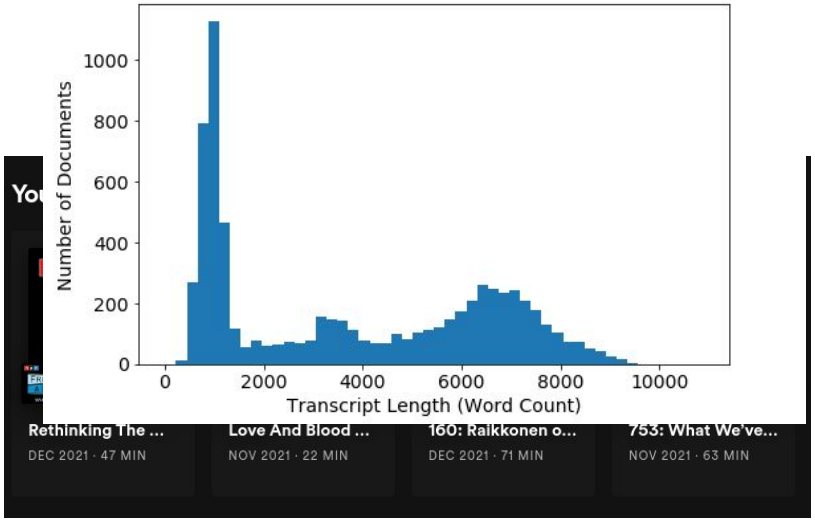
Content-based Fresh Air Episode Recommender

Jan 26, 2022
Metis Online Flex - NLP/Unsupervised
Julianne Skinner

- Motivation: Podcast episode suggestions based on content
- Data: Fresh Air transcripts ranging from 2009-2021
  - Interviews and reviews ranging from current events to arts and pop culture
  - Including some personal narratives
  - 7000 documents with an average of ~1300 tokens after processing



