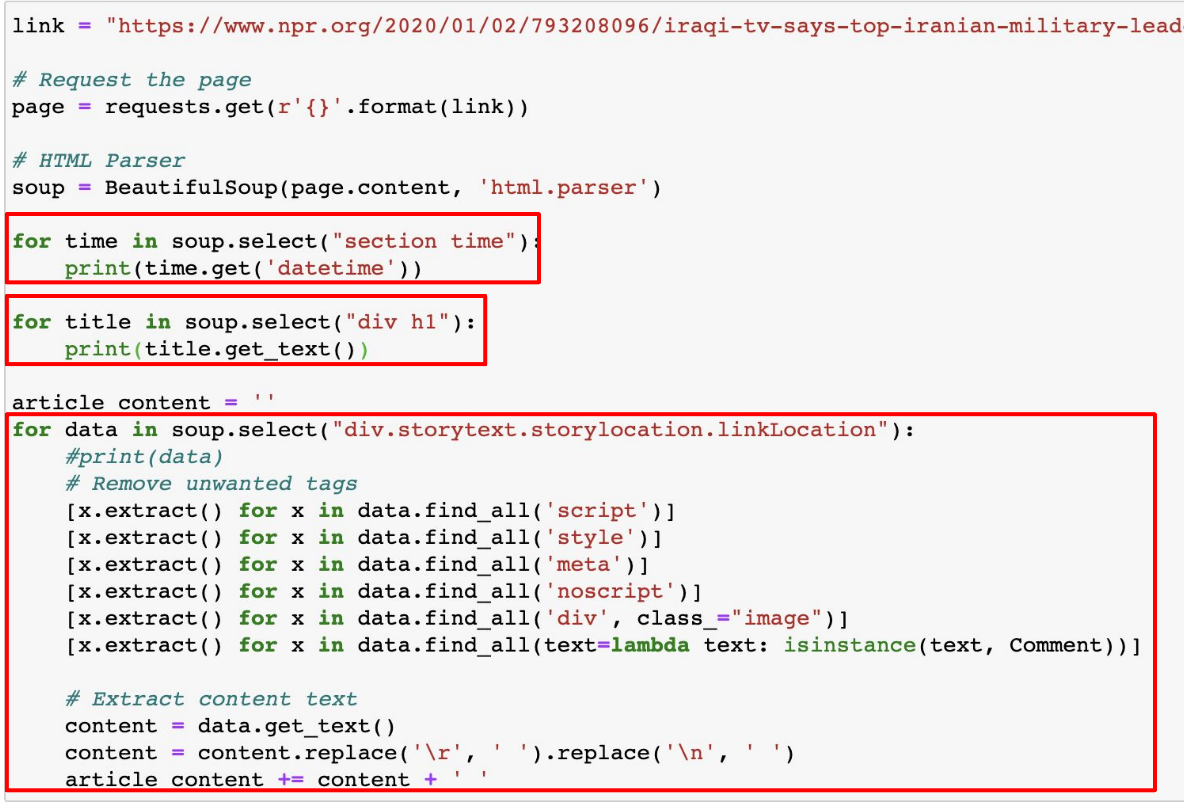
The attack, an action previous presidents have resisted, was undertaken at the direction of President Trump.

For two decades, Soleimani led the elite Quds Force, a branch of Iran's Islamic Revolutionary Guard Corps, responsible for the country's intelligence and military operations outside of Iran.

MIDDLE EAST

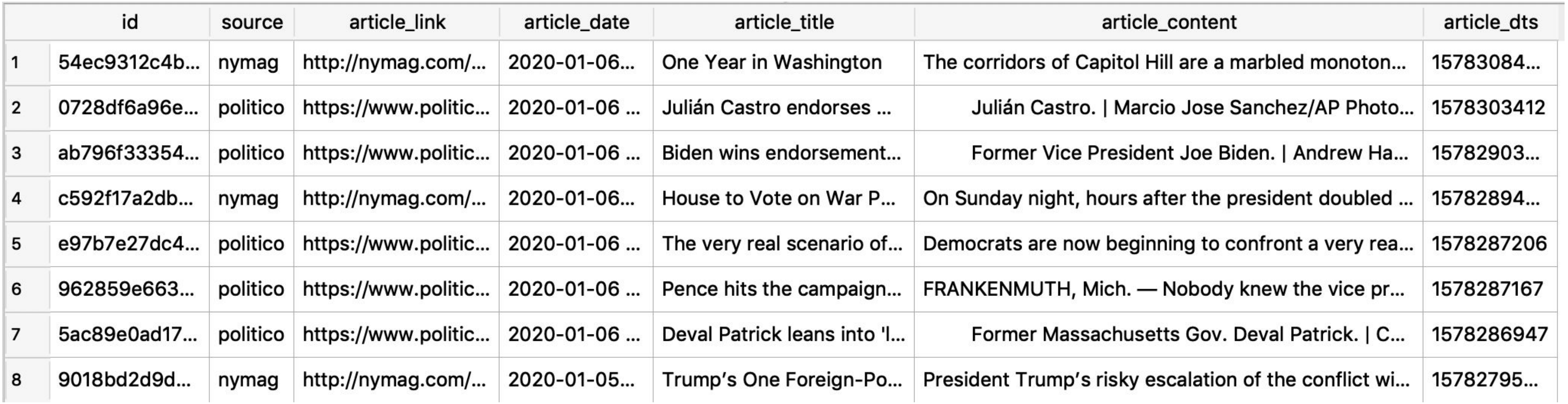
Qassem Soleimani's Enduring Legacy Across The Middle East

