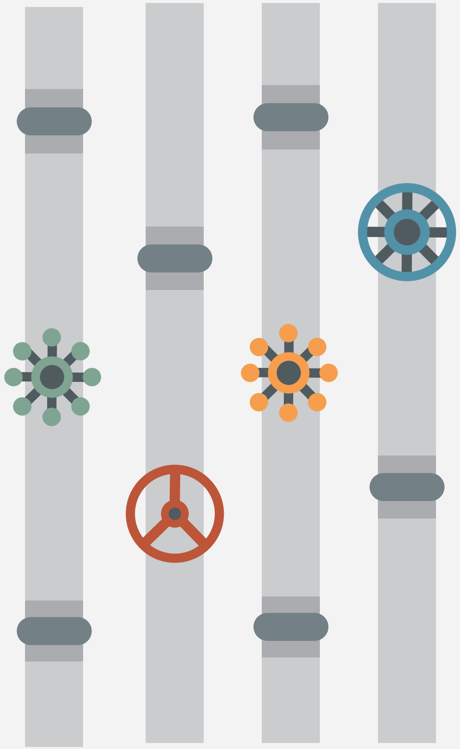
New York Times Topic Modeling End-to-end Data Pipeline Design

Jenica Andersen
Metis DSML, Data Engineering

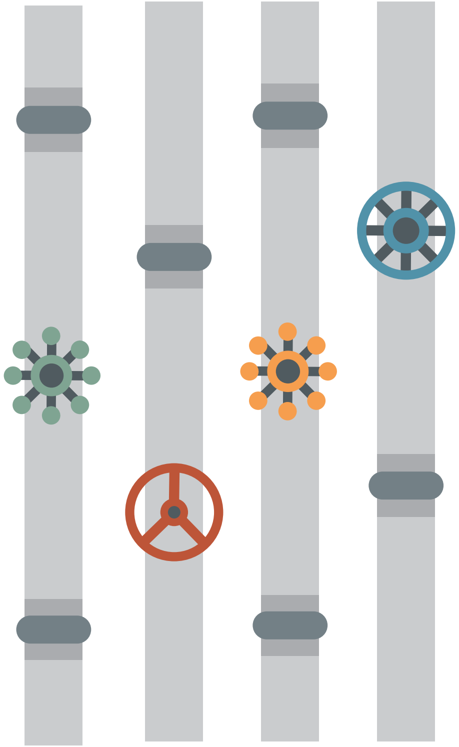
August 10, 2022



Introduction

(After defining question)
Standard Workflow:

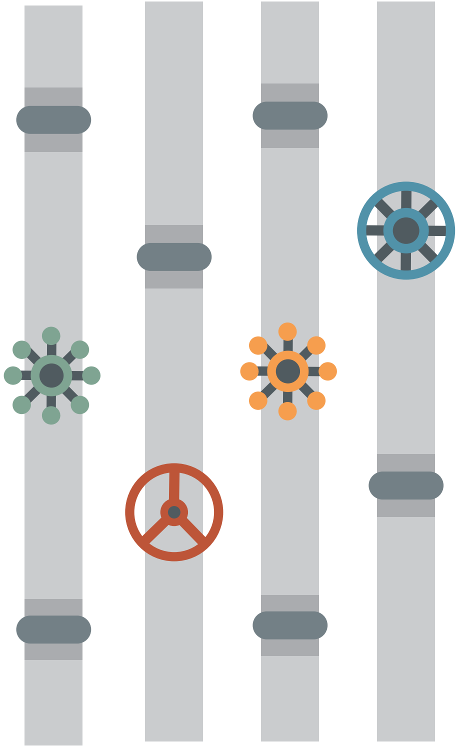
- Create or scrape data
- Pandas DataFrame
- Matplotlib Plots
- Static Slideshow



Problems

Standard Workflow:

- Create or scrape data
- Pandas DataFrame
- Matplotlib Plots
- Static Slideshow

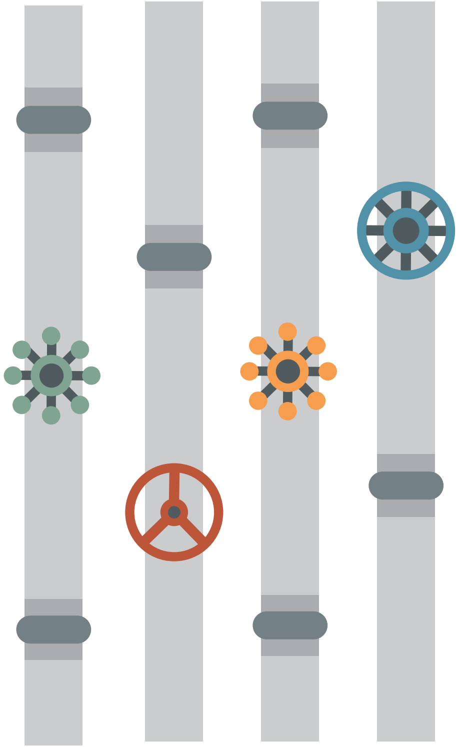


Problems

Standard Workflow:

time consuming, limited data

- Pandas DataFrame
- Matplotlib Plots
- Static Slideshow



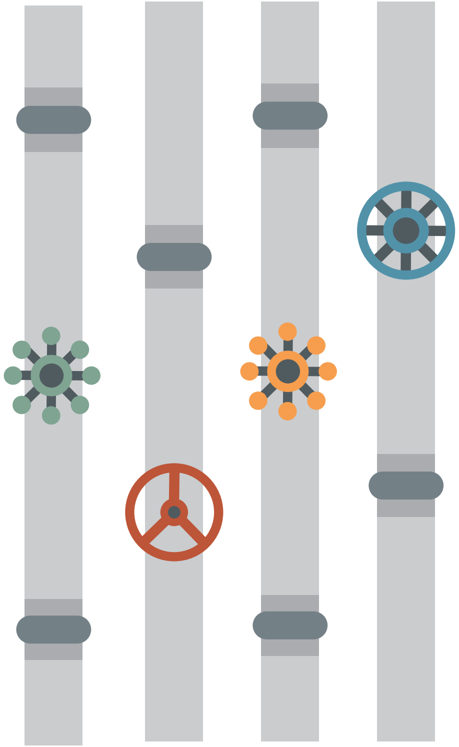
Problems

Standard Workflow:

time consuming, limited data

computationally expensive

- Matplotlib Plots
- Static Slideshow



Problems

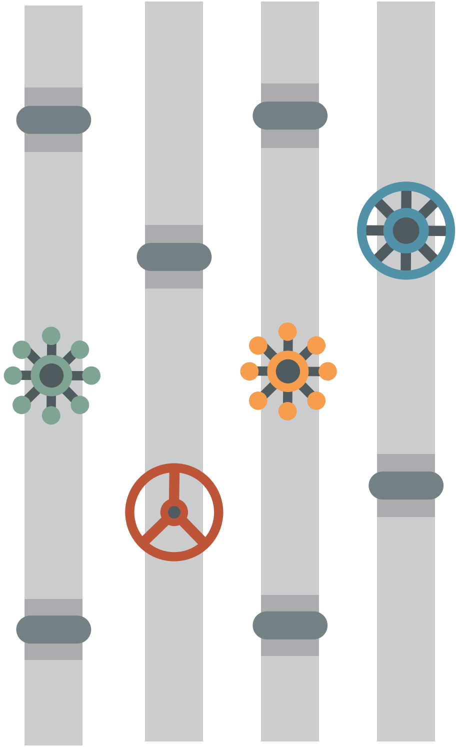
Standard Workflow:

time consuming, limited data

computationally expensive

Static Graphics

- M[illegible]es
- Static Slideshow



Problems

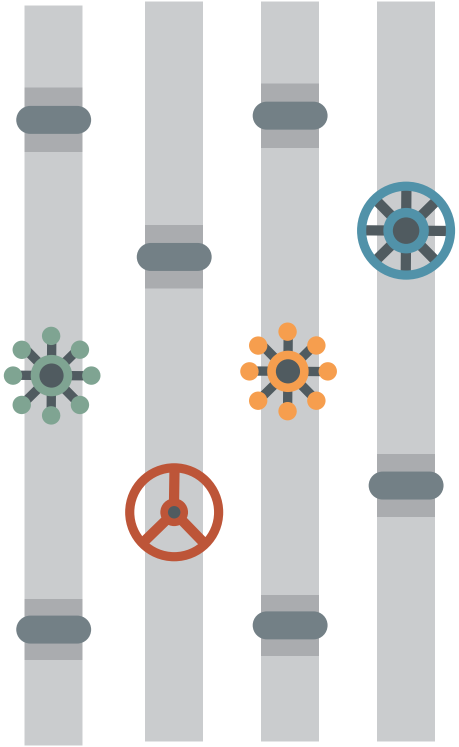
Standard Workflow:

time consuming, limited data

computationally expensive

Static Graphics

Passive deliverables

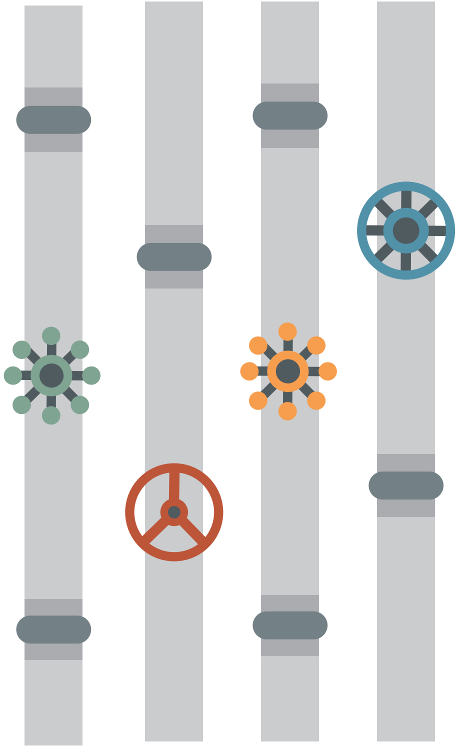


A better way to do things:

