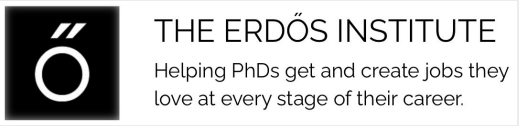
Topic recognition on New York Times articles

Ravi Tripathi, Touseef Haider, Ping Wan, Schinella D'Souza, Alessandro Malusà, Craig Franze



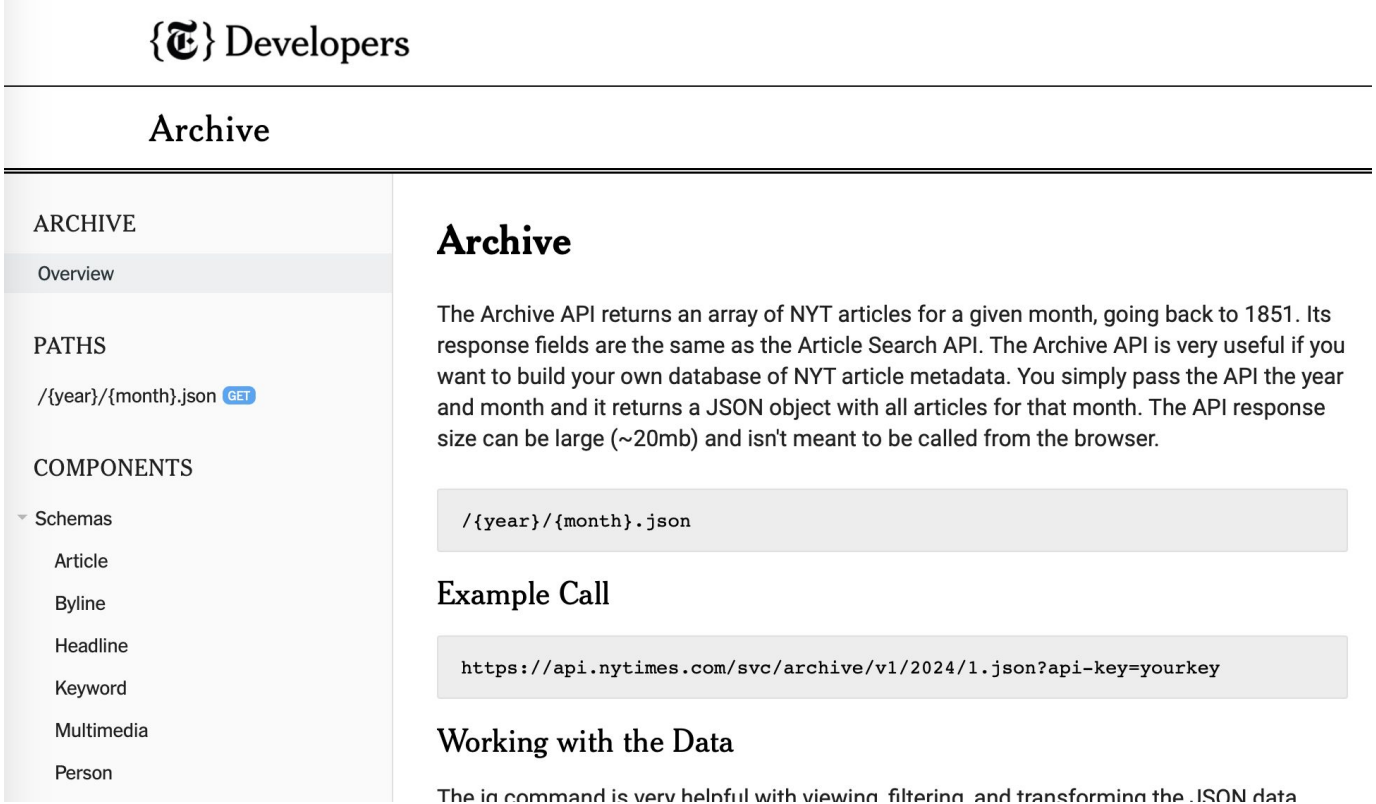
Goals

- Identify topics from a collection of articles
- Assign each article one or more topics, weighted by relevance

Applications:

