Extra! Extra! Read All Articles About It!

Topic Modeling and Clustering for News Aggregation

By Matt Fein and Polly McKim

News

NEWS
MONDAY
November 26, 2012
Daily informational digest
264/37
Today
28 °F / -2°C 56% Fair
BUSINESS • POLITICS • ECONOMICS • SCIENCE • ENGINEERING • MEDICINE • INDUSTRY • ECOLOGY • TECHNOLOGY
ANALYTICS
TOP NEWS
TECHNOLOGY
FINANCIAL REPORTING NEWS
GLOBAL ECOLOGY: IT'S NOT TIME TO FALL BACK
MOVE YOUR BUSINESS TO CLOUD
Enter your address here
Search
64GB
New technologies are coming soon
Medical research aspects
METAL SWITCHING

Now more than ever, people rely on news to remain informed about important world issues.

However, multiple sources cover the same topic...

- *Can we build a model to tell when this occurs?*
- *Why is this important?*

Topic Modeling + Clustering News Topics!

By using topic modeling we can work to cluster news articles about the same event from different sources

Use Cases

Politics

Politicians could easily track public opinions of constituents, enabling them to modify messaging or voting behaviors

Business

Companies could quickly synthesize information about their own business, competitor, or events that impact strategy

Government Intelligence

Rapid aggregation of all current news about an event could replace manual gathering; articles about the same event that differ widely could also indicate misinformation

Use Cases

Combat Media Bias/Misinformation

Aggregating all news articles available about an event gives a more robust picture of an event; ideally aggregating all articles about an event would work towards counteracting the media biased silos that currently exist

Goals for this Project

- Obtain a dataset of current news from large media sources
- Use topic modeling and clustering to group articles about the same event together
- Obtain interpretable results
- Explore results

Next we will discuss our methodology and results

Sources Used

Guardian API

- British news source articles
- Fields used:
  - Article Title
  - Body of article
  - Publication date

NYT API

- American news source Articles
- Fields used:
  - Article Title
  - Leading Paragraph
  - Snippet
  - Publication date

Preprocessing After Obtaining Dataset

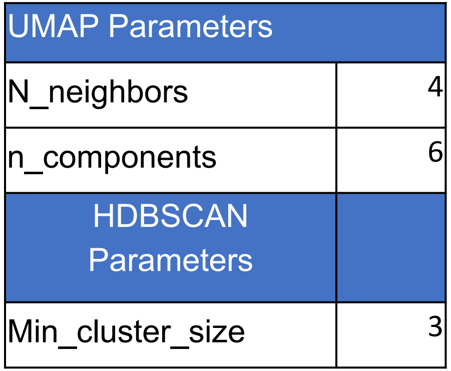
- Converted Json data into a single dataframe
- Cleaned dataset and removed extremely short news articles (ie. articles that only said “Election Updates Streaming Live”) & removed “titleless” articles
- Lots of data cleaning!
- Final Data set had 3073 entries and covered 10/1/2020-12/5/2020

Decision to use Bert and HBSCAN/ Processing

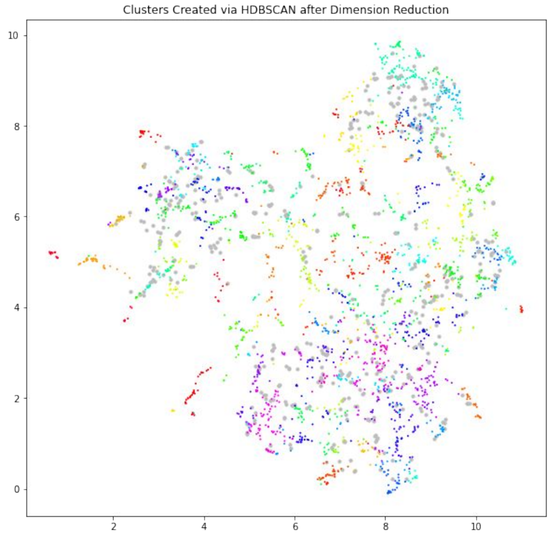
- BERT for topic modeling and HDBSCAN for clustering
- BERT is a good at recognizing context
- HDBSCAN does not force outliers into clusters
- Converted the documents to numerical data at the article level with “Sentence-Transformers” package/ distilbert-base-nli-mean-tokens sentence transformer
- Used UMAP dimension reduction to simplify the embeddings to facilitate making clusters
- Then used two rounds of TF-IDF to reduce and refine topics

BERT & HBDSCAN Results (Whole Dataset)

- Using the parameters pictured below returned the following results (after running the data through topic reduction



UMAP Parameters	
N_neighbors	4
n_components	6
HDBSCAN Parameters	
Min_cluster_size	3



Clusters Created via HDBSCAN after Dimension Reduction

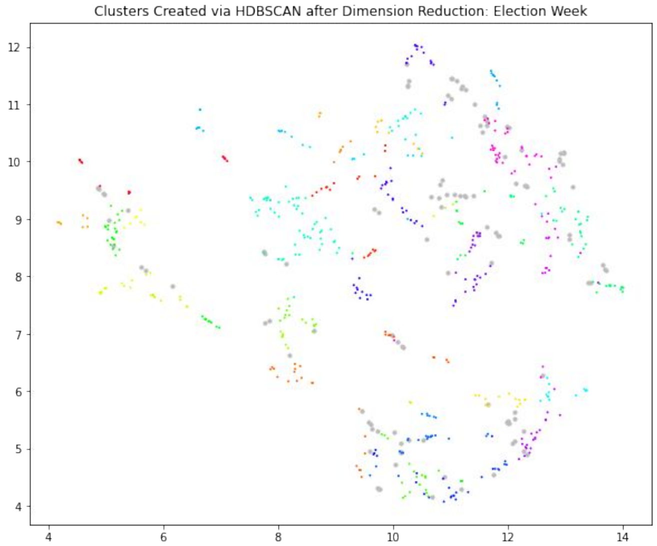
Top Clusters Found for Entire Dataset

- Pandemic
- Hurricane coverage
- Wildfire coverage
- Austrian terror attacks
- Turkish earthquake
- Election
- Vaccine news

BERT & HDBSCAN Results (Election Week)

- Changed parameters to get most frequent stories/ topics

UMAP Parameters	
N_neighbors	3
n_components	5
HDBSCAN Parameters	
Min_cluster_size	5


Clusters Created via HDBSCAN after Dimension Reduction: Election Week

Interesting Findings During Election Week According to the Model

- Michael Gove & his childcare policy gaffe
- Other top clusters were about the pandemic and of course… the elections- both in US and Burma

Missteps Taken and Lessons Learned

Underestimating time needed to acquire dataset:

APIs were much more complicated/ time intensive to work with than originally anticipated & often lacked the info needed

Underestimating time needed to clean dataset:

The dataset retrieved from the API was very messy and data irregularities negatively impacted our initial attempts at topic modeling

Using Packages that are newer/not debugged/hard to configure: Originally tried Top2Vec for topic modeling and clustering; it was a nightmare to get it to run/integrate; Also ran into problems with HSBM during installation

Possible Next Steps

- Further develop the model to make more accurate hyperlocal clusters
- Pair this model with sentiment analysis
- Pair model findings with similarity measurements to try to detect if an article contains inaccurate information

The possibilities are endless!!!!

THANKS!

Any questions?

github.com/pmckim1/NLP_News_Project

Polly McKim: pmckim1@gwu.edu

Matt Fein: mfein28@gmail.com