Multilabel Topic Model and User-Item Representation for Personalized Display of Review

Chin-Wei Huang*[1] and Pierre-André Brousseau†[2]

[1]Department of Mechanical Engineering, Polytechnique Montréal
[2]Department of Computer Science, University of Montréal

Abstract

On an online review platform, it is common for a user to be overwhelmed by the presence of too much information when browsing through the reviews regarding an item. To address this problem, in this paper, we reinterpret a supervised variant of LDA, the author topic model, to model the reviews as a corpus along with side information regarding an item and a user. We justified the use of such a model by comparing it to unigram as a benchmark, and discussed some of its advantages over other LDA variants. We also gave an illustrative example of personalized active highlighting of review to help a user navigate through a web-page of reviews.

1. Introduction

Community-driven online review platforms provide an abundant amount of information about products, businesses, restaurants, etc. Here, we borrow the collaborative filtering terminology, "item" and "user", since websites of these kinds usually serve as data sources that support a recommendation system. It is common, however, for a user to be overwhelmed by the presence of too much information when browsing through the reviews regarding an item. It is then possible to model the reviews as a corpus to address this problem.

As an example to model a corpus of documents along with multiple labels, we have collected reviews from *TripAdvisor*, a community-driven review platform for travelers. On *TripAdvisor*, it is the usual practice for a user to browse through multiple reviews regarding an attraction spot to get a sense of whether or not they would like to visit the place. Each of the reviews is written by a reviewer (user) about an attraction spot (item). Each reviewer can have multiple attributes, up to a maximum number of 19 user tags, and each attraction spot too has multiple tags (max. 237 from the data we scraped). We filtered out the data points whose corresponding user or attraction has no tags, (describe data, i.e. number) Our goal is to personalize the display of the reviews regarding an attraction by actively highlighting the parts expected to be of a user's interest.

The rest of this paper is organized as follows: In Section 2, we first briefly discuss the related work and justify our use of a specific form of topic model (i.e. Author topic model). Section 3 summarizes the notation and terminology. We describe some models (both with and without latent variables) that can potentially achieve personalized display of reviews in Section 4. In Section 5, we present our representation of user-item attributes in the author topic model framework, and demonstrate the applicative results. Finally, we summarize in Section 6 our extended interpretation of the model and some potential future work.

2. Related work

Latent Dirichlet Allocation (LDA) is a latent variable model designed to model collections of discrete data (Blei et al., 2003). Some common uses of LDA include document modeling and classification. In the former case, one seeks to project the word features onto a lower dimensional semantic space that is known as topics. The latter case can thus be thought of as a pre-training step that reduces the dimensionality of the feature space, and was shown to have better or equal predictive power compared to a model that was trained on features with full dimensionality. In (Mcauliffe & Blei, 2008), the authors used the principal component analysis (PCA) and its supervised counterparts, such as partial least square, as an continuous-variable analog of the relationship between LDA and supervised LDA (sLDA). It was shown that, using topic vector as explana-

*cw.huang427@gmail
†pierre-andre.brousseau@umontreal.ca

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

tory variable for regression, sLDA has much better predictive performance than the unsupervised LDA.

Other variants of LDA were also proposed to model documents along with side information, such as DiscLDA (Lacoste-julien et al.). As opposed to sLDA, DiscLDA is trained discriminatively via conditional likelihood. The author topic model (Rosen-Zvi et al., 2004), which our study is based on, provides an alternative way to model the dependencies among variables. Like in DiscLDA, the topic vector is input dependent, but the distribution over topic is no longer document-specific as in the above-mentioned models. This gives several advantages and will be discussed throughout this paper. Also as will be proved, when Gibb's sampling is used to approximate the posterior distribution on the topic vector, the transitional probability of the Markov chain is proportional to the posterior predictive distribution conditioned on the author responsible for the generation of a word in a document. This is an interesting perspective that inspired us, since its approximate inference using Gibb's sampling is similar to the work done by (Griffiths & Steyvers, 2004) but provides a different way to compress the information. We will continue with a more detailed discussion in Section 4.2.2.