- Study topic trends
- Build a recommendation engine

Data


{T} Developers
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 API response size can be large (~20mb) and isn't meant to be called from the browser.
/{year}/{month}.json
Example Call
https://api.nytimes.com/svc/archive/v1/2024/1.json?api-key=yourkey
Working with the Data
The jq command is very helpful with viewing, filtering, and transforming the JSON data

Data

```
import pandas as pd
df = pd.read_csv('data/nyt_metadata.csv')
df.head(10)
```

[1] ✓ 2.2s Python

abstract	web_url	snippet	lead_paragraph	print_section	print_page	source	multimedia	headlin
Economic hardship, climate change, political i...	https://www.nytimes.com/2023/05/14/us/migrants...	Economic hardship, climate change, political i...	Relative quiet has prevailed along the souther...	A	14.0	The New York Times	[{'rank': 0, 'subtype': 'xlarge', 'caption': N...	{'main': 'Titl 42 Is Gone but Not th Condit.
It's election night in America. Stay away from...	https://www.nytimes.com/2023/05/14/arts/televi...	It's election night in America. Stay away from...	The day before Logan Roy died, he delivered a ...	NaN	NaN	The New York Times	[{'rank': 0, 'subtype': 'xlarge', 'caption': N...	{'main ''Succession Season 4 Episode Rec.
Tom is stressed in dress shoes, Shiv hides ben...	https://www.nytimes.com/2023/05/14/style/succe...	Tom is stressed in dress shoes, Shiv hides ben...	This article contains spoilers for Episode 8 o...	NaN	NaN	The New York Times	[{'rank': 0, 'subtype': 'xlarge', 'caption': N...	{'main ''Succession Style, Episod 8: Some .
No corrections appeared in print on Monday, Ma...	https://www.nytimes.com/2023/05/14/pageoneplus...	No corrections appeared in print on Monday, Ma...	Errors are corrected during the press run when...	NaN	NaN	The New York Times	[]	{'main': 'N Corrections May 15, 2023 'kick.
Quotation of the Day for Monday, May 15, 2023.	https://www.nytimes.com/2023/05/14/pageoneplus...	Quotation of the Day for Monday, May 15, 2023.	"For me, it was time to give back the love the...	A	2.0	The New York Times	[]	{'main 'Quotation c the Day When You Cham.
The 19-year-old French basketball star is the ...	https://www.nytimes.com/2023/05/15/sports/bask...	The 19-year-old French basketball star is the ...	Boris Diaw was passing through Paris in late S...	D	1.0	The New York Times	[{'rank': 0, 'subtype': 'xlarge', 'caption': N...	{'main 'Everybod Wants Victo Wembanyama H.

Data

```
import pandas as pd
df = pd.read_csv('data/nyt_metadata.csv')
df.head(10)
```

[1] ✓ 2.2s Python

```
df[['abstract', 'lead_paragraph']].head(10)
```

✓ 0.0s Python

	abstract	lead_paragraph
0	Economic hardship, climate change, political i...	Relative quiet has prevailed along the souther...
1	It’s election night in America. Stay away from...	The day before Logan Roy died, he delivered a ...
2	Tom is stressed in dress shoes, Shiv hides ben...	This article contains spoilers for Episode 8 o...
3	No corrections appeared in print on Monday, Ma...	Errors are corrected during the press run when...
4	Quotation of the Day for Monday, May 15, 2023.	“For me, it was time to give back the love the...
5	The 19-year-old French basketball star is the ...	Boris Diaw was passing through Paris in late S...
6	New York City students are struggling with rea...	Good morning. It’s Monday. We’ll look at somet...
7	Results of Turkey’s election.	Turkey’s presidential election appears to be d...
8	Shouldn’t a protest movement led by unions be ...	For three months, France has been in revolt: D...
9	A spy drama based on a decades-long manhunt co...	Between network, cable and streaming, the mode...

The 19-year-old French basketball star is the ...	https://www.nytimes.com/2023/05/15/sports/bask...	The 19-year-old French basketball star is the ...	Boris Diaw was passing through Paris in late S...	D	1.0	The New York Times	[{'rank': 0, 'subtype': 'xlarge', 'caption': N...	{'main 'Everybod Wants Victo Wembanyama H.