Hamid Mousavi, a political science professor at the University of Tehran, said the strike stunned many Iranians who revere the military leader.



```
link = "https://www.npr.org/2020/01/02/793208096/iraqi-tv-says-top-iranian-military-lead

# Request the page
page = requests.get(r'{}'.format(link))

# HTML Parser
soup = BeautifulSoup(page.content, 'html.parser')

for time in soup.select("section time"):
    print(time.get('datetime'))

for title in soup.select("div h1"):
    print(title.get_text())

article content = ''
for data in soup.select("div.storytext.storylocation.linkLocation"):
    #print(data)
    # Remove unwanted tags
    [x.extract() for x in data.find_all('script')]
    [x.extract() for x in data.find_all('style')]
    [x.extract() for x in data.find_all('meta')]
    [x.extract() for x in data.find_all('noscript')]
    [x.extract() for x in data.find_all('div', class_="image")]
    [x.extract() for x in data.find_all(text=lambda text: isinstance(text, Comment))]

    # Extract content text
    content = data.get_text()
    content = content.replace('\r', ' ').replace('\n', ' ')
    article content += content + ' '
```

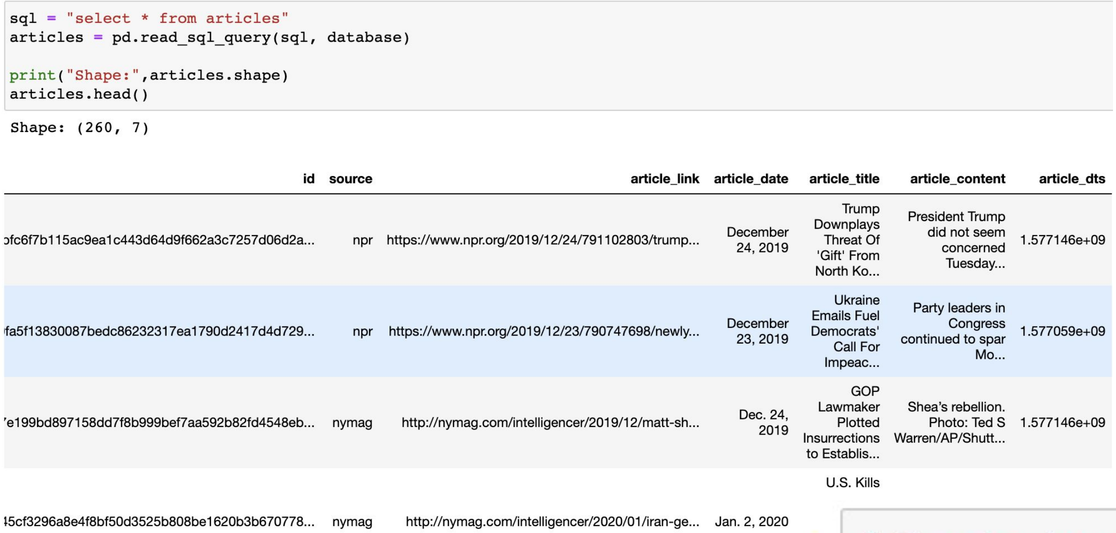
Save Content



	id	source	article_link	article_date	article_title	article_content	article_dts
1	54ec9312c4b...	nymag	http://nymag.com/...	2020-01-06...	One Year in Washington	The corridors of Capitol Hill are a marbled monoton...	15783084...
2	0728df6a96e...	politico	https://www.politic...	2020-01-06 ...	Julián Castro endorses ...	Julián Castro. | Marcio Jose Sanchez/AP Photo...	1578303412
3	ab796f33354...	politico	https://www.politic...	2020-01-06 ...	Biden wins endorsement...	Former Vice President Joe Biden. | Andrew Ha...	15782903...
4	c592f17a2db...	nymag	http://nymag.com/...	2020-01-06...	House to Vote on War P...	On Sunday night, hours after the president doubled ...	15782894...
5	e97b7e27dc4...	politico	https://www.politic...	2020-01-06 ...	The very real scenario of...	Democrats are now beginning to confront a very rea...	1578287206
6	962859e663...	politico	https://www.politic...	2020-01-06 ...	Pence hits the campaign...	FRANKENMUTH, Mich. — Nobody knew the vice pr...	1578287167
7	5ac89e0ad17...	politico	https://www.politic...	2020-01-06 ...	Deval Patrick leans into 'l...	Former Massachusetts Gov. Deval Patrick. | C...	1578286947
8	9018bd2d9d...	nymag	http://nymag.com/...	2020-01-05...	Trump's One Foreign-Po...	President Trump's risky escalation of the conflict wi...	15782795...

