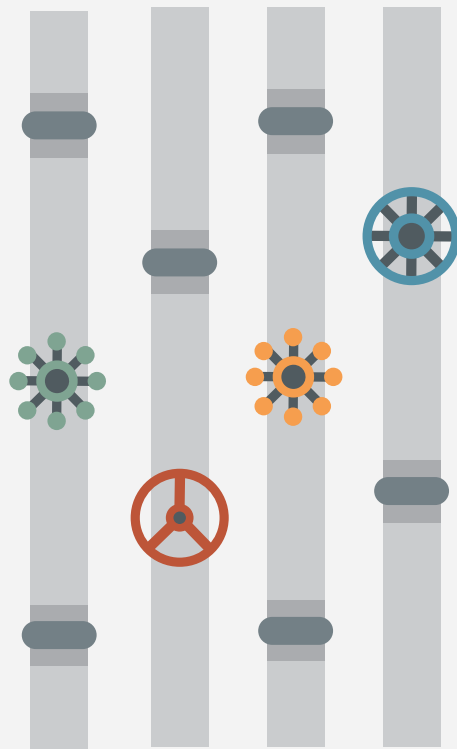


New York Times Topic Modeling End-to-end Data Pipeline Design

Jenica Andersen

Metis DSML, Data Engineering

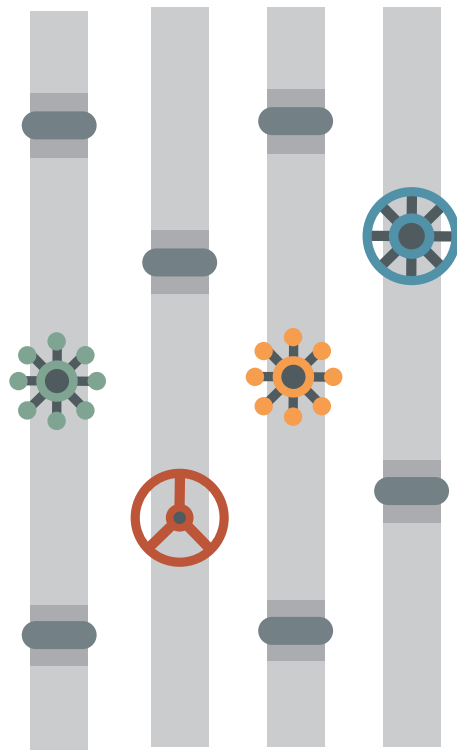
August 10, 2022



Introduction

Standard Workflow:

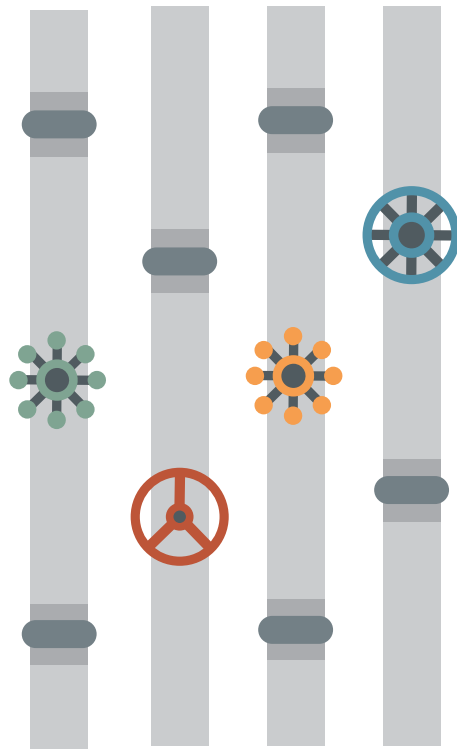
- Create or scrape data
- Pandas DataFrame
- Matplotlib Plots
- Static Slideshow



Problems

Standard Workflow:

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

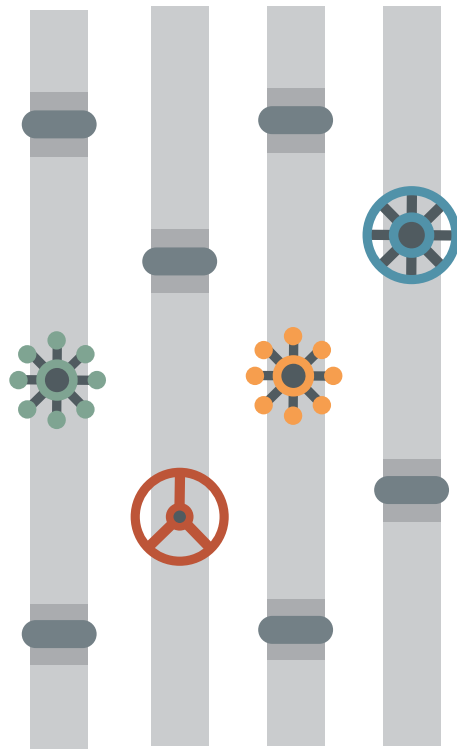


Problems

Standard Workflow:

time consuming, limited data

- Pandas DataFrame
- Matplotlib Plots
- Static Slideshow



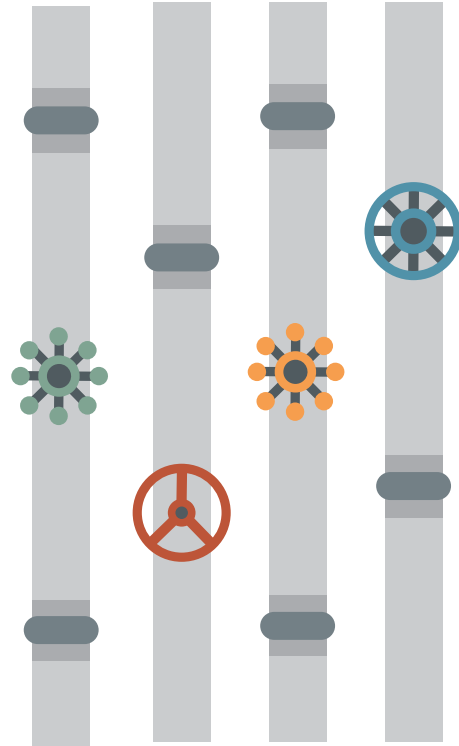
Problems

Standard Workflow:

time consuming, limited data

computationally expensive

- Matplotlib Plots
- Static Slideshow



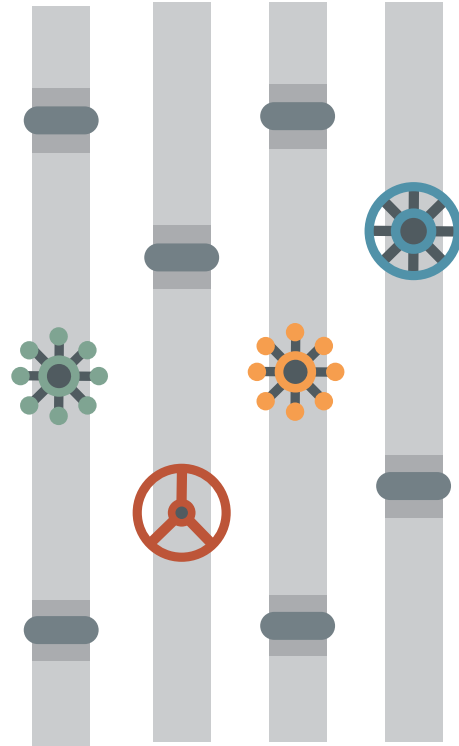
Problems

Standard Workflow:

time consuming, limited data

computationally expensive

- Static Graphics
- Static Slideshow



Problems

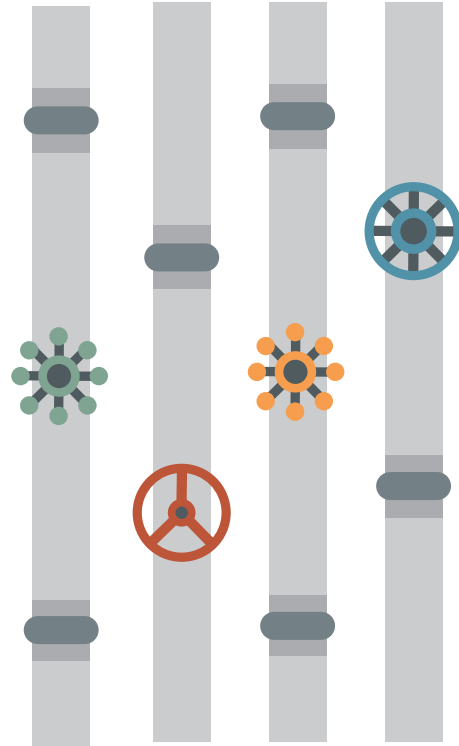
Standard Workflow:

time consuming, limited data

computationally expensive

- ~~ML~~ Static Graphics

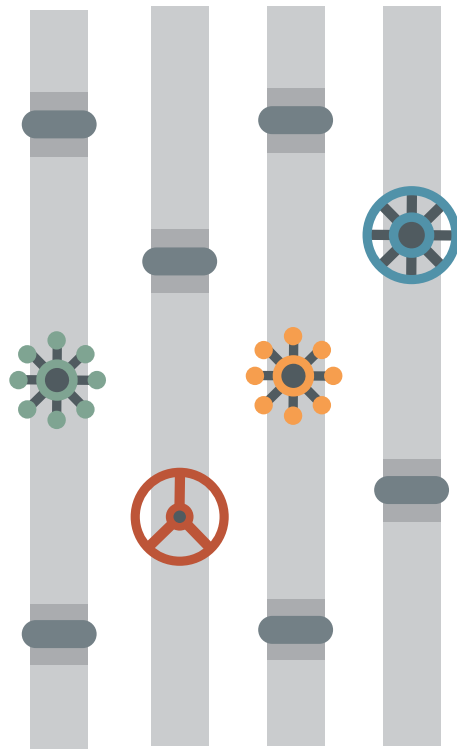
Passive deliverables



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



API-Application Programming Interface

Process



Manage data.
Implement Topic Modeling (NMF)

Application

 Streamlit



Interactive results

New End-to-End Data Pipeline

Data

{T} Developers



So much more data
available this way

Process

SQL

Amazon EC2

Amazon S3

pandas

Application



Streamlit

2020/01/01

End date

2020/03/08

100

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

[illegible]

New End-to-End Data Pipeline

Data

{T} Developers

So much more data available this way

Process

SQL



Scalable solutions

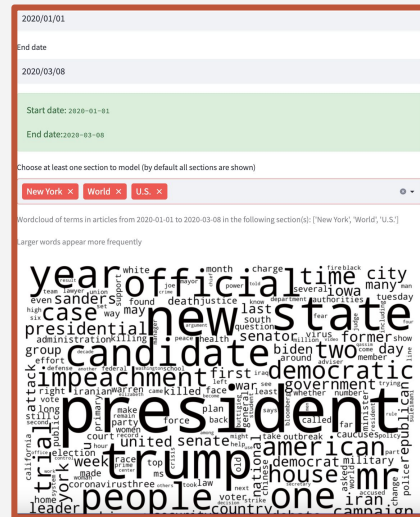
Amazon EC2 Amazon S3



Application



Streamlit



New End-to-End Data Pipeline

Data

{T} Developers

So much more data
available this way



Process



Scalable solutions

Amazon EC2 Amazon S3

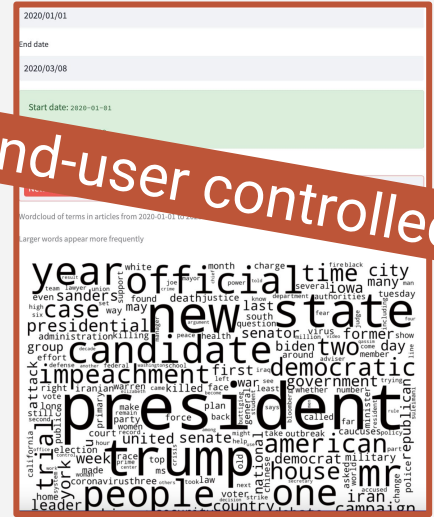
pandas

Application



Streamlit

End-user controlled



The Data: New York Times Archive API

{E} Developers

HomeAPIs

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()