Data Cleaning

- Abstracts were converted to lowercase
- Punctuation was stripped
- Acronyms were kept in
  - Ex: U.S. -> us
- Words in abstracts were put into lists
- Common phrases were kept together
  - Ex: climate_change
- Dates kept in

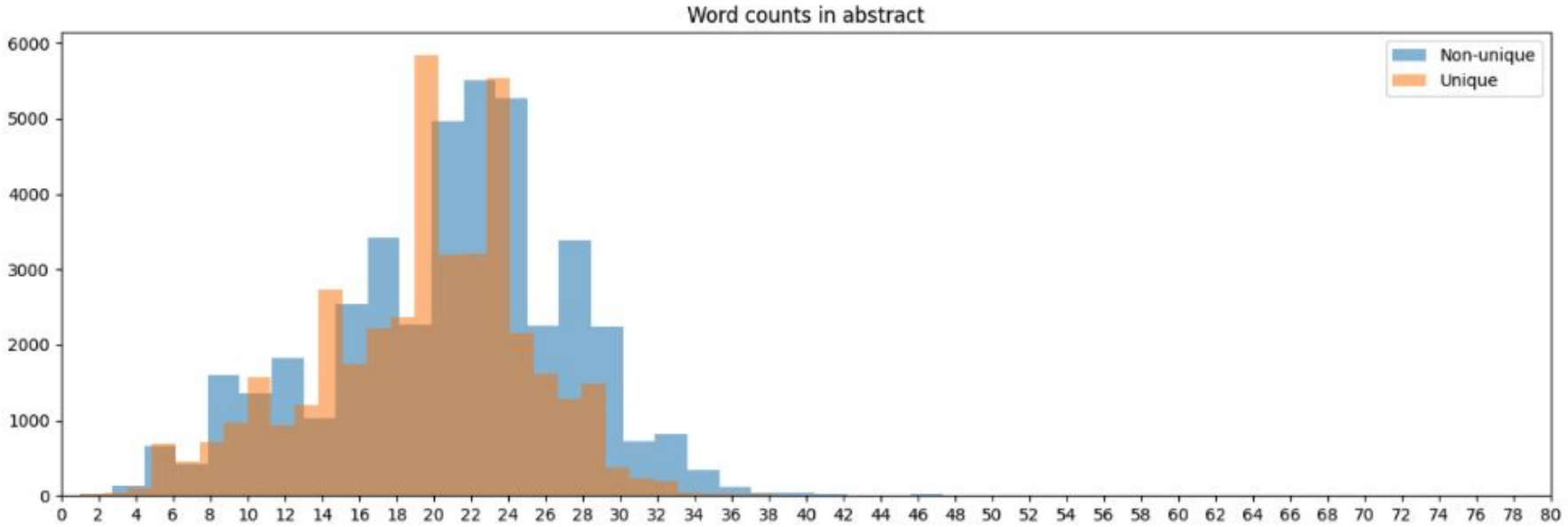
Cleaned Data

```
df = pd.read_csv('data/nyt_metadata_cleaned.csv')
df[['abstract', 'lead_paragraph']].head(10)
```

✓ 2.1s Python

	abstract	lead_paragraph
0	economic hardship climate change political ins...	relative quiet has prevailed along the souther...
1	election night america stay away from the bode...	the day before logan roy died delivered fiery ...
2	tom stressed dress shoes shiv hides beneath la...	this article contains spoilers for episode the...
3	corrections appeared print monday may 2023	errors are corrected during the press run when...
4	the year old french basketball star the most h...	boris diaw was passing through paris late sept...
5	new york city students are struggling with rea...	good morning monday we'll look something funda...
6	results turkey election	turkey presidential election appears destined ...
7	shouldn protest movement led unions benefiting...	for three months france has been revolt demons...
8	spy drama based decades long manhunt comes sho...	between network cable and streaming the modern...
9	vice which had wooed media giants has struggle...	vice media filed for bankruptcy monday punctua...