Also worth noting is the labeled-LDA (l-LDA) model devised by (Ramage et al., 2009), a model that treats the generation of words as a random process distributed by a mixture of topics as well, which is conditioned on the multivalued properties associated with the document. Unlike author topic model, l-LDA requires the document-specific topic simplex to be conditioned on the labels as a constraint on topic dimensionality; prior probability, thus, is usually projected onto a much lower dimensional space. The intuition of this model allows it to involve multilabel information, but the advantage of using a latent variable model such as LDA to expand the dimensionality of unobserved topics to model larger corpus is sacrificed.

The l-LDA, for example, was used to fulfill a task known as snippet extraction (Ramage et al., 2009) in search engine optimization. Their work achieved passive highlighting of important parts of a document upon tag specification made by the user as a search query. In our case, tags are not associated with the documents, i.e. reviews, themselves. Instead, we know exactly who wrote the review and which place it was about. Here, using Author topic model, we treat the latent variable as the semantic representation of reviews, and the generation of each review is conditioned on the combination of user and attraction tags as labels.

Last, we would like to point out that the models discussed above all assume exchangeability for the words in a document. Research has been conducted to include markovianity in hopes of capturing the sequential nature of textual information, in terms of the syntactic dependencies (Griffiths et al., 2005), the local similarity of semantics within a sentence (Gruber et al., 2007) or across multiple segments (Du et al., 2012), the temporal evolution of topics (Blei & Lafferty, 2006), etc. In our study, though, we consider the simplest case of bag-of-words representation, which relies on the naive yet reasonable assumption of exchangeability that is considered to be useful for the task that we have described. Markov models, however, are powerful variants that can be used to capture a reviewer's sentiments, and thus are considered a potential extension to our research if one seeks to include rating information as an extra source of side information.

3. Notation

In this paper, the author topic model is reinterpreted to incorporate multilabel information that is tied with a document. As an application example, we analyze the reviews collected from the online tourist review platform TripAdvisor. The term review and document are thus used interchangeably. We also refer to each pair of attraction tag and user tag as a label (similar to the author of a document).

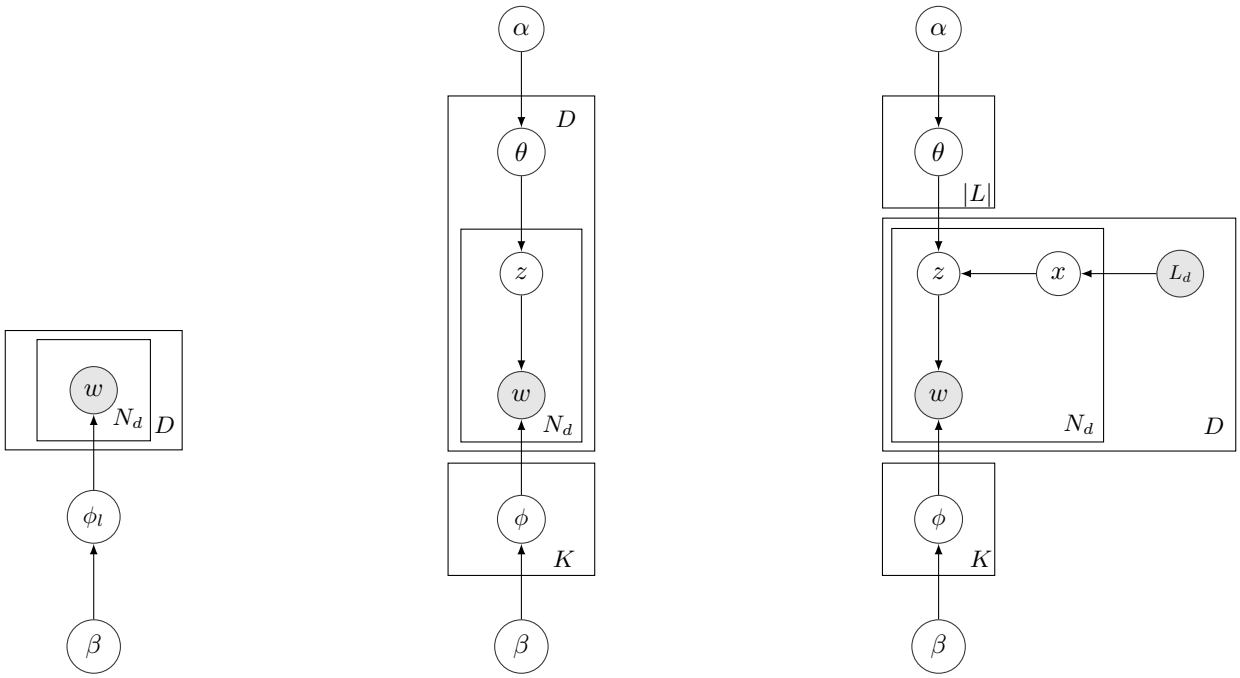
- A vocabulary is the full set of word tokens of size V.
- A word is the smallest unit comprising a document. Its vector, of length V, contains only one element equal to 1, with the others equal to 0. The number of words in a document is denoted by N_d.
- A document is one data point of length V. Its j^{th} elements is the word frequency that corresponds to the j^{th} word token in the vocabulary.
- A corpus is the data set consisting of several documents. It is denoted by D, with subscript indicating the partitioning purpose along with it, e.g. D_{train}.
- A label is a piece of side information tied with a document. It is represented by a binary vector of length $|L|$.

