Evolution of News Over Time

Atharva Mahesh Kavitkar
Department of Computer Science
Technische Universität Kaiserslautern
kavitkar@rhrk.uni-kl.de

Abstract

Topic Modelling is a useful technique to gather insights from a huge corpus of data. It is capable of organising the corpus in terms of topics that the documents belong to. But the temporal aspect of identified topics is not acknowledged in static topic modelling techniques. The evolution of a topic over time cannot be mapped using static topic modelling technique. In this project, a dynamic topic modelling technique is used based on the paper from Blei et. al. to analyse a dataset of New York Times news articles. This dataset spans from 1920 to 2020 and would be used to study the evolution and usage of topics over the 100 years.

1 Introduction

Over the years, we have seen rapid evolution of data storage technology. This has allowed us to store huge amounts of textual data for personal as well as professional applications. The current challenge is the analysis of such data and extract useful insights from it. NLP techniques such as topic modelling has emerged as an important method for analysis of such massive collection of documents(1). Apart from identifying themes from corpus, the topic modeling method can also be used to model the progress of the topics. In this report, a dynamic topic modeling approach is used to examine the evolution of topics in New York Times news articles over a span of 100 years.

The content of a news article might reveal the topics to which the article concentrates, especially the terms and terminologies being used. The most common topics in a single era might offer a sense of the trend in news at the time. If news topics are put into a time sequence, it could correlate with the corresponding world events happening at the time. It can be demonstrated how the news topics develop over time by using data from New York Times news articles. The remainder of this report is divided into 5 sections. LDA section describes the Latent Dirichlet Allocation (LDA) algorithm which is used as the foundation for the time dependent topic analysis model used in this project. Dynamic Topic Modelling section describes the method built on top of LDA for this project. Training section describes the algorithm used for training of the model. Results section displays the outputs obtained from the model. The report ends with a section for Conclusion and Future Work.

2 LDA: Latent Dirichlet Allocation

The method used in the project was built on top of Latent Dirichlet Allocation (LDA). The news articles are assumed to follow a generative probabilistic process with the topic of the article acting as one of the hidden variable.

LDA assumes that a news article is made up of multiple topics. For example, an article titled "Using Robots in Tennis" could be classified under "technology" and "sports". It is assumed that every word in each of the document belongs to already established set of topics(2). Let V the size of the vocabulary and K be a specific number of topics, α a [1xK] dimensional vector and η be a scalar. $D_V(\eta)$ denote a K dimensional Derichlit with scalar parameter η and $D_V(\alpha)$ denotes a V-dimensional Derichlit with vector parameter α.

Preprint. Under review.

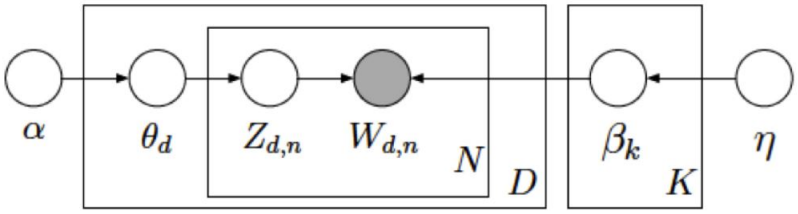

Figure 1: Graphical representation of LDA(2)

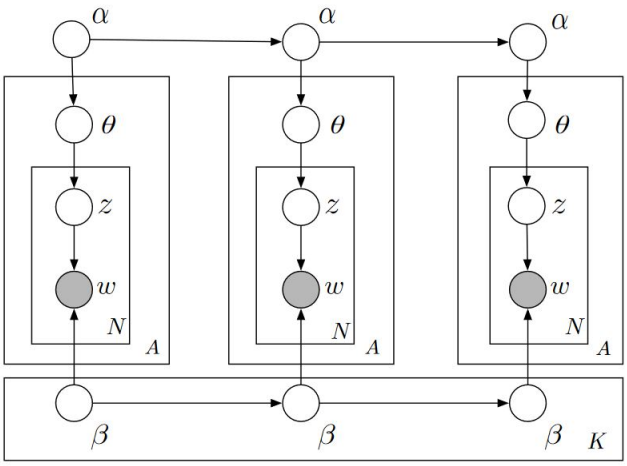

Figure 2: Graphical representation of Dynamic Topic Model(4)

LDA Algorithm(3)(2):

1. For each topic:
   (a) Draw a distribution over words $\beta_k \sim D_V(\eta)$
2. For each news article:
   (a) Draw a vector of topic proportions $\theta_k \sim D(\alpha)$
   (b) For each word:
      i. Draw a topic assignment $Z_{d,n} \sim Mult(\theta), Z_{d,n} \in 1, 2, ..., K$
      ii. Draw a word $W_{d,n} \sim Mult(\beta_{Z_{d,n}}), W_{d,n} \in 1, 2, ..., V$

LDA gives a combined distribution over the hidden and observed variables. The hidden topic decomposition of a certain corpus results from the corresponding distribution of the hidden variables in accordance with the number of documents observed(D)(2). This posterior distribution is our key to the joint distribution followed by the generative process. Fig.1 shows the relation between the random variables in LDA.

