NewsClusters

Topic Modeling News and Clustering Reporters

Jake Stratton, Data Scientist

Background

- Individuals who work in communications and public relations, often need to find reporters to write about issues that are important to their organization.

- But how could they find these new authors without actually reading their stories?

- A machine learning application that would intake text, assign topics based on content, and allow a PR person to search for unknown reporters based on topics they write about would be a good solution.
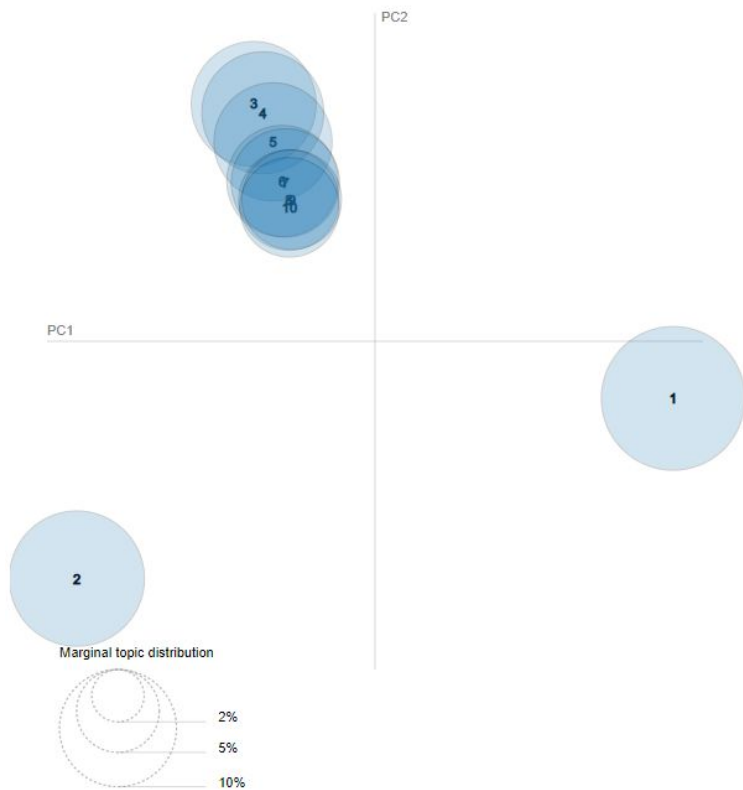
Technology Stack

Raw Data

- Used NY Times Archive API to collect metadata of the last 5 years of NY Times articles
- Data included author, headline, and the first sentence of the article.
- Total articles - 246,603
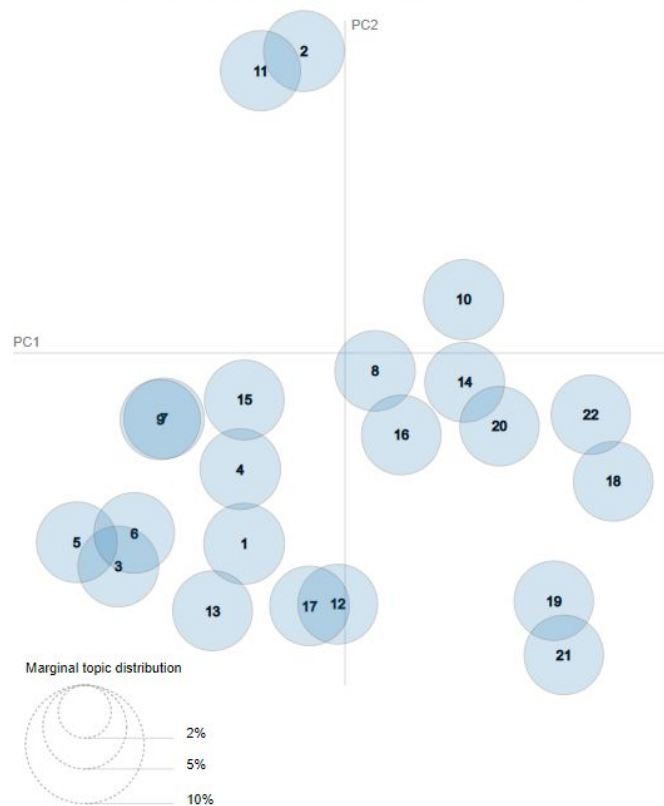- Authors - 13,842