Our Task: Topic Model the New York Times

– Client wants contextualized knowledge of events, in order to connect with consumers

– Organize, search, and understand the most read newspaper in the US



New End-to-End Data Pipeline

Data

{𝕿} Developers

The New York Times

In Germany, cool politics as resistance

'Brexit' poses risk to hard-won Irish peace

Concerns over ethical ivory

To build his label, a designer heads to Italy

INTERNATIONAL LUXURY CONFERENCE NOVEMBER 13-14, 2017 BRUSSELS

WHAT'S NEXT: LUXURY IN A TURBULENT WORLD

API-Application Programming Interface

Process

SQL

Amazon EC2

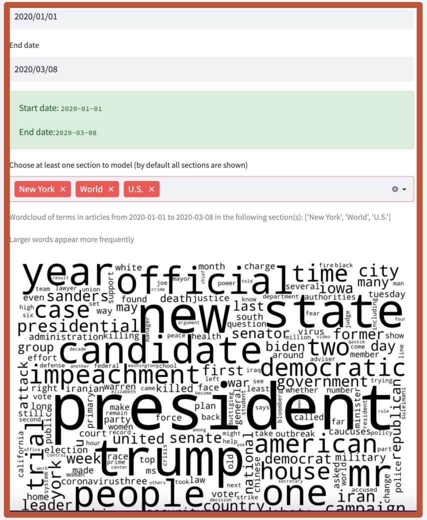
Amazon S3

pandas

Manage data.
Implement Topic Modeling (NMF)

Application

Streamlit



2020/01/01

End date

2020/03/08

Start date: `2020-01-01`

End date:`2020-03-08`

Choose at least one section to model (by default all sections are shown)

New York × World × U.S. ×

Wordcloud of terms in articles from 2020-01-01 to 2020-03-08 in the following section(s): ['New York', 'World', 'U.S.']

Larger words appear more frequently

Interactive results

New End-to-End Data Pipeline

Data

{T} Developers

The New York Times
Concerns over ethical ivory
To build his label, a designer heads to Italy
INTERNATIONAL LUXURY CONFERENCE NOVEMBER 13-14, 2017 BRUSSELS
WHAT'S NEXT: LUXURY IN A TURBULENT WORLD

So much more data available this way

Process

SQL

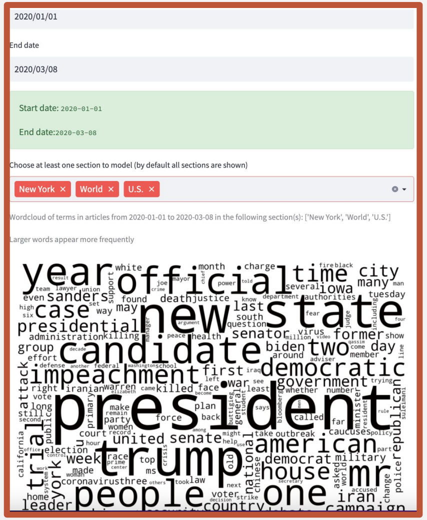
Amazon EC2

Amazon S3

pandas

Application

Streamlit



2020/01/01

End date

2020/03/08

Start date: `2020-01-01`

End date:`2020-03-08`

Choose at least one section to model (by default all sections are shown)

New York × World × U.S. ×

Wordcloud of terms in articles from 2020-01-01 to 2020-03-08 in the following section(s): ['New York', 'World', 'U.S.']

Larger words appear more frequently

New End-to-End Data Pipeline

Data

{T} Developers

The New York Times
Concerns over ethical ivory
To build his label, a designer heads to Italy
INTERNATIONAL LUXURY CONFERENCE NOVEMBER 13-14, 2017 BRUSSELS
WHAT'S NEXT: LUXURY IN A TURBULENT WORLD

So much more data available this way

Process

SQL

Amazon EC2

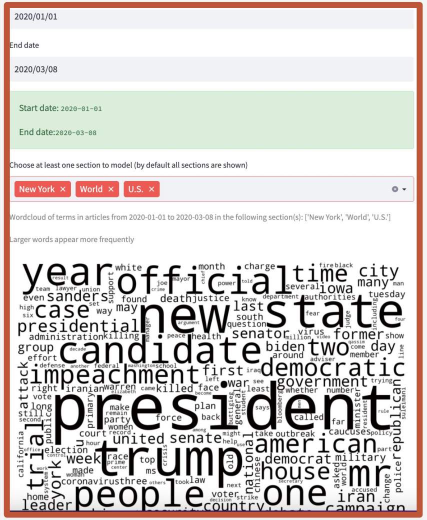
Amazon S3

Scalable solutions

pandas

Application

Streamlit



2020/01/01

End date

2020/03/08

Start date: 2020-01-01

End date:2020-03-08

Choose at least one section to model (by default all sections are shown)

New York × World × U.S. ×

Wordcloud of terms in articles from 2020-01-01 to 2020-03-08 in the following section(s): ['New York', 'World', 'U.S.']