Data Store

Explore Data



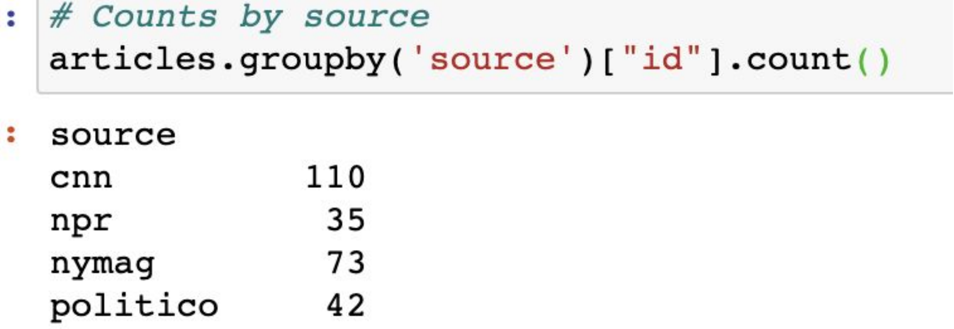
```
sql = "select * from articles"
articles = pd.read_sql_query(sql, database)

print("Shape:",articles.shape)
articles.head()
```

```
Shape: (260, 7)
```

id	source	article_link	article_date	article_title	article_content	article_dts
ɔfc6f7b115ac9ea1c443d64d9f662a3c7257d06d2a...	npr	https://www.npr.org/2019/12/24/791102803/trump...	December 24, 2019	Trump Downplays Threat Of 'Gift' From North Ko...	President Trump did not seem concerned Tuesday...	1.577146e+09
fa5f13830087bedc86232317ea1790d2417d4d729...	npr	https://www.npr.org/2019/12/23/790747698/newly...	December 23, 2019	Ukraine Emails Fuel Democrats' Call For Impeac...	Party leaders in Congress continued to spar Mo...	1.577059e+09
'e199bd897158dd7f8b999bef7aa592b82fd4548eb...	nymag	http://nymag.com/intelligencer/2019/12/matt-sh...	Dec. 24, 2019	GOP Lawmaker Plotted Insurrections to Establis...	Shea's rebellion. Photo: Ted S Warren/AP/Shutt...	1.577146e+09
ł5cf3296a8e4f8bf50d3525b808be1620b3b670778...	nymag	http://nymag.com/intelligencer/2020/01/iran-ge...	Jan. 2, 2020	U.S. Kills		

□ 01_Scrap_Data.ipynb

□ 02_Explore_Data.ipynb

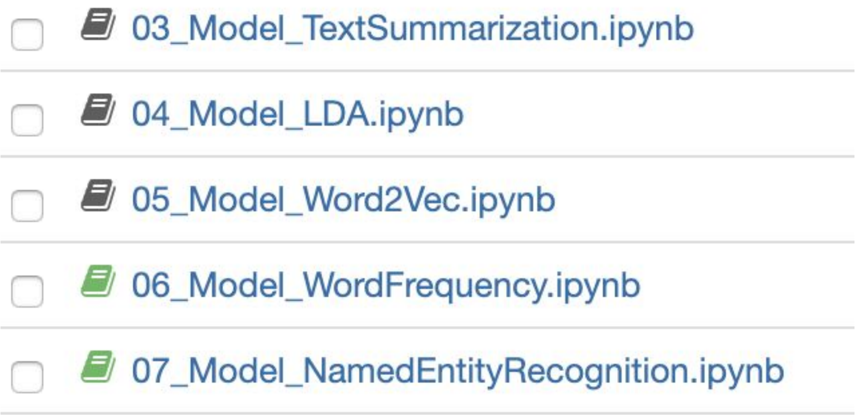


```
# Counts by source
articles.groupby('source')["id"].count()
```

```
source
cnn          110
npr           35
nymag         73
politico      42
```

Explore NLP Models

- Text Summarization
- Named Entity Recognition
- Latent Dirichlet Allocation (LDA)
- Word2Vec
- Word Frequency