Data Exploration



Latent Dirichlet Allocation (LDA) Model

- LDA takes in a corpus of documents and generates representative topics
- Our corpus consists of NYT abstracts
- For each topic, the LDA model generates the top keywords

Optimal Topics

In [10]:

```
# Define Search Param
search_params = {'n_components': [10, 15, 20, 25, 30,35,40], 'learning_decay': [.5, .7, .9]}

# Init the Model
lda = LatentDirichletAllocation()

# Init Grid Search Class
model = GridSearchCV(lda, param_grid=search_params)

# Do the Grid Search
model.fit(X_count)
```

Out[10]:

```
GridSearchCV(estimator=LatentDirichletAllocation(),
             param_grid={'learning_decay': [0.5, 0.7, 0.9],
                         'n_components': [10, 15, 20, 25, 30, 35, 40]})
```

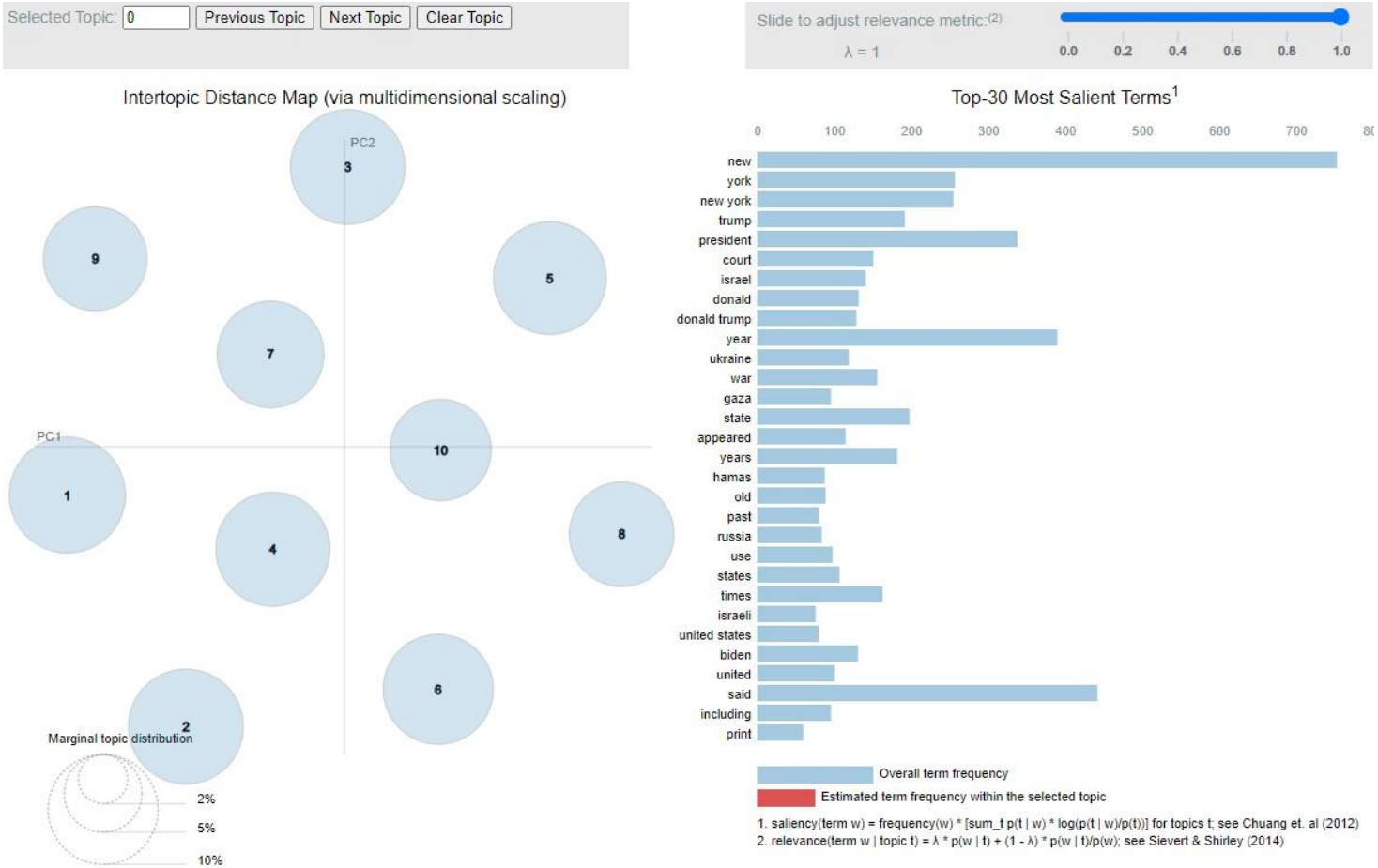
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [12]:

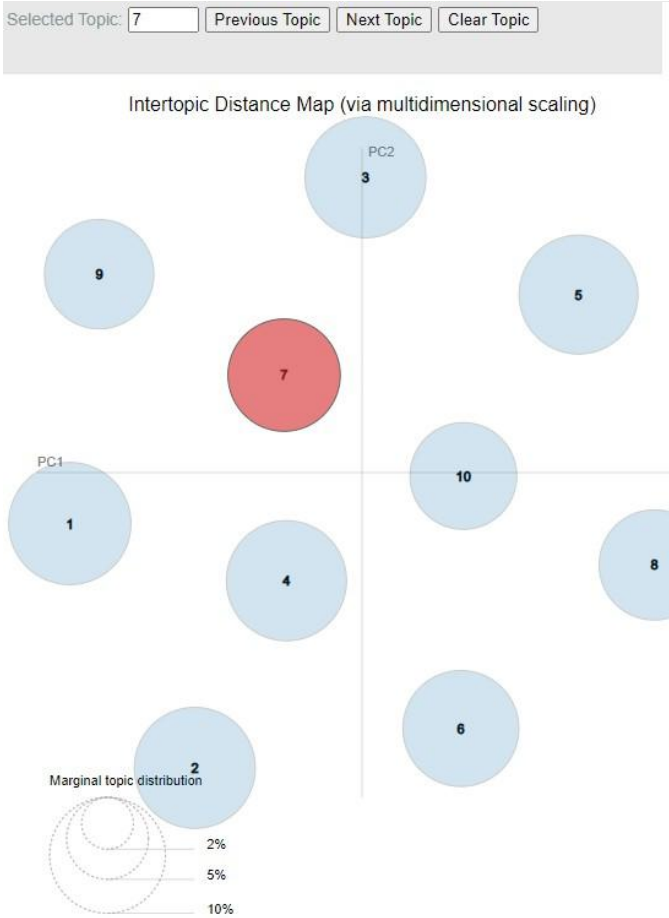
```
# Best Model
best_lda_model = model.best_estimator_

# Model Parameters
print("Best Model's Params: ", model.best_params_)

# Log Likelihood Score
print("Best Log Likelihood Score: ", model.best_score_)

# Perplexity
print("Model Perplexity: ", best_lda_model.perplexity(X_count))
```

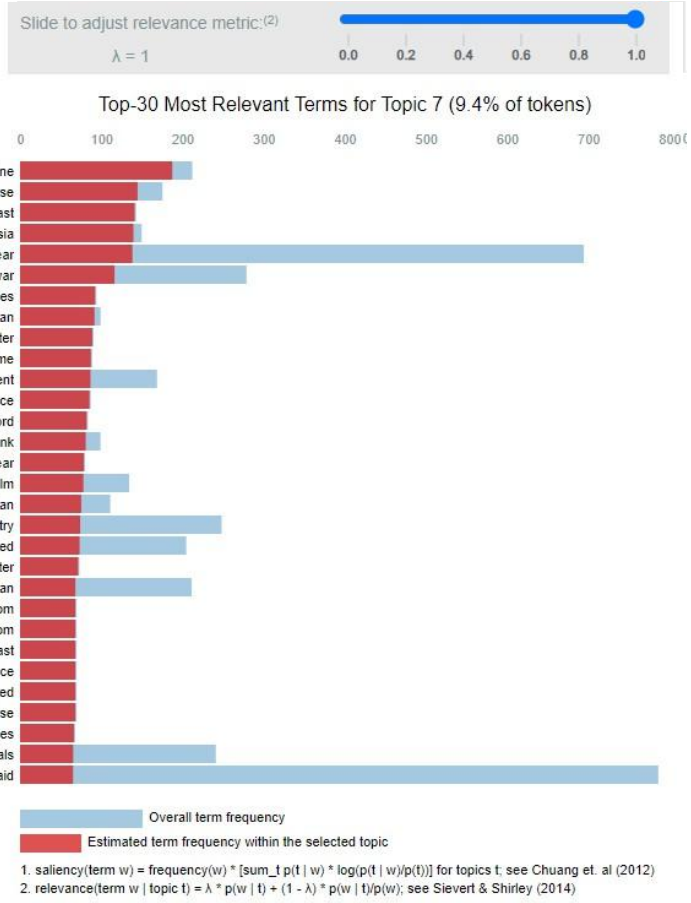
```
Best Model's Params:  {'learning_decay': 0.5, 'n_components': 10}
Best Log Likelihood Score:  -87000.10962360006
Model Perplexity:  1624.4050019802207
```

