Problem

Temporal Evolution Topic Modeling via Non-Negative CP Tensor Decomposition

Topic modeling is a branch of Natural Language Processing (NLP) that is heavily researched, yet can be quite intractable at times. Quantitatively capturing figure of speech, sarcasm, sentiment, etc poses an extreme challenge for even the most controlled experiments. This challenge grows exponentially when accounting for foreign languages, bi-grams and tri-grams, numerical connotations, etc. Currently there are several methods to tackle the problem of topic modeling. The setback of the majority of these models is that they must be applied multiple times in order to extract topic evolution over a temporal domain. Given a corpus of text that spans days, months, years, would require an iterative approach to extract topic evolution using techniques such as LDA, Non Negative Matrix Factorization, and various clustering techniques. This iteration could lead to high variability in results as well as the interpretability of said results. Recently, analytical methods involving high order tensors and their decomposition have been gaining in popularity. This is due to their ability to extract unique latent variables of a large dataset simultaneously in multiple dimensions. Due to constraints naturally imposed on the modeled data as a result of the higher order tensor, the decomposed latent factors model important mixing information across all dimensions of the nth order tensor. Utilizing non negative CP Tensor Decomposition, we intend to extract the most important topics from 2.7 million news articles that spans over 4 years. We will demonstrate not only the speed and efficiency of this method (in contrast to iteratively applying single domain methods noted above), but its accuracy in determining the major topic from every month for each year. This will be done by determining the major topic for a given month from a given year, and cross referencing it with a manual search for that month and year.

Supplemental

CP Decomposition performs better than LDA which is a probabilistic Bayesian approach that relies on a well established prior

CP Decomposition performs better than Non negative matrix factorization because it does not suffer from the rotation problem. NMF has the possibility to produce non unique results during its decomposition. This affects interpretability of the results from NMF since they are not unique

CP Decomposition performs better than clustering techniques because interpretability of said techniques require manual analysis of each cluster for interpretability. Even then, there’s no guarantee that the “semantic gap” is small enough such that the latent information that formed each cluster is even human understandable.

Data

2.7 million news articles and essays from 27 American publications ranging from 2016 to 2020
https://components.one/datasets/all-the-news-2-news-articles-dataset/

2016 – 604,511 articles
2017 – 640,493 articles
2018 – 553,588 articles
2019 – 655,456 articles
2020 – 234,830 articles

138,547 Authors

```
Vox, Business Insider, Reuters, TMZ, Vice, Vice News,
Hyperallergic, TechCrunch, Axios, Refinery 29, The Verge,
Mashable, People, Economist, CNN, Gizmodo, New Yorker,
CNBC, Wired, New Republic, Fox News, The Hill,
Politico, The New York Times, Buzzfeed News,
Washington Post
```

Literature Review

Blei, et. al in 2003 proposed *Latent Dirichlet Allocation (LDA)* which is a “hierarchical Bayesian model”. It models each text in a corpus as a “finite” mixture over an underlying set of topics. Because it is a generative parametric model, it makes base assumptions in establishing a priori in order to produce a posteriori. A weak priori will result in an even weaker posteriori. It also fails to account for temporal topic mixing with respect to it is approach. LDA would have to be applied multiple times over a time domain to produce temporal evolution of topics. This is not efficient.

Lee, et. al in 2012 produced *iVisClustering: An interactive Visual Document Clustering via Topic Modeling*. Their motivations were to counteract the shortcomings of traditional clustering techniques with an interactive view of the topics to be analyzed by an expert with domain knowledge. They correctly pointed out that clustering for topic modeling poses issues due to the high probability of noisy features and outliers. Also, it is not uncommon for clusters to not make sense due to the “semantic gap”. That is, the gap between associating low level latent features with “high-level” human concepts. Because clustering could be utilized to extract temporal evolution of topics, it would be a potential choice in method, however due to the unconstrained nature of how clusters are formed, the semantic gap could widen, leading to results that are nonsensical. It would also prove difficult computationally to produce enough meaningful clusters throughout a large time span such that sufficient topics are represented yet the space isn’t over saturated with topics so that no topic truly dominates all others in high probability.

Shahnaz, et. al in 2004 provided their method for *Document clustering using non negative matrix factorization*. This method is very close to the method we utilize. It represents a set of documents as an m x n matrix where m is the number of terms and n is the number of documents. It then finds a low rank approximation by factoring the m x n matrix into the product of two reduced-dimensional matrices. Two problems with this method however prevent it from being the most viable method. 1.) It is a factorization method in only two dimensions. The reduced rank representations would be with respect to documents vs. topics and words vs topics. Notably, the temporal component is missing. This means, NMF would have to be ran multiple times across the time span. 2.) NMF is not immune to the rotation problem of matrix factorization. The decomposition of an m x n matrix need not be unique. Due to the production of non unique results, that makes the interpretability of the results impossible. Huang, et. al further explored this problem in their 2014 paper: *Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition*. In this paper, they outline several conditions of known unique results to matrix factorization. However there is no guarantee in topic modeling that these constraints can be met.

Method

- Preprocess 2.7 million news articles
    - Canonical format (lower case)
    - Remove stop words
    - Remove numbers
    - Remove extra spaces and new lines
    - Perform Lemmatization and/or Stemming
    - Vectorize text (TF-IDF)

- Aggregate Articles by month and year
    - Store results in SQL DB to offload processing work on the DB engine instead of in-memory processing

- Build functions to construct mode-3 tensors with dimensions (words x documents x months) for a specific year
    - N articles will be sampled from each month for a given year to constrain computational and memory complexity

- Perform Non-Negative CP Tensor Decomposition to extract low rank latent information.