□ 03_Model_TextSummarization.ipynb

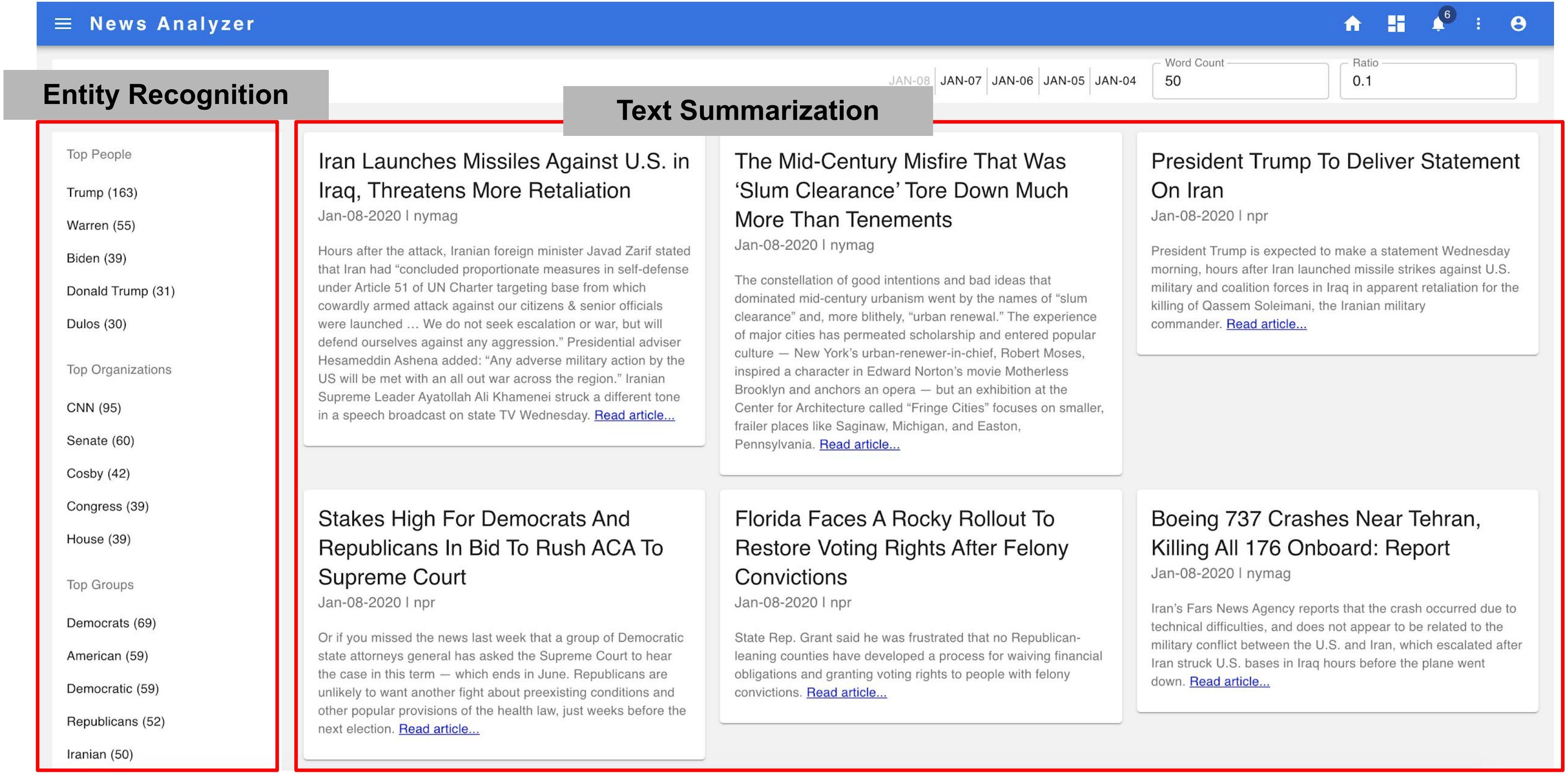
□ 04_Model_LDA.ipynb

□ 05_Model_Word2Vec.ipynb

□ 06_Model_WordFrequency.ipynb

□ 07_Model_NamedEntityRecognition.ipynb

End Use Solution



JAN-08 | JAN-07 | JAN-06 | JAN-05 | JAN-04

Word Count: 50

Ratio: 0.1

Top People

Trump	(163)
Warren	(55)
Biden	(39)
Donald Trump	(31)
Dulos	(30)

Top Organizations

CNN	(95)
Senate	(60)
Cosby	(42)
Congress	(39)
House	(39)

Top Groups

Democrats	(69)
American	(59)
Democratic	(59)
Republicans	(52)
Iranian	(50)

Iran Launches Missiles Against U.S. in Iraq, Threatens More Retaliation

Jan-08-2020 I nymag

Hours after the attack, Iranian foreign minister Javad Zarif stated that Iran had “concluded proportionate measures in self-defense under Article 51 of UN Charter targeting base from which cowardly armed attack against our citizens & senior officials were launched … We do not seek escalation or war, but will defend ourselves against any aggression.” Presidential adviser Hesameddin Ashena added: “Any adverse military action by the US will be met with an all out war across the region.” Iranian Supreme Leader Ayatollah Ali Khamenei struck a different tone in a speech broadcast on state TV Wednesday. Read article...

The Mid-Century Misfire That Was ‘Slum Clearance’ Tore Down Much More Than Tenements

Jan-08-2020 I nymag

The constellation of good intentions and bad ideas that dominated mid-century urbanism went by the names of “slum clearance” and, more blithely, “urban renewal.” The experience of major cities has permeated scholarship and entered popular culture — New York’s urban-renewer-in-chief, Robert Moses, inspired a character in Edward Norton’s movie Motherless Brooklyn and anchors an opera — but an exhibition at the Center for Architecture called “Fringe Cities” focuses on smaller, frailer places like Saginaw, Michigan, and Easton, Pennsylvania. Read article...

President Trump To Deliver Statement On Iran

Jan-08-2020 I npr

President Trump is expected to make a statement Wednesday morning, hours after Iran launched missile strikes against U.S. military and coalition forces in Iraq in apparent retaliation for the killing of Qassem Soleimani, the Iranian military commander. Read article...

Stakes High For Democrats And Republicans In Bid To Rush ACA To Supreme Court

Jan-08-2020 I npr

Or if you missed the news last week that a group of Democratic state attorneys general has asked the Supreme Court to hear the case in this term — which ends in June. Republicans are unlikely to want another fight about preexisting conditions and other popular provisions of the health law, just weeks before the next election. Read article...

Florida Faces A Rocky Rollout To Restore Voting Rights After Felony Convictions

Jan-08-2020 I npr

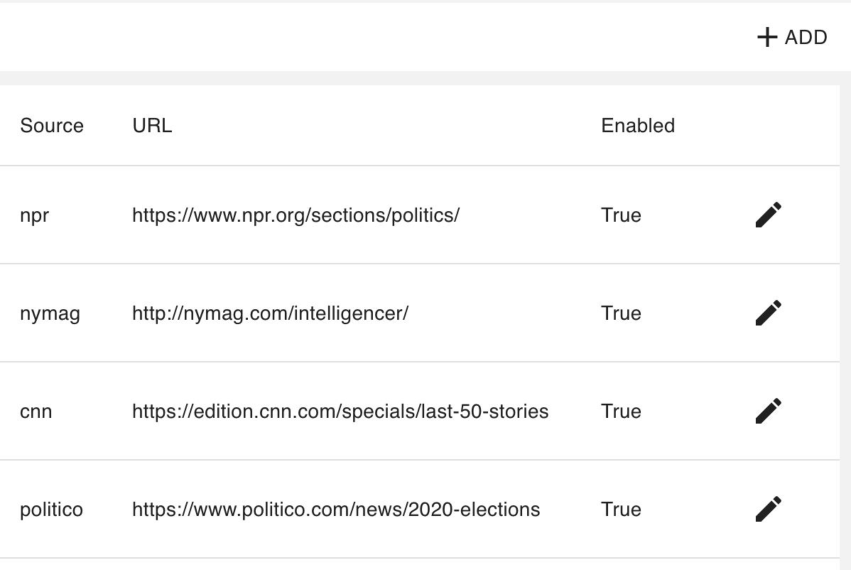
State Rep. Grant said he was frustrated that no Republican-leaning counties have developed a process for waiving financial obligations and granting voting rights to people with felony convictions. Read article...