LDA Model

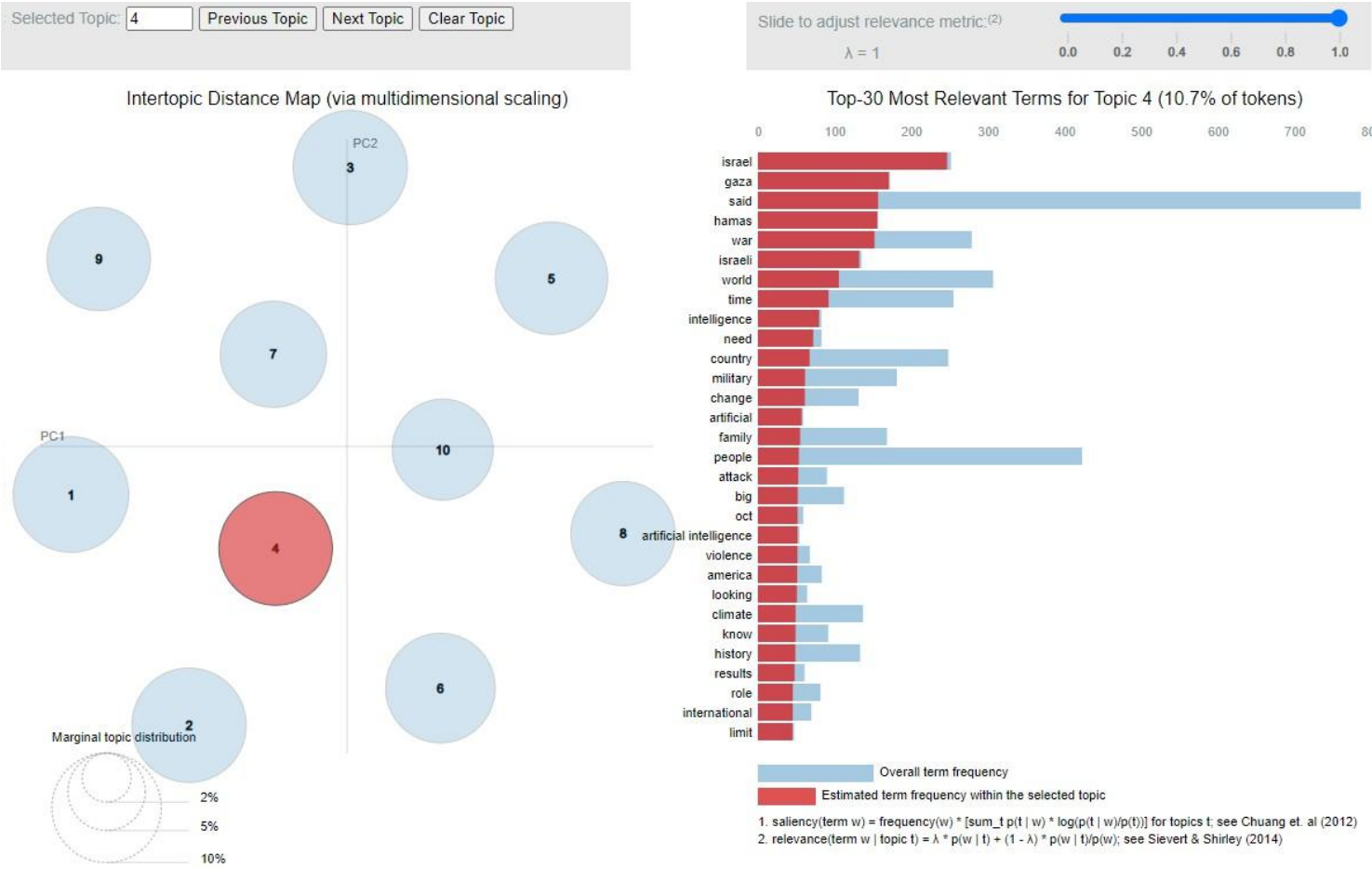

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0 100 200 300 400 500 600 700
new
york
new york
trump
president
court
israel
donald
donald trump
year
ukraine
war
gaza
state
appeared
years
hamas
old
past
russia
use
states
times
israeli
united states
biden
united
said
including
print
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