Larger words appear more frequently

New End-to-End Data Pipeline

Data

{T} Developers

So much more data available this way

Process

SQL

Amazon EC2

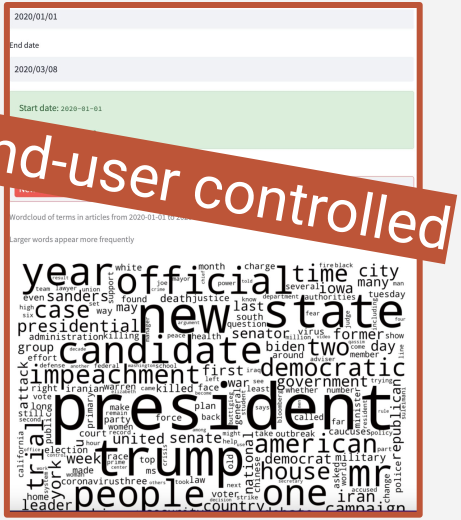
Amazon S3

Scalable solutions

pandas

Application

Streamlit



2020/01/01

End date

2020/03/08

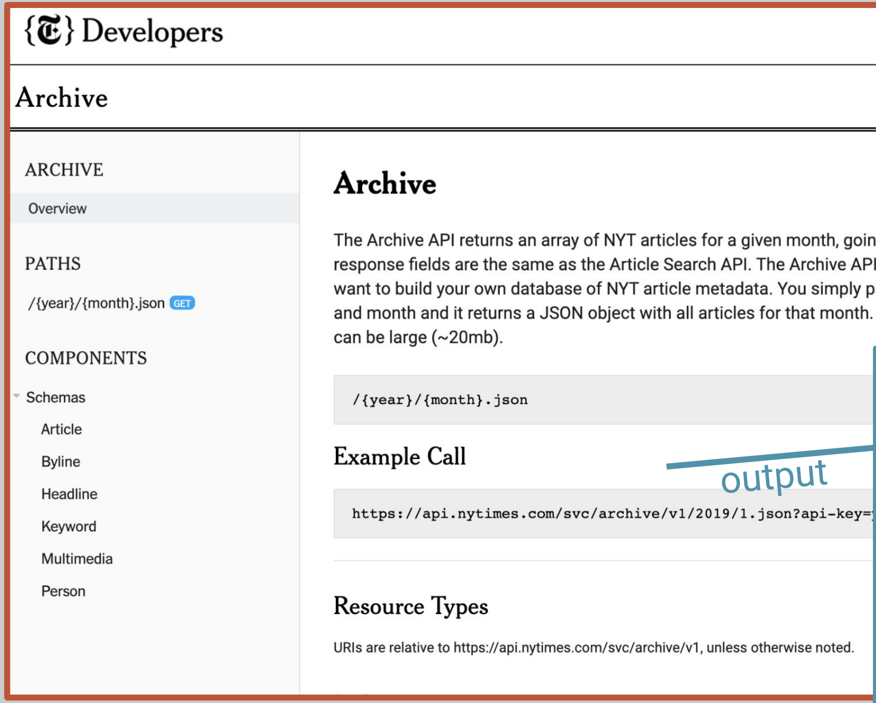
Start date: 2020-01-01

Wordcloud of terms in articles from 2020-01-01 to 2020

Larger words appear more frequently

End-user controlled

The Data: New York Times Archive API



Archive

ARCHIVE

Overview

PATHS

/{year}/{month}.json GET

COMPONENTS

Schemas

- Article
- Byline
- Headline
- Keyword
- Multimedia
- Person

Archive

The Archive API returns an array of NYT articles for a given month, going back to 1851. Its response fields are the same as the Article Search API. The Archive API is very useful if you want to build your own database of NYT article metadata. You simply pass the API the year and month and it returns a JSON object with all articles for that month. The response size can be large (~20mb).

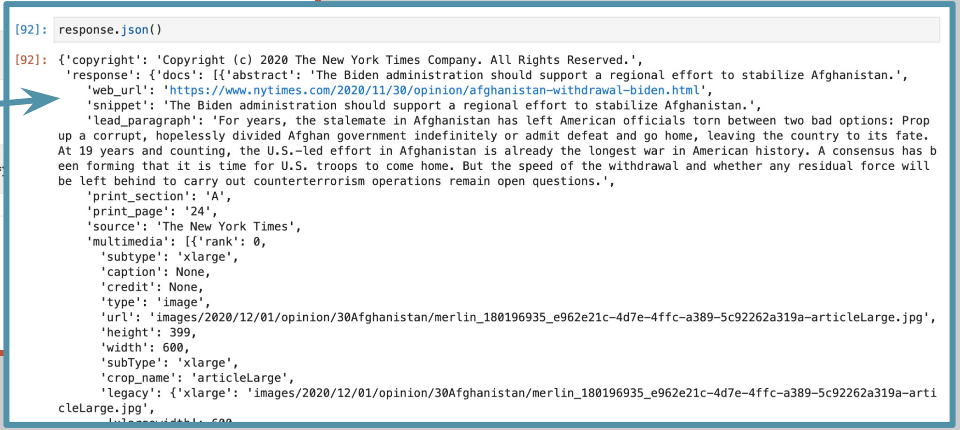
```
/{year}/{month}.json
```

Example Call

```
https://api.nytimes.com/svc/archive/v1/2019/1.json?api-key=
```

Resource Types

URIs are relative to https://api.nytimes.com/svc/archive/v1, unless otherwise noted.