4. Modeling documents with multilabel information

4.1. Models without latent variables

4.1.1. Unigram model

Here we introduce the simplest approach to model text corpora. A unigram model, as shown in Figure 1a, is a distribution over word tokens. As a special case of the larger N-gram family, in which case one assumes the probability of a word given its contextual history can be approximated by the probability of observing it given the preceding $n-1$ words, the unigram model treats each word as independent of all the other words. We forgo the history of each word, and we will make the same assumption for the LDA model and its variants, i.e. conditional exchangeability later on.



(a) Unigram (b) LDA (c) Author Topic Model

Figure 1: Graphical Representation of Probabilistic Models discussed or used in this paper.

At the same time, we seek to model the side information that comes along with each document. Suppose a document is associated with a set of labels, and there are in total $|L|$ labels. The easiest way to do this is to build $|L|$ unigram models. In each one, we count the times a word occurs in all documents with the same label as the maximum likelihood estimate for the multinomial distribution over term frequency. We add a uniform prior and compute the posterior mean as a point estimate of the predictive distribution to avoid instances of zero probability.

The naive way to do prediction given new labels is to average out the probabilities of words associated with the labels. It is often the case that simple approaches like this might work at times, whereas one needs to know that this is not based on proper probabilistic assumption. A similar and supervised model is the multi-label mixture model propsed by (McCallum, 1999). The author model in (Rosen-Zvi et al., 2004), which assumes each author is drawn uniformly, is even more similar to our approach. In these cases one has to do EM or sampling to infer the posterior of the latent variable, and the posterior predictive distribution is computed by taking the weighted or arithmetic average of probabilities associated with labels/authors, which is more probabilistically formalized than our approach.

4.2. Models with latent variables

4.2.1. Latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003), in its most general form, can be used to compactly represent the semantics of words in a document by introducing a latent variable of dimensionality much smaller than the size of a corpus. In the introduction, we have seen many of its variants in the literature. In this section, we describe the generative model, and inherit some of the notations used in this general case to describe the model that we use in the following sections.

The graphical model of LDA is shown in Figure (1b). First, the distribution over topics per document (θ) is sampled from the Dirichlet prior (α) known as topic simplex. For each word in the document, we draw a sample from the chosen parameter as the topic (z) of the word. And finally each word (w) is sampled from a corpus probability (ϕ) conditioned on the topic. The corpus probability is a multinomial distribution per topic, and we can add a Dirichlet prior to make it fully Bayesian.

The computation of the posterior of word topic is intractable, since one needs to integrate over an infinite number of possible values of the parameters θ and ϕ. One can choose to do approximate inference, such as variational EM (Blei et al., 2003), Markov Chain Monte Carlo (MCMC) method (Griffiths & Steyvers, 2004) and expectation prop-

agation (Minka & Lafferty, 2002).

In this general form, one can model the document in an unsupervised manner. The strength of it is its compact document-wise semantic representation, as the topic distribution is sampled per document. With this, one can compare the similarity between documents or find out the underlying category of a document (i.e. clustering). To do prediction, one might be able to marginalize out the topics and integrate out the posterior. This is useful in the collaborative filtering example shown in (Blei et al., 2003). But in document modeling, (Rosen-Zvi et al., 2004) showed that the predictive power of LDA is limited by the amount of pre-trained part of a document. Also, since the side information provided by the labels or tags cannot be incorporated, it will be less useful and harder to do prediction using LDA for a new document unseen in the training set. The Author Topic Model in the next section will enrich our inferential toolkit in this regard.