3 Dynamic Topic Modelling

The LDA model does not consider the temporal aspect of documents (news articles) provided to it. Apart from that it assumes that the same words will belong to a certain topic irrespective of the year(3). This assumption is not suitable in our use case because the news the articles span decades, during which the language of a topic itself are very likely to have changed. Moreover, events happened in an earlier time are likely to have an impact on subsequent events. Dynamic topic models (DTM) are capable of tracking the evolution of topics organised sequentially with respect to time(4). In Dynamic Topic Modelling, the data is devided into time slices, e.g., by year. The news articles of each slice are modelled with a K-component topic model, where the topics associated with slice t evolve from the

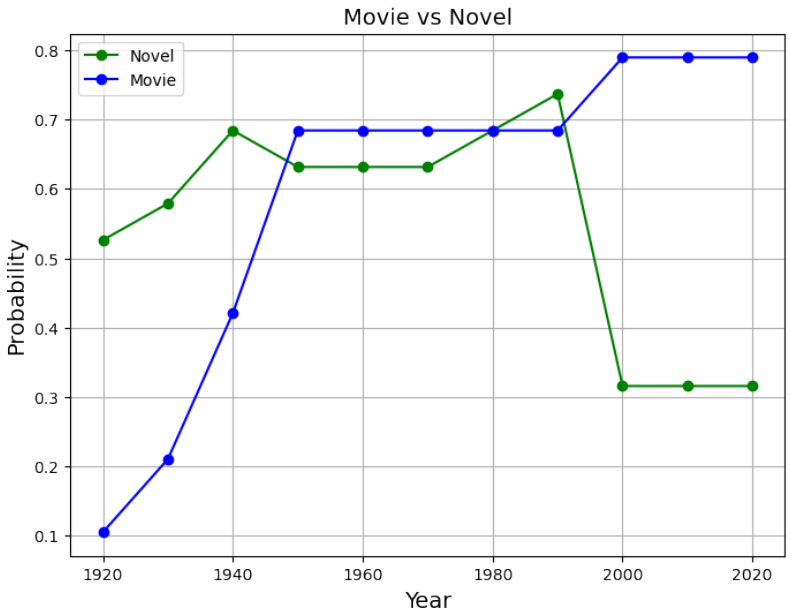

Figure 3: Movie vs. Novel

topics associated with slice t-1. The time-series topics are modeled by a logistic normal distribution of $\beta_{k,1} \rightarrow \beta_{k,2} \rightarrow ... \rightarrow \beta_{k,T}$ The posterior is approximated over the topic decomposition with variational methods explained in upcoming section.

Each subject is now a sequence of words distributed at the topic level. For each subject and year, then, the likelihood of words may be calculated and the complete subject can be viewed in the course of time with its top words(4). This offers a general idea of how the key words of a subject have evolved over the collection. We can analyze their score inside each subject for particular terms of interest. The total popularity of any issue may also be examined annually.

4 Training Algorithm

Approximating the posterior distribution is the key computing challenge for topic modeling using LDA. In this project, mean variable field inference was applied to do this. Variational inference is preferred over sample-based techniques(such as Gibbs sampling) because it can be faster(5). The core notion of variational inference is approximating an intractable posterior distribution over hidden variables with a simpler distribution. These parameters are then fitted such that the approximation is as near to the true posterior as possible. KL-divergence is used to find the difference between true posterior and the approximation. The hidden variables are initialised randomly and the parameters are updated till evidence lower bound is optimised.

5 Results

This section discusses about the results obtained from multiple runs of the dynamic topic modelling algorithm on the News Articles dataset.

5.1 Movie vs. Novel over decades

Fig. 3 describes the usage of word 'movie' as compared to the word 'novel' over a span of 100 years from 1920 to 2020. It can be seen that in the 1920s, the word 'movie' was used a lot less than the word 'novel'. But over the decades, 'movie' has seen an exponential rise in its usage while 'novel' slowly decreasing below 'movie'.

