Conspiracy Theories Project Summary

Puns

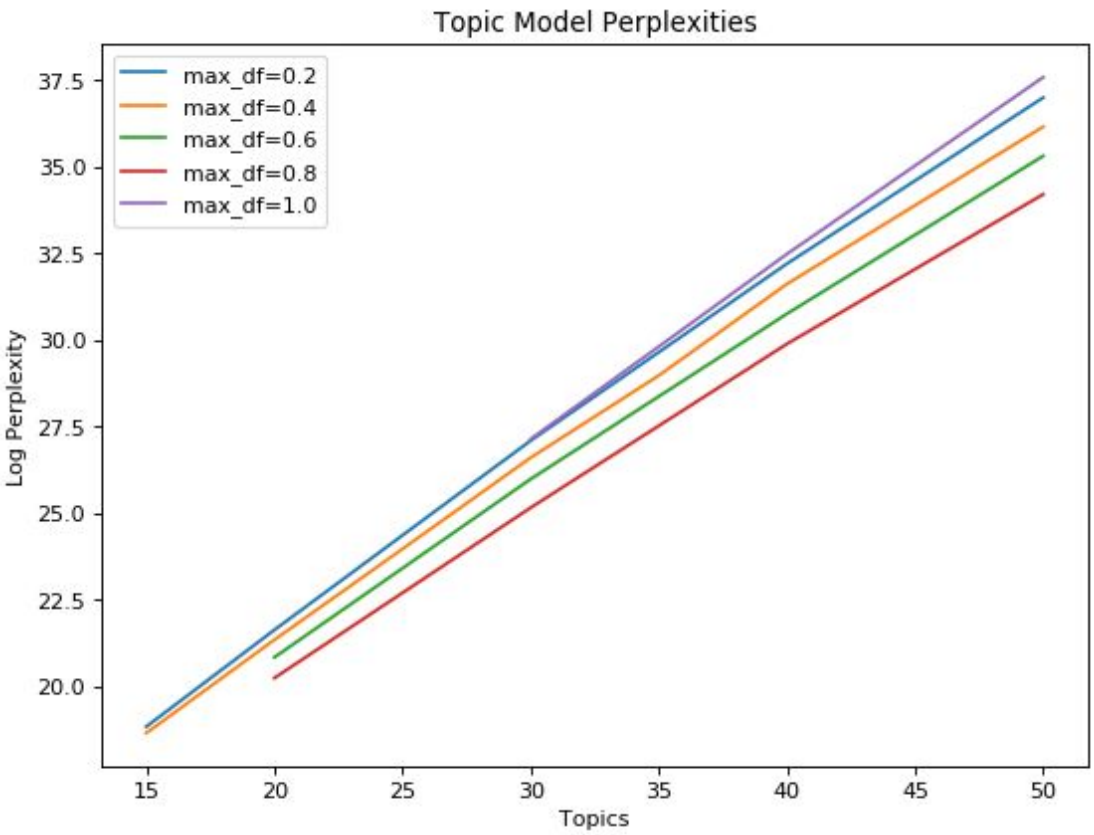
1. “Hopefully you’ll find this presentation illuminating.”
2. 