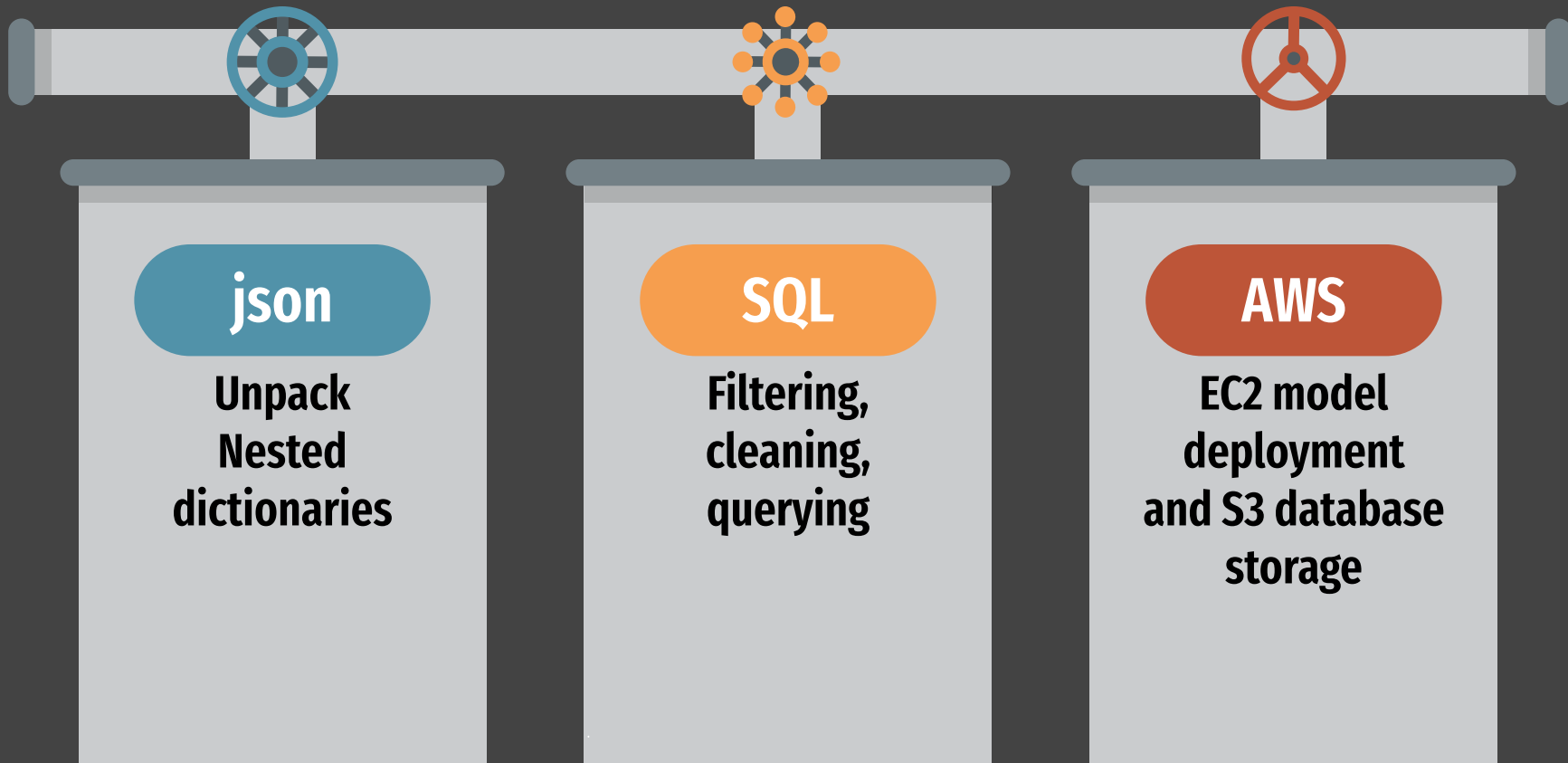
[92]: {'copyright': 'Copyright (c) 2020 The New York Times Company', 'response': {'docs': [{'abstract': 'The Biden administration has taken up a corrupt, hopelessly divided Afghanistan at 19 years and counting, the U.S.-led effort forming that it is time for U.S. troops to be left behind to carry out counterterrorism operations.', '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-columnists-on-the-biden-administration.html', 'height': 399, 'width': 600}], 'lead_paragraph': 'For years, the United States has been bogged down by wars in Iraq and Afghanistan, and now the Biden administration has taken up a corrupt, hopelessly divided Afghanistan at 19 years and counting, the U.S.-led effort forming that it is time for U.S. troops to be left behind to carry out counterterrorism operations.'}, {'abstract': 'The Biden administration has taken up a corrupt, hopelessly divided Afghanistan at 19 years and counting, the U.S.-led effort forming that it is time for U.S. troops to be left behind to carry out counterterrorism operations.', '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-columnists-on-the-biden-administration.html', 'height': 399, 'width': 600}], 'lead_paragraph': 'For years, the United States has been bogged down by wars in Iraq and Afghanistan, and now the Biden administration has taken up a corrupt, hopelessly divided Afghanistan at 19 years and counting, the U.S.-led effort forming that it is time for U.S. troops to be left behind to carry out counterterrorism operations.'}]}