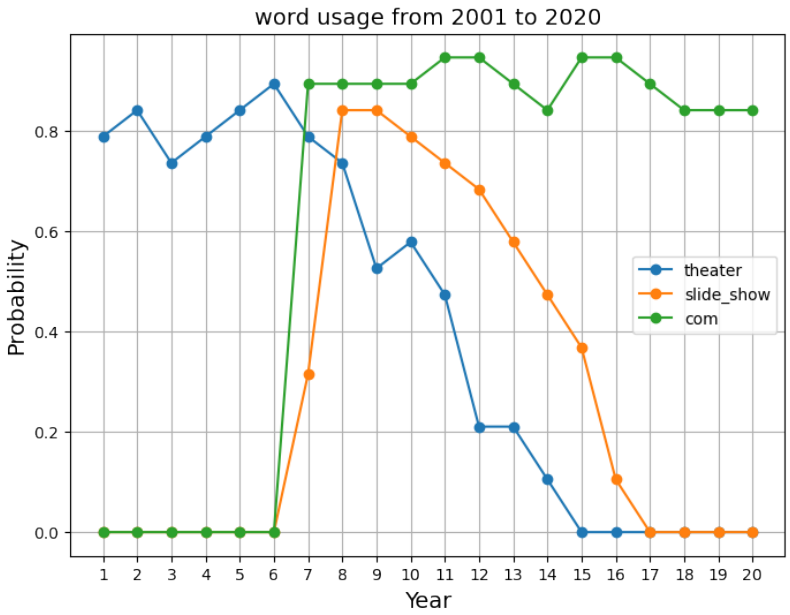

Figure 4: Technology from 2000 to 2020

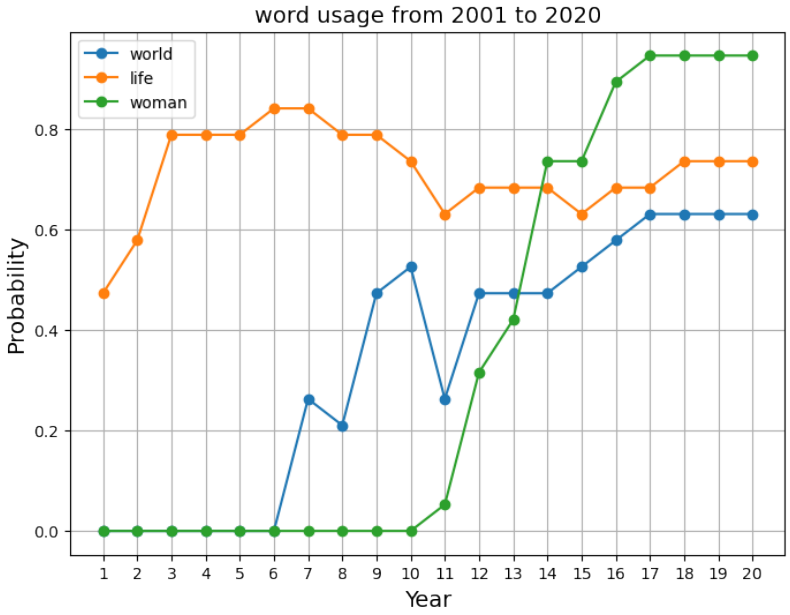

Figure 5: Gender Awareness from 2000 to 2020

5.2 21st century

As shown in Fig.4 and Fig.5, topics such as globalisation,equality and technology have increased in importance. A steady decline can be seen in the use of the word 'theater' over the years, as people have preferred online streaming services over going to theatres. An interesting curve can be seen in the usage of the word 'slideshow', as it rose rapidly in early 2000s but has been on a decline ever since. On the other hand, the usage of the words 'woman' and 'world' have seen an increase in importance over the years showing increased awareness about globalisation and equality.

6 Conclusion and Future Work

Information was extracted on key topics in news articles as well as how they change over time by using Dynamic Topic Modelling over a New York Times news articles data set. By examining the likelihood of the top key words in each subject, their popularity was tracked and observed as to how it changes over time. Although the model was able capture the trend in topics as well as words over time. There as some flaws in the model in terms of rigidity towards topics over time. The model assumes that a topic once formed will remain important over the complete span of time. It does not allow for topics to move in and out of importance over time. Rather than assuming a constant number of subjects, the most intriguing expansion to the approach given here is to integrate a model of how new topics in the collection emerge or fade away over time.

References

[1] "Ways to compute topics over time," https://jeriwieringa.com/2017/06/21/Calculating-and-Visualizing-Topic-Significance-over-Time-Part-1/, accessed: 2021-09-26.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[3] C. Meng, M. Zhang, and W. Guo, "Evolution of movie topics over time," *UR L: http://cs229. stanford. edu/proj2012/Meng ZhangGuo-EvolutionofMovieTopicsOverTi me. pdf. Cited April*, vol. 1, p. 2014, 2012.

[4] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[5] H. Zhou, H. Yu, and R. Hu, "Topic evolution based on the probabilistic topic model: a review," *Frontiers of Computer Science*, vol. 11, no. 5, pp. 786–802, 2017.