Boeing 737 Crashes Near Tehran, Killing All 176 Onboard: Report

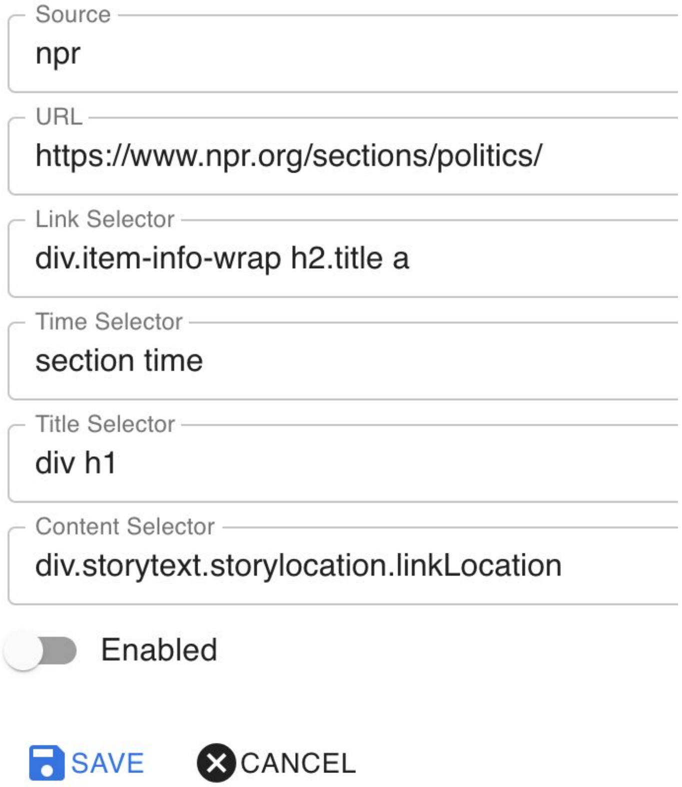
Jan-08-2020 I nymag

Iran’s Fars News Agency reports that the crash occurred due to technical difficulties, and does not appear to be related to the military conflict between the U.S. and Iran, which escalated after Iran struck U.S. bases in Iraq hours before the plane went down. Read article...

End Use Solution



Source	URL	Enabled
npr	https://www.npr.org/sections/politics/	True
nymag	http://nymag.com/intelligencer/	True
cnn	https://edition.cnn.com/specials/last-50-stories	True
politico	https://www.politico.com/news/2020-elections	True