LDA Model


Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%


Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Top-30 Most Relevant Terms for Topic 7 (9.4% of tokens)
0
100
200
300
400
500
600
700
ukraine
use
past
russia
year
war
articles
ukrainian
minister
prime
government
sentence
word
bank
past year
film
russian
country
appeared
prime minister
american
articles nytimescom
nytimescom
nytimescom past
use sentence
word appeared
year use
appeared articles
officials
said
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

LDA Model


Selected Topic: 4
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 4 (10.7% of tokens)
0
100
200
300
400
500
600
700
israel
gaza
said
hamas
war
israeli
world
time
intelligence
need
country
military
change
artificial
family
people
attack
big
oct
artificial intelligence
violence
america
looking
climate
know
history
results
role
international
limit
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Recommendation System

- LDA model returns a vector of probabilities for each topic

Recommendation System

- LDA model returns a vector of probabilities for each topic

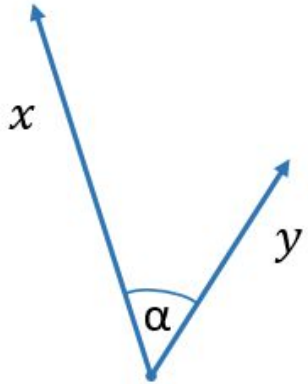
	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	dominant_topic
Article0	0.010000	0.010000	0.010000	0.010000	0.100000	0.010000	**0.240000**	**0.250000**	0.010000	**0.350000**	**9**
Article1	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	**0.740000**	0.010000	**0.160000**	0.010000	**6**
Article2	0.030000	0.030000	0.030000	0.030000	0.030000	0.030000	0.030000	**0.770000**	0.030000	0.030000	**7**
Article3	0.010000	**0.110000**	0.010000	**0.550000**	0.010000	0.010000	**0.260000**	0.010000	0.010000	0.010000	**3**
Article4	0.030000	0.030000	0.030000	0.030000	0.030000	0.030000	0.030000	0.030000	**0.700000**	0.030000	**8**
Article5	**0.670000**	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	**0.230000**	0.010000	0.010000	0
Article6	0.010000	0.010000	0.010000	0.010000	**0.870000**	0.010000	0.010000	0.010000	0.010000	0.010000	**4**
Article7	**0.230000**	0.010000	0.010000	**0.470000**	0.010000	0.010000	0.010000	**0.210000**	0.010000	0.010000	**3**
Article8	0.010000	0.010000	0.010000	0.010000	**0.940000**	0.010000	0.010000	0.010000	0.010000	0.010000	**4**
Article9	0.010000	**0.410000**	0.010000	0.010000	0.010000	0.010000	0.010000	**0.490000**	0.010000	0.010000	**7**

Recommendation System

- LDA model returns a vector of probabilities for each topic
- Cosine similarity is the cosine of the angle between the vectors

Recommendation System

- LDA model returns a vector of probabilities for each topic
- Cosine similarity is the cosine of the angle between the vectors



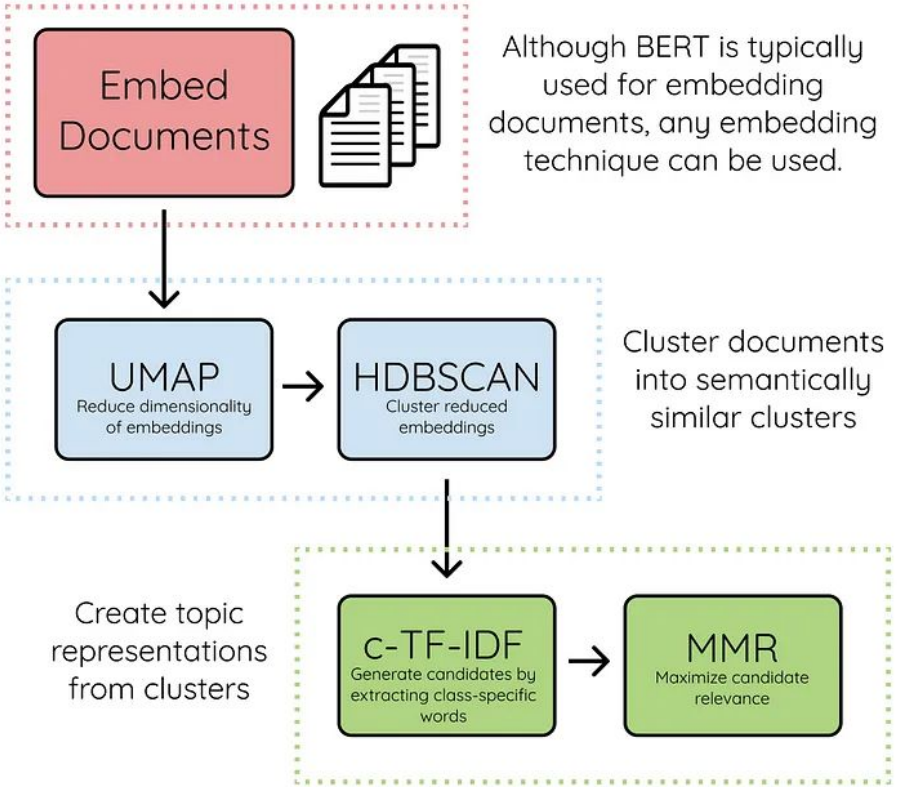
$$\cos(\alpha) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