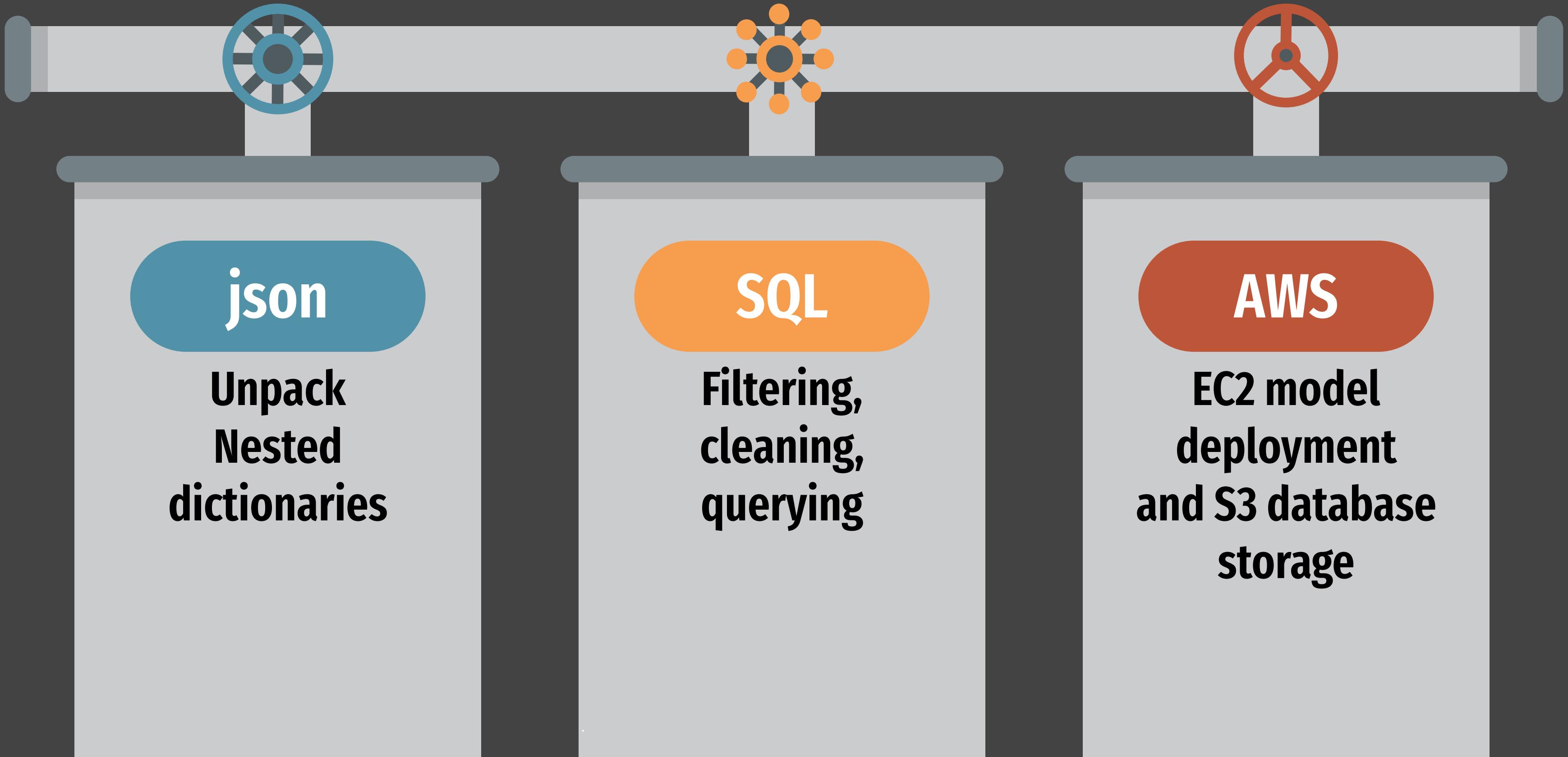


```
[92]: response.json()
```

```
[92]: {'copyright': 'Copyright (c) 2020 The New York Times Company. All Rights Reserved.',
       'response': {'docs': [{'abstract': 'The Biden administration should support a regional effort to stabilize Afghanistan.',
          'web_url': 'https://www.nytimes.com/2020/11/30/opinion/afghanistan-withdrawal-biden.html',
          'snippet': 'The Biden administration should support a regional effort to stabilize Afghanistan.',
          'lead_paragraph': 'For years, the stalemate in Afghanistan has left American officials torn between two bad options: Prop
      up a corrupt, hopelessly divided Afghan government indefinitely or admit defeat and go home, leaving the country to its fate.
      At 19 years and counting, the U.S.-led effort in Afghanistan is already the longest war in American history. A consensus has b
      een forming that it is time for U.S. troops to come home. But the speed of the withdrawal and whether any residual force will
      be left behind to carry out counterterrorism operations remain open questions.',
          'print_section': 'A',
          'print_page': '24',
          'source': 'The New York Times',
          'multimedia': [{'rank': 0,
            'subtype': 'xlarge',
            'caption': None,
            'credit': None,
            'type': 'image',
            'url': 'images/2020/12/01/opinion/30Afghanistan/merlin_180196935_e962e21c-4d7e-4ffc-a389-5c92262a319a-articleLarge.jpg',
            'height': 399,
            'width': 600,
            'subType': 'xlarge',
            'crop_name': 'articleLarge',
            'legacy': {'xlarge': 'images/2020/12/01/opinion/30Afghanistan/merlin_180196935_e962e21c-4d7e-4ffc-a389-5c92262a319a-arti
      cleLarge.jpg',
             'xlargewidth': 600
```