Scope

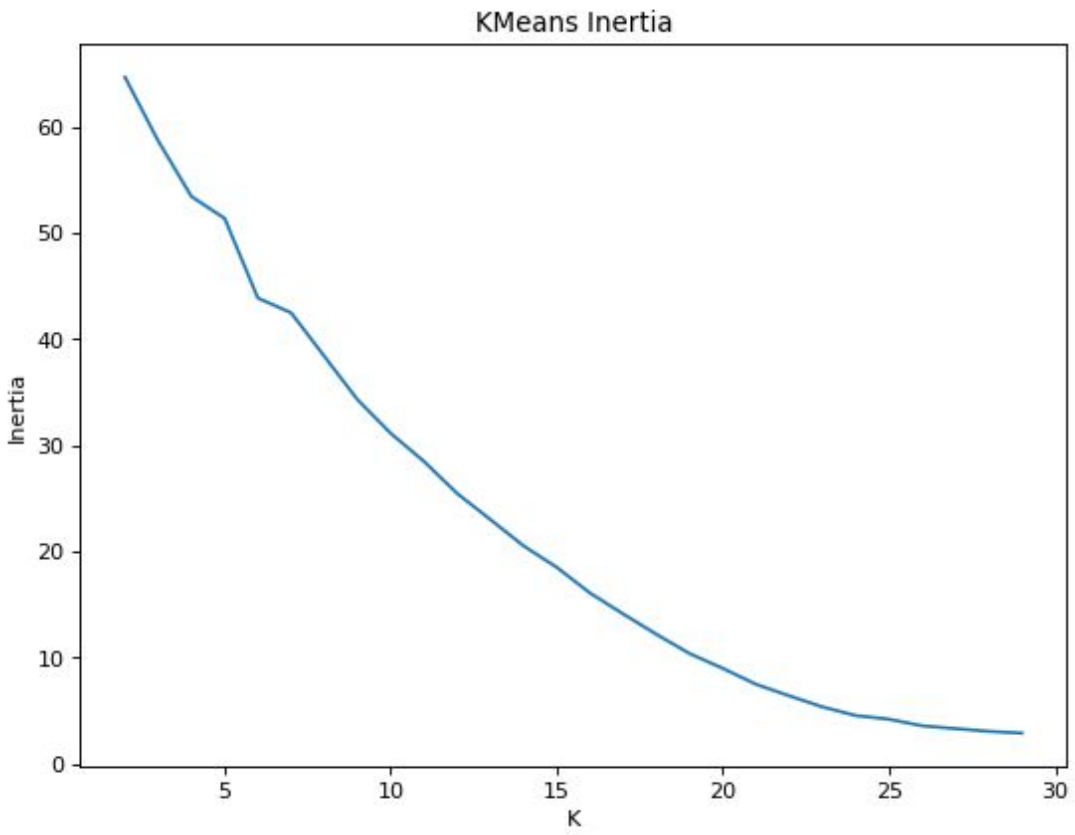
Discovering characteristics of different types of conspiracy theory texts.

Design

1. I built a bibliography from academic works on conspiracy theories and wikipedia of the most significant and influential conspiratorial works, and scoured the internet to find 175+ texts -- primarily books and a few articles. I then converted those pdfs to txt, and cleaned them.
2. NLTK vectorized. I chose not to stem/lemma because it’s not relevant for most conspiracy domain-terms; for style, stemming would have been counterproductive -- e.x. ‘couldn’t’ (negative certainty) is not the same as ‘could’ (positive uncertainty) and ‘n’t’ (negative unclear.
3. LDA topic modeling.
   The nature of conspiracy theories makes it particularly tricky to tell how connected topics are, how many topics there are, and how good topic models are. So I needed to test a broad range of max_df and num_topic combinations. I created a VectorWrap class and an LDAWrap class to produce a kind of hyperparameter grid search and keep my test models/results tidy. I produced multiple runs on the search using perplexity and diff as heuristics to help narrow down the search. But ultimately I still had to assess topic coherence for each model by hand, by looking up exactly how all these hundreds/thousands of obscure topic terms were being used in individual texts, which was extremely difficult and time-consuming. I ultimately settled on 25 topics, with max_df set to 0.8.



4. I clustered off my LDA topics. Both DBSCAN and KMeans both suggested roughly the same number of clusters as topics (19 and 24ish, respectively).



5. I created a cosine similarity text summarizer so that for each topic I could produce summaries of the closest-matching texts.

6. I also really wanted to topic model on paragraphs, as well as topic model on punched out domain terms to model style and epistemic style. While I didn't have time to run the modeling, I completed code for these parts so I'll be able to complete that modeling in the future.

Tools/Algorithms

- NLTK
- Gensim LDA
- Cosine Similarity
- DBSCAN, KMeans

Topic Results

<u>Coherent topics</u>

- Anti-Government
- Mind-Control
- New_Age
- NWO

- Noosphere
- Illuminati
- Reptilian_Masters
- Aliens
- Anthrax/Biothreats
- Islam
- Birthers
- Holocaust
- Govt_Internment

Surprising topics -- uncertain coherence

- Obama-JFK-Catholicism
- Antichrist/Catholicism
- I_AM_and_Mind_Control
- Illuminati and Hallucinogens?
- Pale_Horse (Alien+Multiple overlapping conspiracies *a la* Cooper's *Behold a Pale Horse)*

Probably incoherent topics

Six of them.

What I'd do differently next time…

- Not get sick.
- Be less picky about getting the exact right topic model, or subsample my data when I'm tuning topic modeling hyperparameters.