Source: npr

URL: https://www.npr.org/sections/politics/

Link Selector: div.item-info-wrap h2.title a

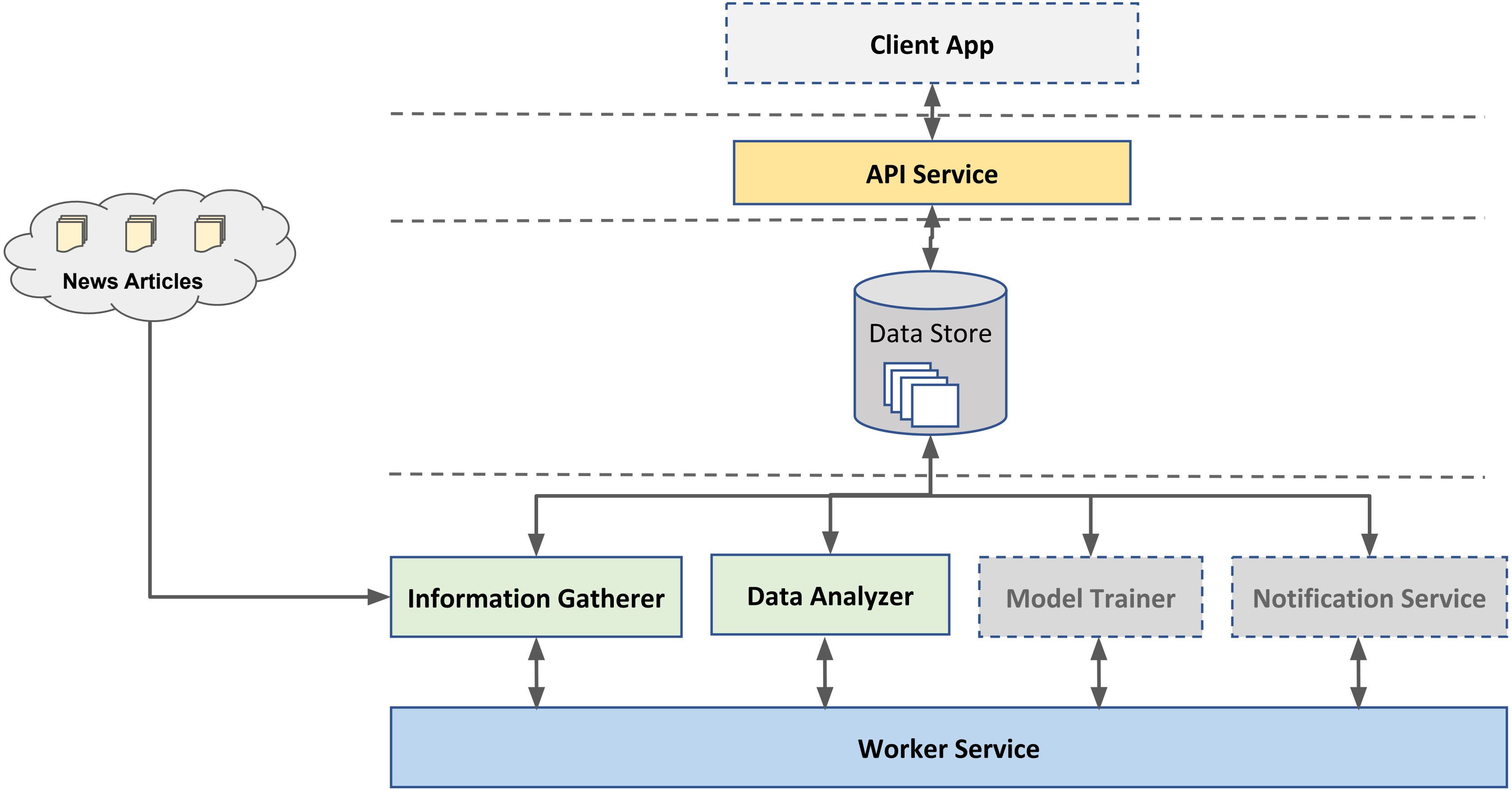
Time Selector: section time

Title Selector: div h1

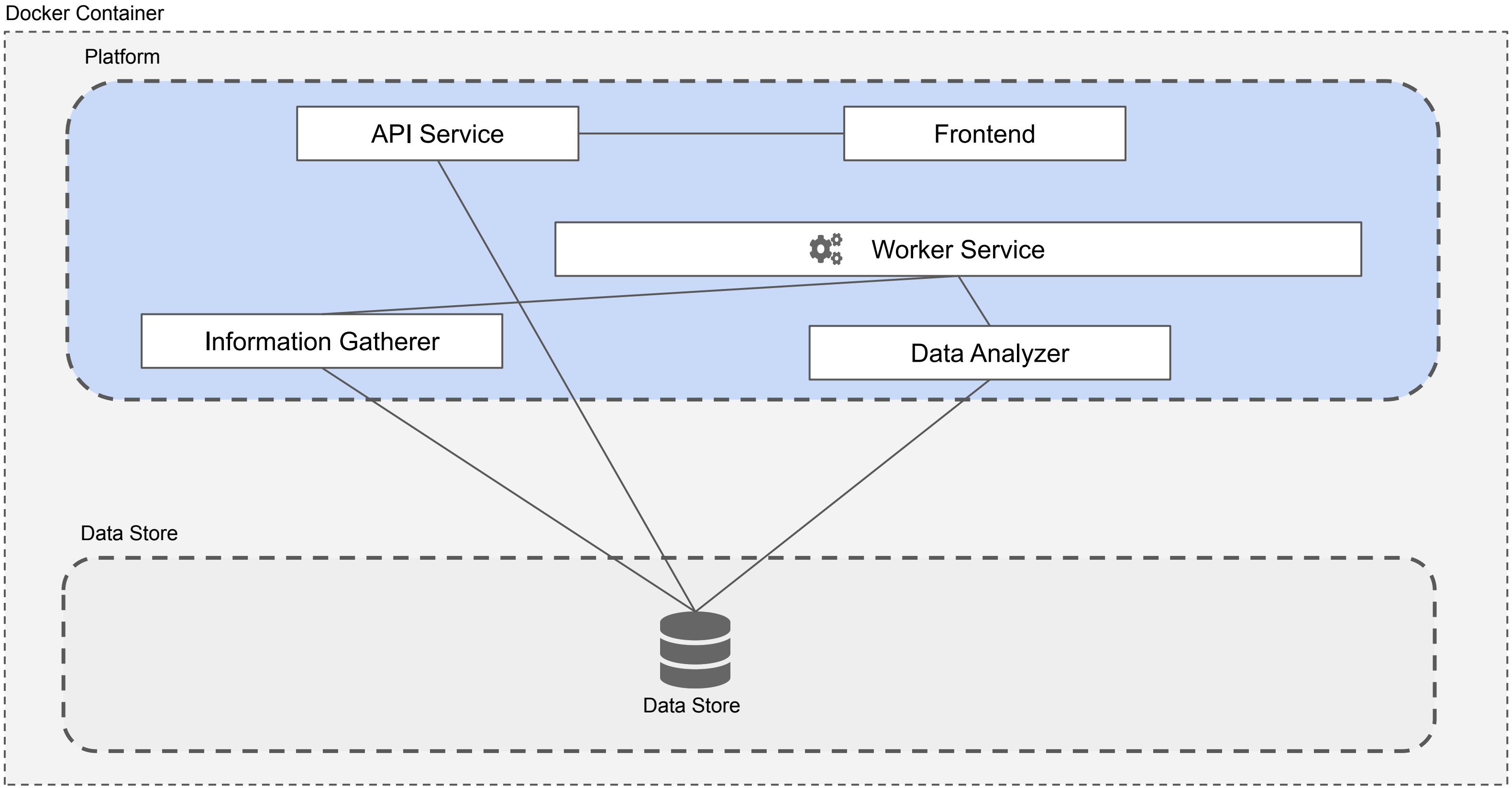
Content Selector: div.storytext.storylocation.linkLocation