- 10 years of articles (747,468)
- Raw data: 3.8 GB
- Data of Interest: 577 MB
  - Date Published
  - Headline
  - Snippet
  - URL

Processing



json

Unpack Nested dictionaries

SQL

Filtering, cleaning, querying

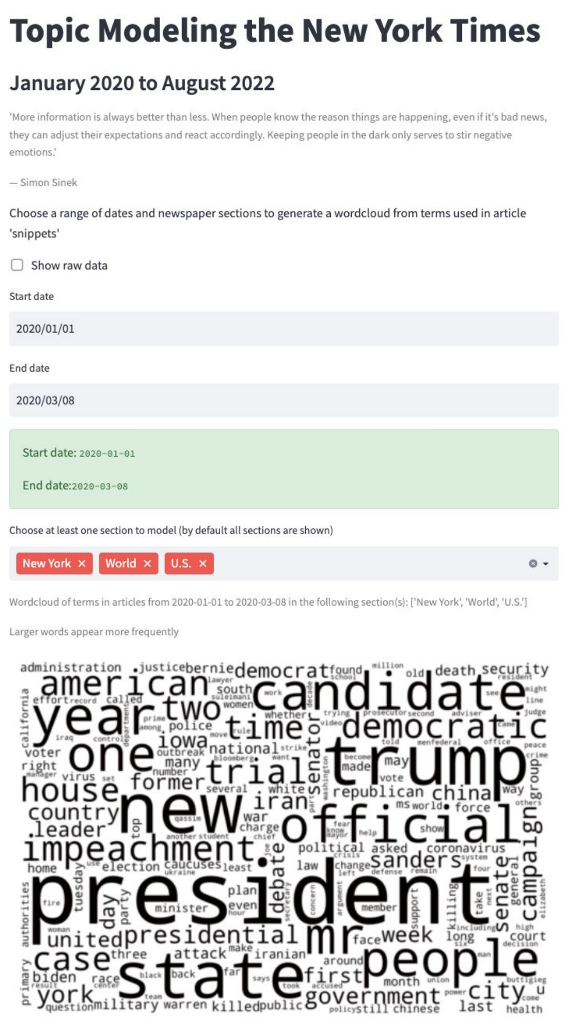
AWS

EC2 model deployment and S3 database storage

Streamlit + GitHub

Live Application Deliverables

https://jenica-a-nyt-wordcloud-nyt-app-uqe5rx.streamlitapp.com/



Topic Modeling the New York Times

January 2020 to August 2022

'More information is always better than less. When people know the reason things are happening, even if it's bad news, they can adjust their expectations and react accordingly. Keeping people in the dark only serves to stir negative emotions.'

— Simon Sinek

Choose a range of dates and newspaper sections to generate a wordcloud from terms used in article 'snippets'