Recommendation System

- LDA model returns a vector of probabilities for each topic
- Cosine similarity is the cosine of the angle between the vectors
- Recommendation system based on LDA model

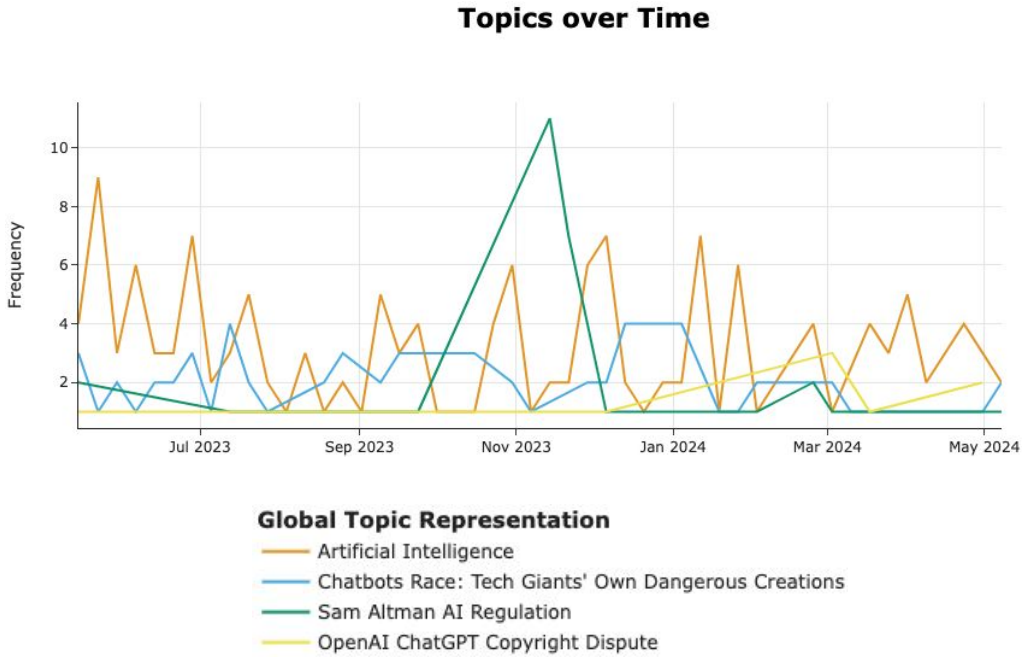
BERTopic

- Embeds docs using transformers to (default is ‘all-MiniLM-L6-v2’ which outputs 384 dimensional vectors)
- Reduces dimensionality using UMAP and clusters with HDBSCAN
- Identified over 400 emerging topics over the course of the past year
- Ollama was used to give sensible labels to these topics



BERTopic

- Supports dynamic topic modeling to understand the popularity of an identified topic over time
- Provides built-in search functions to go from topics to documents that could be useful for a recommender



Future Directions

- Enhance Model Performance with Hyperparameter Tuning
- Include additional data (lead paragraph, keywords etc) into the analysis
- Analyze Temporal Trends
- Integrate User interaction Data
- …and more!

Acknowledgements

- The Erdős Institute
- Our mentor Matthew Graham
- The New York Times

THE ERDŐS INSTITUTE

Helping PhDs get and create jobs they love at every stage of their career.