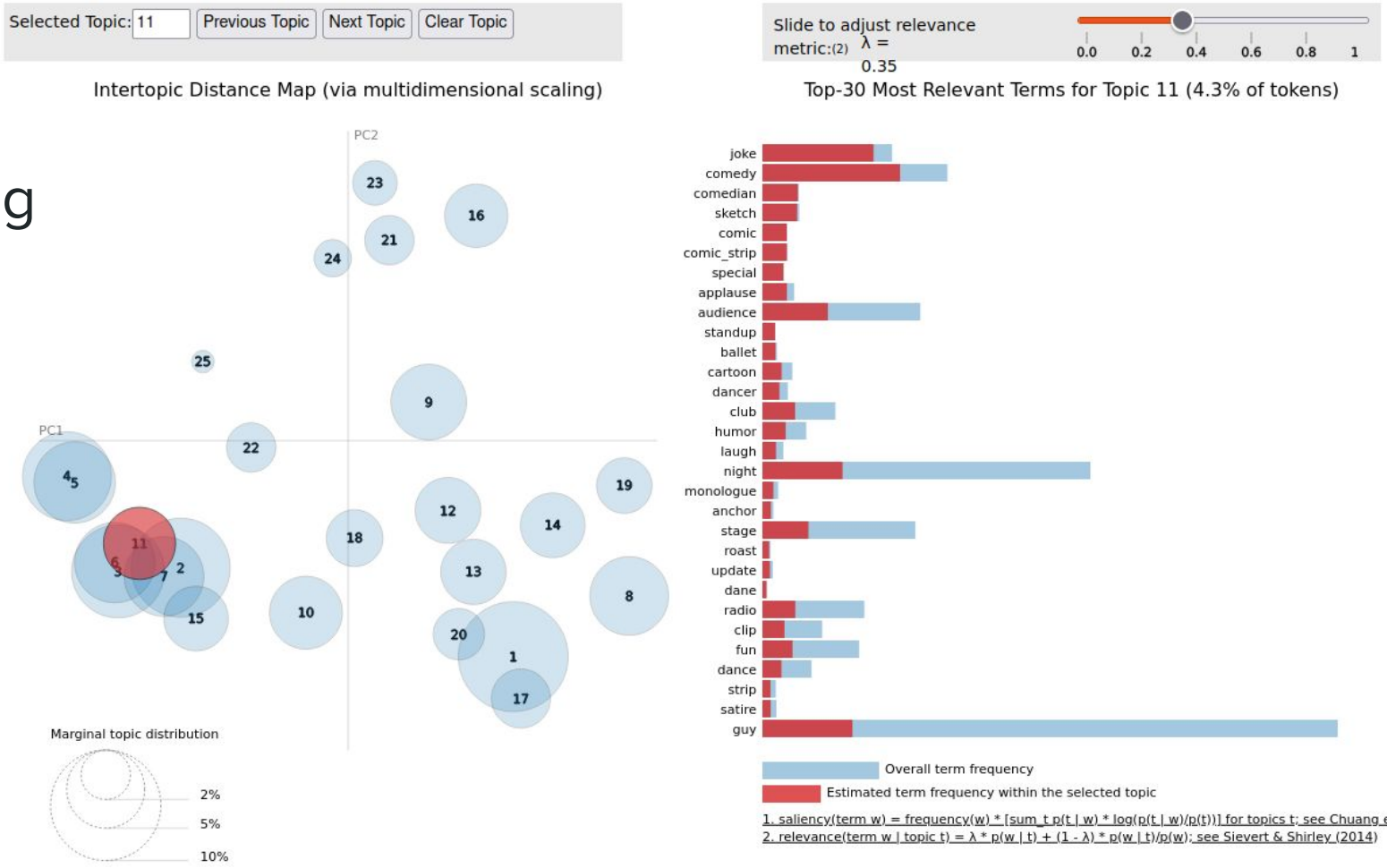
Methodology

- Webscraping
- Preprocessing
- Topic Modeling
  - LDA - assumes multiple topics per document
- Episode Recommender
  - Episode similarity based on topic probabilities per document
  - Other features: segment length and recency

Preprocessing

- SpaCy
  - lemmatization
  - stop word removal - amended with common phrases from the corpus
  - removal of nouns and named people
- CountVectorizer
  - Adjusting min_df and max_df
  - Ngrams (eventually discarded bigrams)

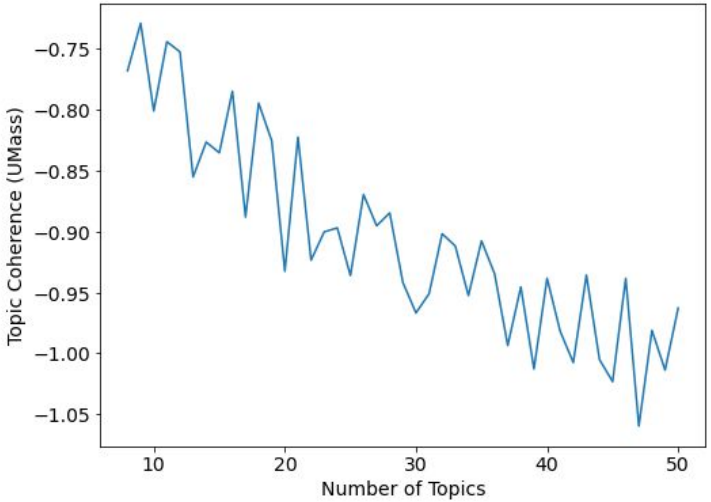
Topic Modeling



Topic Modeling

Top 8 words for the 10 most frequent topics:

0: film movie scene director character actor story documentary
1: company information computer internet book technology world system
2: book world church religion idea sense novel question
3: character series scene actor season episode role man
4: school student college class teacher education community kid
5: food book virus oil restaurant fish water climate
6: game guy player team ball baseball sport pitch
7: money state company business bank tax law election
8: song music singing album jazz band singer recording
9: woman right law court case child abortion decision