Cleaned Data

- Words in corpus - 5,963,685
- Total documents - 197,603
- Authors - 12,703
- Most commonly used words:
  - york
  - trump
  - president

Gensim LDA, 10 Topics

Mallet LDA, 22 Topics, No Adjectives or Adverbs

Coherence Score per Number of Topics
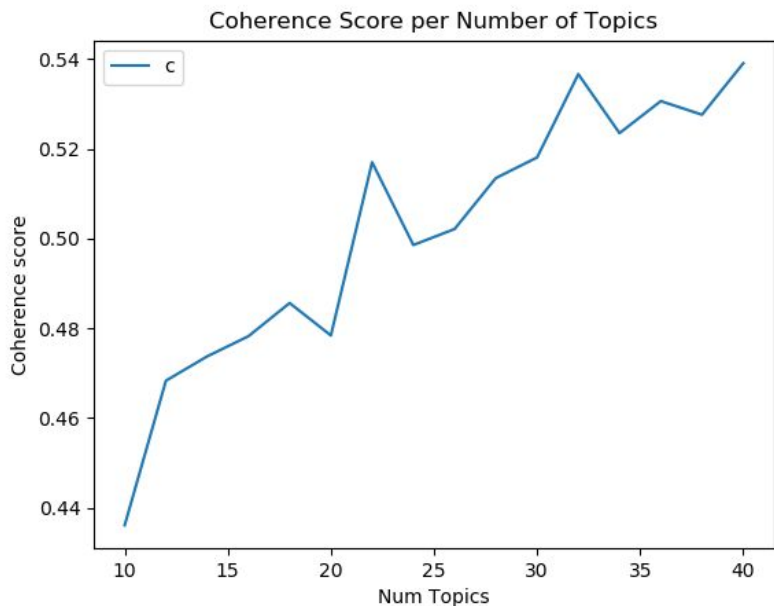
NY Local - york, city, brooklyn, manhattan, water, home, park, resident...
Crime - man, kill, police, attack, charge, death, fire, accuse, shoot...
Politics & Elections - house, campaign, election, vote, donald_trump...
Foreign Affairs - china, power, move, face, trade, russia, france...
War and Conflict - country, fight, war, force, europe, fear, battle, syria...
Health & Science - find, study, health, test, risk, drug...
News About News - time, talk, news, week, report, discuss...
Tech & Business - company, apple, amazon, facebook, investor...
Economy - pay, raise, market, bank, deal, sale, money...
Labor - make, open, job, line, spend, worker, growth, increase...
Travel - home, food, offer, summer, hotel, tour, travel, island, holiday...
Fashion & Art - show, art, museum, paris, designer, wear, collection...
Local Sports - run, win, yankees, ranger, giant, series, loss, strike, series...
National Sports - game, team, win, player, coach, world_cup, soccer...
World News - world, year, britain, leave, office, begin, scandal...
Education - school, program, student, question, college, university, child...
Music - music, year, career, record, album, jazz, singer, band, pop...
Government - call, leader, state, plan, senate, group, government...
Law & Judicial - case, rule, court, law, charge, judge, accuse...
Film & TV - film, review, series, movie, star, tv, show, night, host...
Theater - theater, broadway, role, festival, review, dance, stage...
Family & Social - woman, book, family, child, love, story, mother, son...

NY Times Article Database

Search for Reporters Using:

Reporter Name ▼

anthony tommasini

Search

| Author	Headline	Topic
anthony tommasini	review the met operas tosca overcomes months of chaos	Theater
anthony tommasini	review a nostalgia trip at the new york philharmonic	Theater
anthony tommasini	review risktaking new opera tells a tragic 1950s gay love story	Social
anthony tommasini	what a hologram of maria callas can teach us about opera	Theater
anthony tommasini	review concertgebouw orchestra shows off its legacy at carnegie	Theater

NY Times Article Database

Search for Reporters Using:

Reporter Name ▼

brian seibert

Search

Author	Headline	Topic
brian seibert	for michelle ellsworth practice makes more practice	Theater
brian seibert	on artists and audiences at american realness	Theater
brian seibert	review the bitter tears of dean moss co	Theater
brian seibert	a storytelling choreographer who listens to her audience	Theater
brian seibert	review on siri bodies and masks in he his own mythical beast	Theater

Top 100 Authors based on Total Articles

Distance
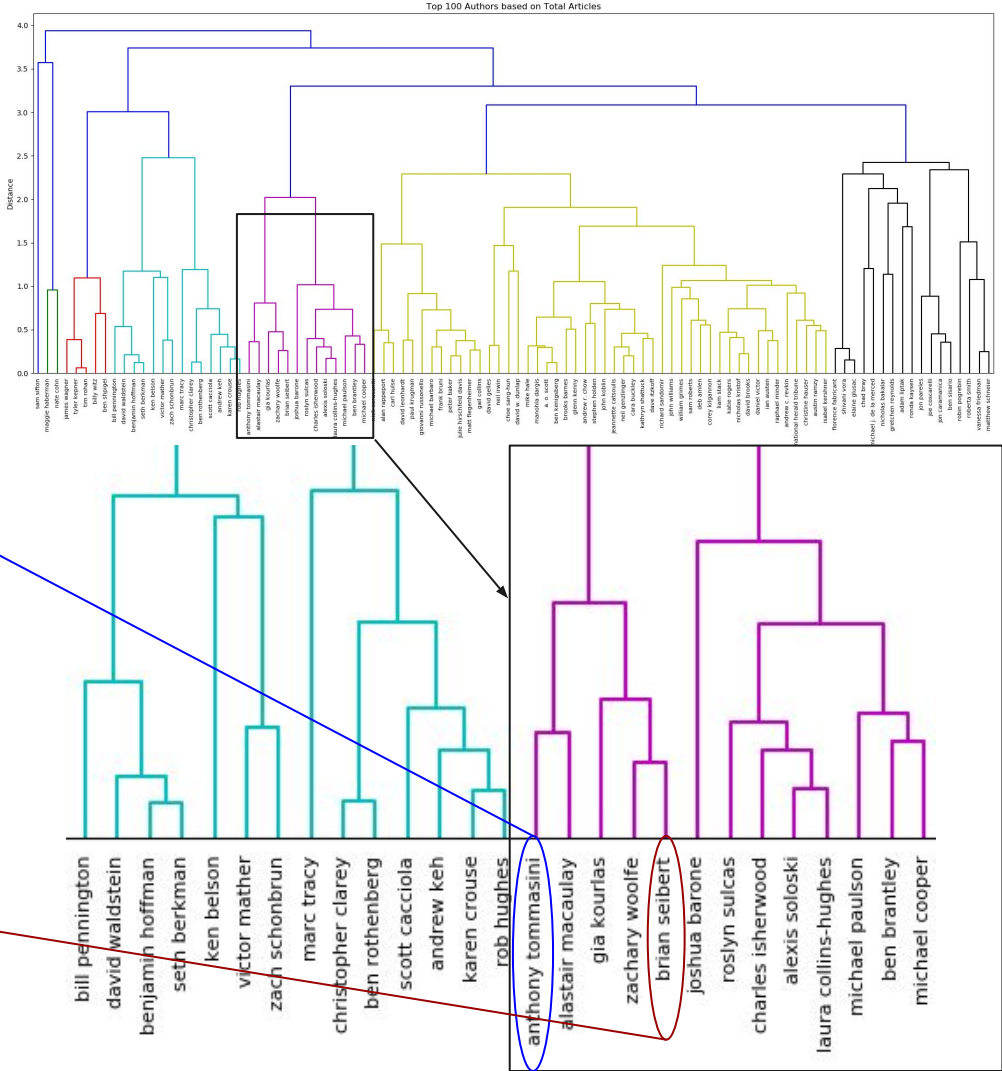
Next Steps

- Add author to author search functionality
- Add search by topic functionality
- Include multiple news outlets
- Clean up Flask interface

Contact

- Jake Stratton
- jacobcstratton@gmail.com
- https://github.com/JakeStratton
- https://www.linkedin.com/in/jacobcstratton/