Enabled

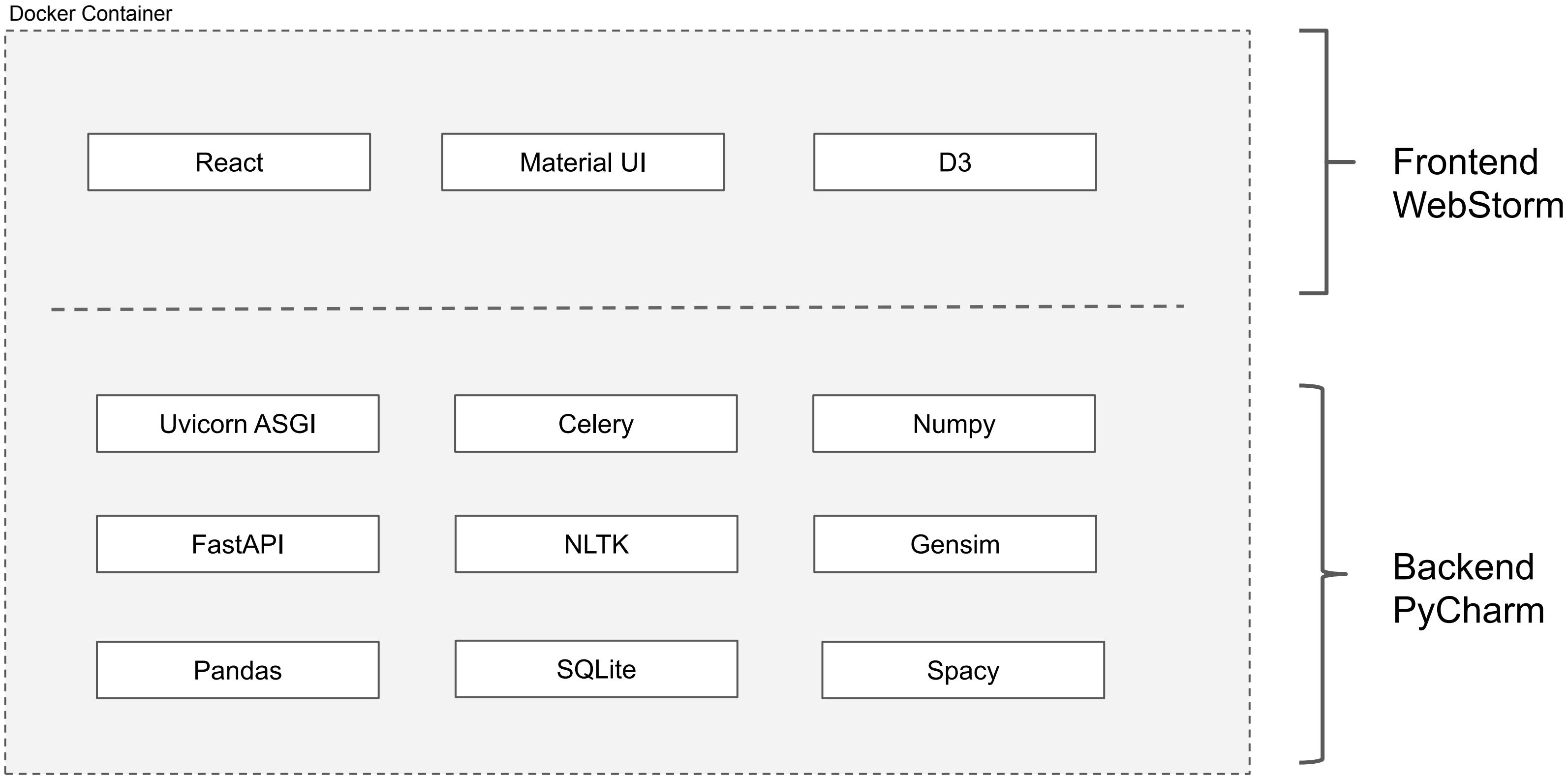
Solution Architecture



Technical Architecture



Technology Stack



Docker Container

React

Material UI

D3

Frontend
WebStorm

Uvicorn ASGI

Celery

Numpy

FastAPI

NLTK

Gensim

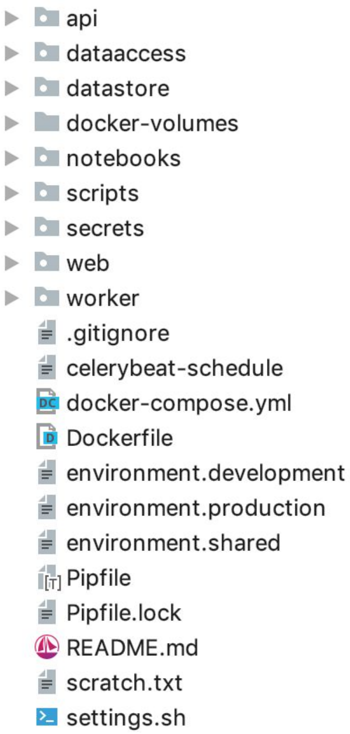
Pandas

SQLite

Spacy

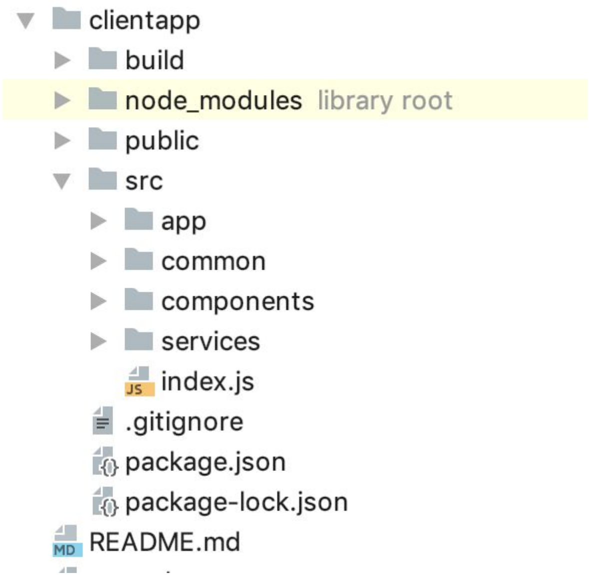
Backend
PyCharm

Development Environment



```
▶ api
▶ dataaccess
▶ datastore
▶ docker-volumes
▶ notebooks
▶ scripts
▶ secrets
▶ web
▶ worker
  .gitignore
  celerybeat-schedule
  docker-compose.yml
  Dockerfile
  environment.development
  environment.production
  environment.shared
  Pipfile
  Pipfile.lock
  README.md
  scratch.txt
  settings.sh
```

Platform Container



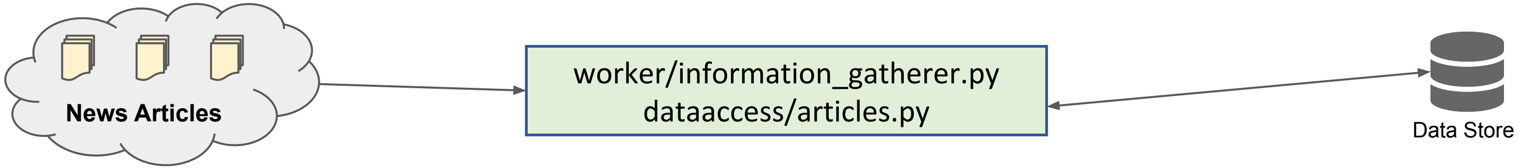
```
▼ clientapp
    ▶ build
    ▶ node_modules  library root
    ▶ public
    ▼ src
        ▶ app
        ▶ common
        ▶ components
        ▶ services
          index.js
      .gitignore
      package.json
      package-lock.json
  README.md
```

Platform Client

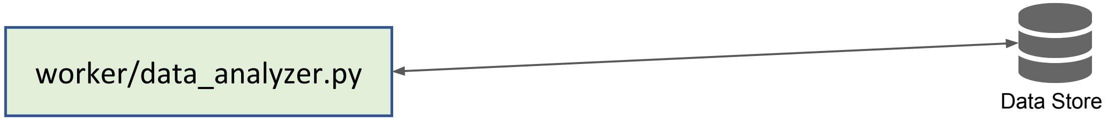
Development Environment

- Starting up Docker container