☐ Show raw data

Start date

2020/01/01

End date

2020/03/08

Start date: `2020-01-01`

End date:`2020-03-08`

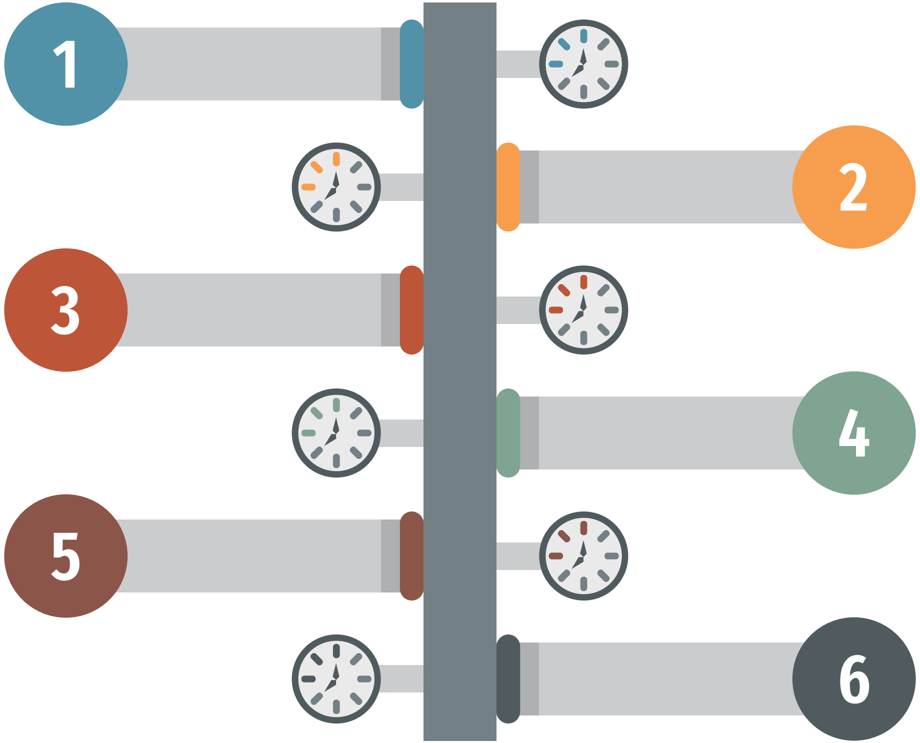
Choose at least one section to model (by default all sections are shown)

New York ✕ World ✕ U.S. ✕

Wordcloud of terms in articles from 2020-01-01 to 2020-03-08 in the following section(s): ['New York', 'World', 'U.S.']

Larger words appear more frequently

Next Steps



More Data & EC2

Get entire NYT archive, 1851 to present and use larger EC2 instance type

NMF

Move results from log to application interface and customize error response

End-user word entry

Add more powerful app features like links to URL

Clean “keywords”

And other fields, like “headline”, to be usable in app or model

MongoDB

Streamline json unpacking

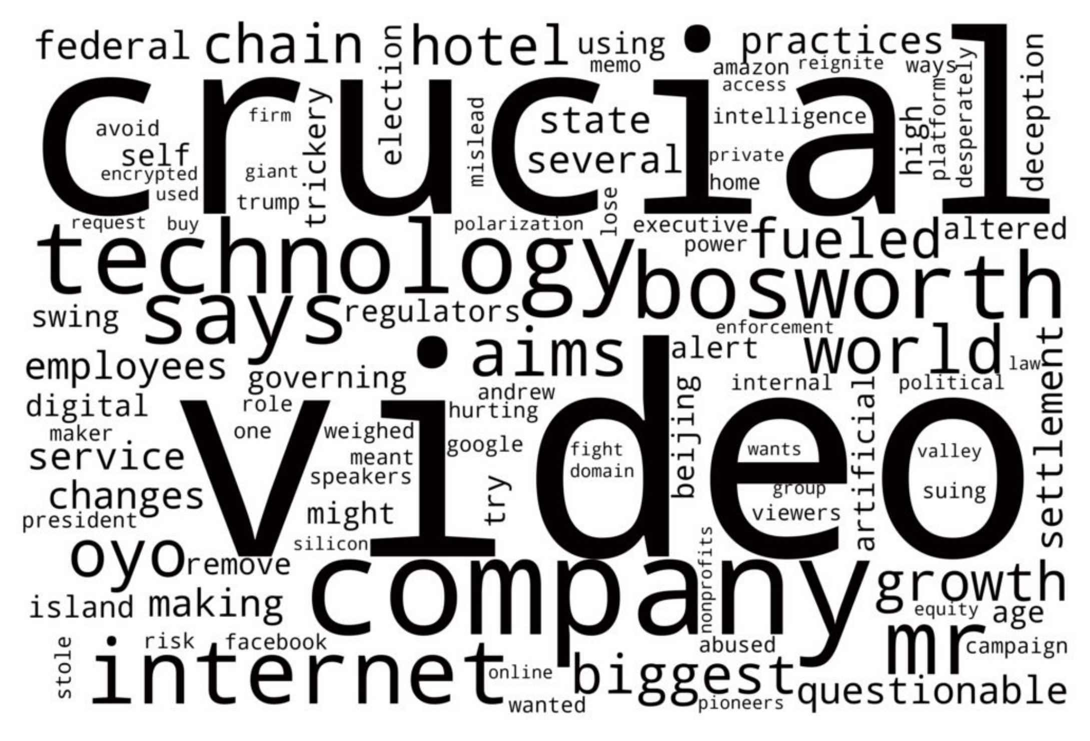
Spark

Leverage cluster computing for quicker modeling

Data Pipeline Summary

✗	**Previous Workflow**	Good but Limited
✓	**Better Pipeline**	Scalable, Dynamic, Iterative Improvements
✓	**Data Acquisition**	Website API
✓	**Data Processing**	json→sql→aws→NLP modeling
✗	**Data Application**	End-user controlled (Streamlit) App

Thank you!