Episode Recommendations

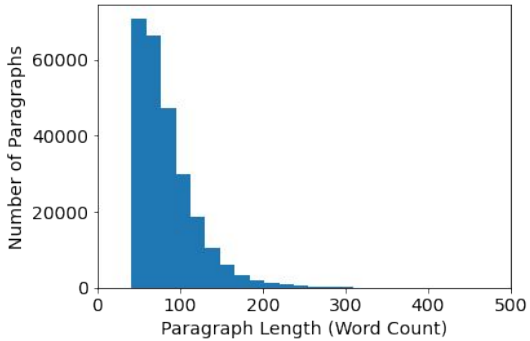
Input: < How a group of online sleuths are helping the FBI track down Jan. 6 rioters

Recommendations:

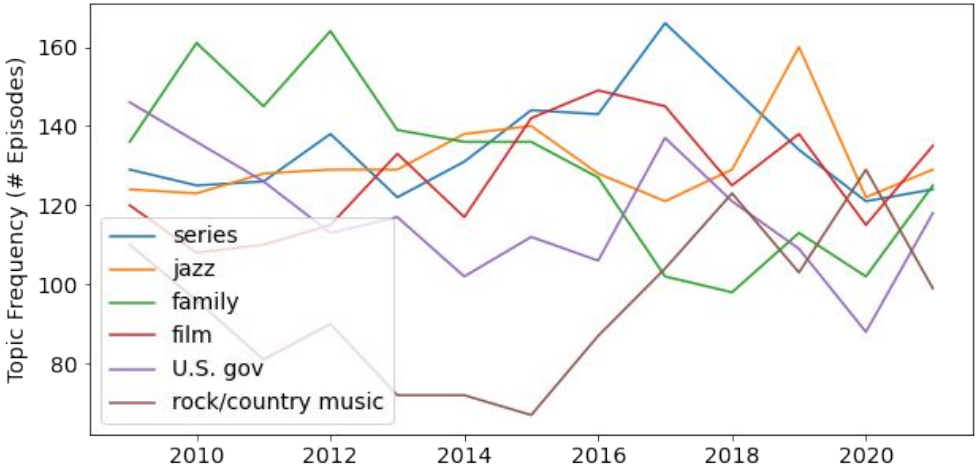
Date	Title
2021-01-27	< Surveillance And Local Police: How Technology Is Evolving Faster Than Regulation
2014-03-04	< By The Time Your Car Goes Driverless, You Won't Know The Difference
2019-02-27	< 'Consumer Reports' Director Offers An Inside Take On The Car Testing Process
2018-12-10	< The Revolution Will Be Driverless: Autonomous Cars Usher In Big Changes
2017-08-22	< FBI Profiler Says Linguistic Work Was Pivotal In Capture Of Unabomber

Conclusions / Future work

- Using the relevance parameter for the episode recommender
- Dynamic topic modeling (initial results were not coherent, and the process was time-consuming)
- Seed topics using CorEx
- Putting named entities (people's names) back in the corpus for more fine-tuned topics
- Splitting up documents into overlapping segments
    - Paragraphs > BERTopic



Appendix


Topic Frequency (# Episodes)
160
140
120
100
80
2010
2012
2014
2016
2018
2020
series
jazz
family
film
U.S. gov
rock/country music

Appendix

	Topic	Count	Name
0	-1	5023	-1_war_family_friend_love
1	88	368	88_mother_mom_wife_parent
2	6	277	6_author_poem_novel_memoir
3	60	257	60_tax_bank_insurance_debt
4	95	195	95_father_dad_son_brother
5	52	189	52_film_movie_screenplay_video
6	98	180	98_jazz_piano_pianist_musician
7	62	178	62_cancer_surgery_hospital_doctor
8	43	134	43_religion_church_faith_gospel
9	70	130	70_voter_vote_voting_ballot
10	49	122	49_comedy_joke_comedian_sketch
11	31	111	31_restaurant_chef_cooking_cookbook
12	37	105	37_child_kid_baby_parent
13	24	77	24_dog_animal_smell_puppy
14	71	76	71_internet_hacker_secret_conspiracy
15	3	70	3_prison_prisoner_jail_inmate