- 10 years of article
- Raw data: ~3.8 GB
- Data of Interest: 577 MB
 - Date Published
 - Headline
 - Snippet
 - URL

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

```
[2]: {'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 been 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-articleLarge.jpg',
                                                         'medium': 'images/2020/12/01/opinion/30Afghanistan/merlin_180196935_e962e21c-4d7e-4ffc-a389-5c92262a319a-articleMedium.jpg',
                                                         'small': 'images/2020/12/01/opinion/30Afghanistan/merlin_180196935_e962e21c-4d7e-4ffc-a389-5c92262a319a-articleSmall.jpg'}]}]}
```

Processing





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 & EC@

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

1

End-user word entry

Add more powerful app features like links to ULR

3

MongoDB

Streamline json unpacking

5

2

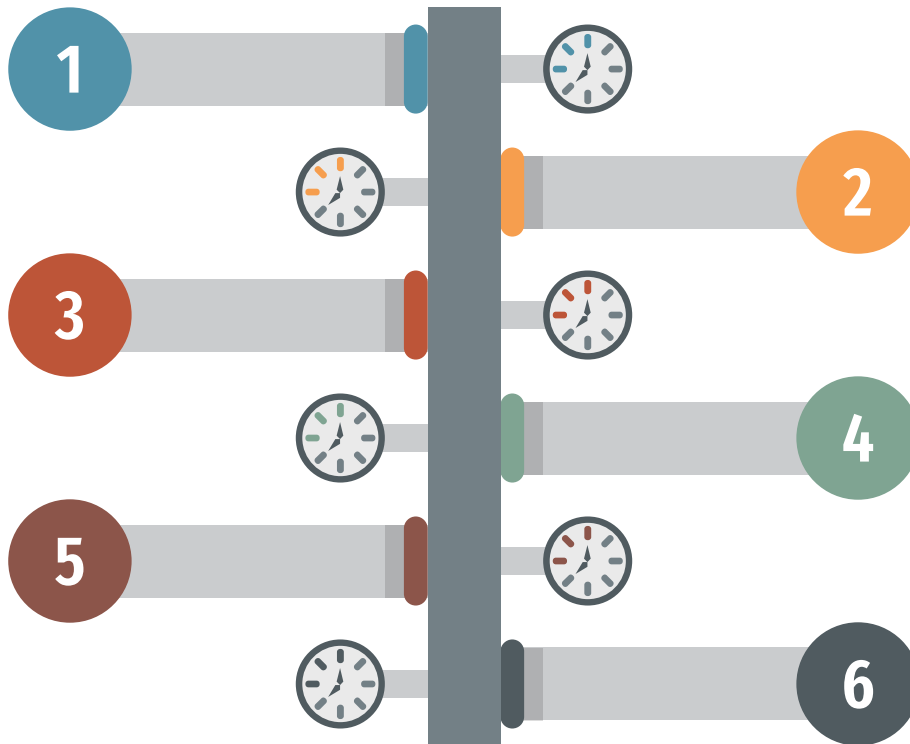
NMF
Move results from log to application interface