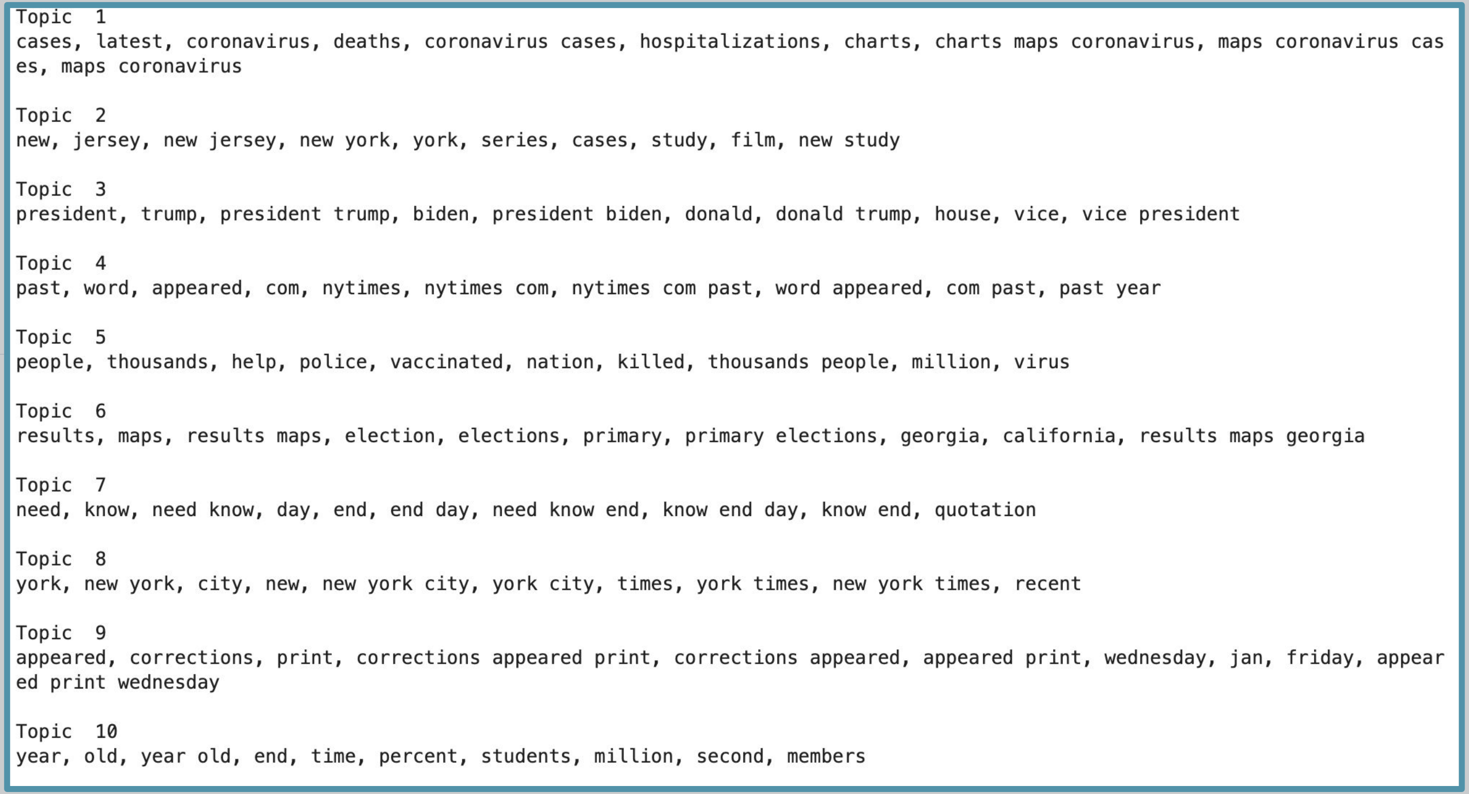


WordCloud of Jan 2020 Technology Article Snippets

Infographics and slides by SlidesGo

New York Times Topic Model

From random sample of 5000 articles from the last 2 years



```
Topic  1
cases, latest, coronavirus, deaths, coronavirus cases, hospitalizations, charts, charts maps coronavirus, maps coronavirus cas
es, maps coronavirus

Topic  2
new, jersey, new jersey, new york, york, series, cases, study, film, new study

Topic  3
president, trump, president trump, biden, president biden, donald, donald trump, house, vice, vice president

Topic  4
past, word, appeared, com, nytimes, nytimes com, nytimes com past, word appeared, com past, past year

Topic  5
people, thousands, help, police, vaccinated, nation, killed, thousands people, million, virus

Topic  6
results, maps, results maps, election, elections, primary, primary elections, georgia, california, results maps georgia

Topic  7
need, know, need know, day, end, end day, need know end, know end day, know end, quotation

Topic  8
york, new york, city, new, new york city, york city, times, york times, new york times, recent

Topic  9
appeared, corrections, print, corrections appeared print, corrections appeared, appeared print, wednesday, jan, friday, appear
ed print wednesday

Topic  10
year, old, year old, end, time, percent, students, million, second, members
```