- Running Jupyter notebooks
- Starting Web server
- Running React App

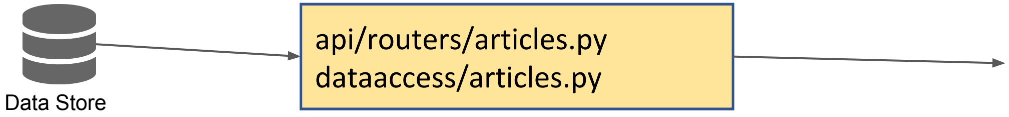
Extracting Content Flow



Named Entity Recognition Flow

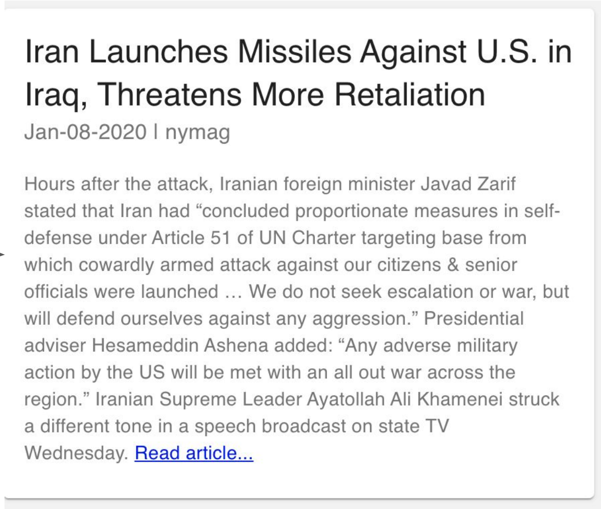


New Analyzer Home Page Flow



Data Store

api/routers/articles.py
dataaccess/articles.py



Iran Launches Missiles Against U.S. in Iraq, Threatens More Retaliation

Jan-08-2020 I nymag

Hours after the attack, Iranian foreign minister Javad Zarif stated that Iran had “concluded proportionate measures in self-defense under Article 51 of UN Charter targeting base from which cowardly armed attack against our citizens & senior officials were launched … We do not seek escalation or war, but will defend ourselves against any aggression.” Presidential adviser Hesameddin Ashena added: “Any adverse military action by the US will be met with an all out war across the region.” Iranian Supreme Leader Ayatollah Ali Khamenei struck a different tone in a speech broadcast on state TV Wednesday. Read article...

/api/v1/articles?summarize=true&day=0&hours=24&summarize_ratio=0.1&summarize_word_count=75

Questions

- ...