4

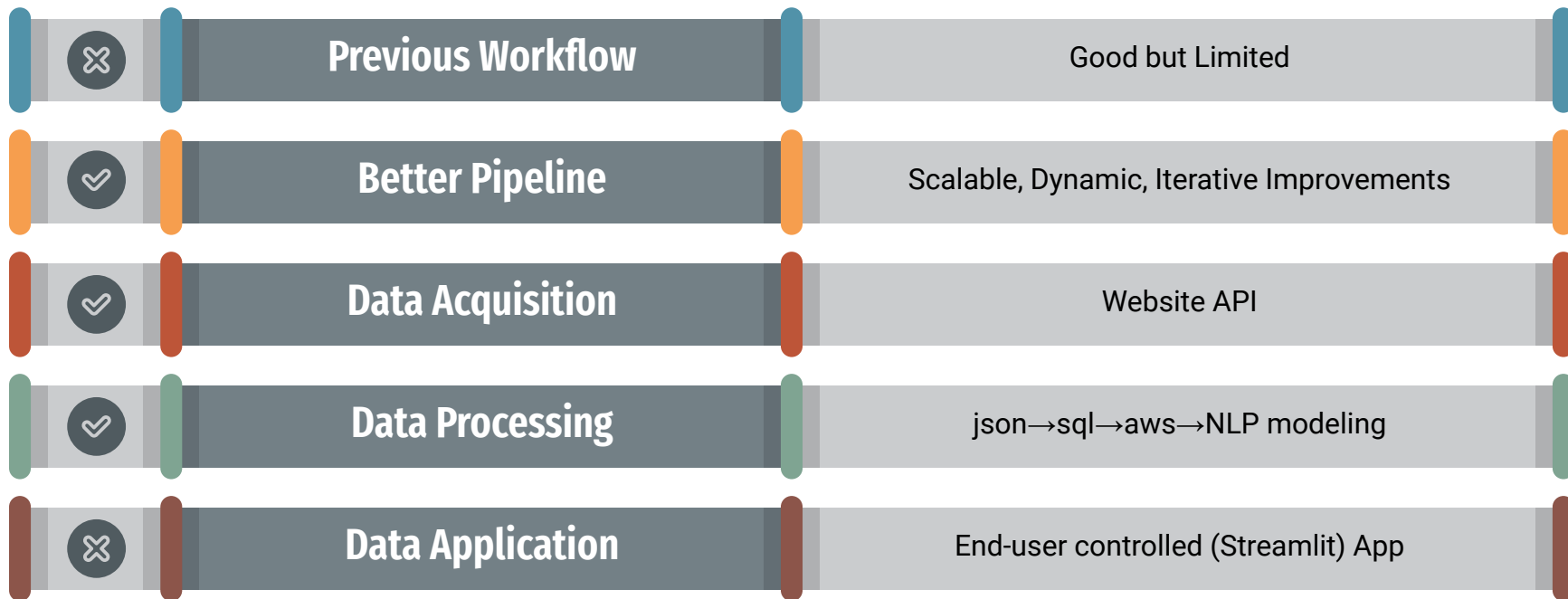
Clean "keywords"
And other fields, like "headline", to be usable in app or model

6

Spark
Leverage cluster computing for quicker modeling



Data Pipeline Summary



Thank you!



WordCloud of Jan 2020 Technology Article Snippets