4.2.2. AUTHOR TOPIC MODEL

In some cases, one might be more interested in a third variable that is associated with the generation of the document, such as the authors of a document. Each document is generated by a set of authors, denoted as L_d (as we are in favor of the term label in a generalized sense), where the subscript d is the index of a document. The author topic model as a graphical model shown in Figure (1c) describes the following dependencies.

1. For the topic density corresponding to each author:
   (a) Choose $\theta_l \sim \text{Dir}(\alpha)$
2. For the vocabulary density corresponding to each topic
   (a) Choose $\phi_j \sim \text{Dir}(\beta)$
3. For each document d where $d = 1{:}D$,
   (a) A group of authors L_d decide to write the document.
   (b) For each word i in the document where $i = 1{:}N_d$,
      i. An author x_i is chosen uniformly at random from L_d.
      ii. A topic z_i is chosen from the topic distribution specific to the author X_i.
      iii. A word w_i is chosen from the vocabulary distribution specific to the topic z_i.

Since the computation of the posteriors of the latent variables z and x is intractable, we need to do approximate inference. We approach it using the MCMC method (Griffiths & Steyvers, 2004; Rosen-Zvi et al., 2004). With the preset meta-parameters α as a conjugate prior over multinomial topics and β over vocabulary, we do this by integrating out θ and ϕ. This allows us to draw samples from the full conditional joint:

$$\begin{aligned} &p(z_i = j, x_i = l | z_{\neg i}, x_{\neg i}, w_i = v, w_{\neg i}, L_d) \\ &\propto p(w_i = v, w_{\neg i} | z) p(z_i = j, x_i = l, z_{\neg i}, x_{\neg i} | L_d) \\ &\propto \frac{n_{vj} + \beta}{\sum_{v'} n_{v'j} + V\beta} \frac{n_{lj} + \alpha}{\sum_{j'} n_{vj'} + K\beta} \end{aligned} \tag{1}$$

which is a product of two marginal joint distributions. See Appendix A for the derivation.

Specifically, the second term of Eq (1) is conditioned on the author of the document, which is similar to the class-dependent linear transformation seen in DiscLDA (Lacoste-julien et al.), but with different parameterization. In the case of DiscLDA, the distribution of topic per document is drawn from the prior, and each of its element corresponds to one class, which will be transformed via a stochastic matrix that introduces a mixture of multinouilli distributions. Differently parameterized, author topic model can be interpreted similarly if we marginalize out the author responsible for a word, which yields $\frac{1}{L'_d L_d}\Theta L_d$, where $L'_d L_d$, a normalizing factor, gives us the number of non-zero elements in L_d and Θ, a K-by-$|L|$ matrix with K being the total number of topics, has its $|L|$ columns drawn from the Dirichlet prior.

With a different structure, author topic model allows topic distributions to be learned per author. As in LDA, we can count the number of times a topic is assigned to a word token to compute a rough point estimate of the posterior mean of ϕ_j. Counting the number of times an author is associated with a topic to estimate the posterior mean of θ_l serves as a second time compression of the information. This is the strength of author topic model, as instead of having a point estimate of distribution over topic to describe a document on a semantic simplex, we can learn the topic distribution associated with an author collectively across all the documents related to this author. The approximate inference we use can thus be considered as a shared parameter training. And the document-specific representation can also be derived by marginalizing out the authors to get a uniform mixture of multinouilli distributions. With this, we can have the model generalize to unseen data points to do prediction based on a new set of authors.

5. Multilabel topic modeling and user-item representation

The Author Topic Model discussed in the previous section can be generalized to model the semantics of words conditioned on other relevant multi-valued properties, other than authors. Probability of a certain topic can then be estimated with respect to one of the many labels instead of a specific

User\Item Tags	Outdoor	Shopping	Museums	...	Mountains
Thrill Seeker	θ_{11}	θ_{12}	θ_{13}	...	θ_{1I}
History Buff	θ_{21}	θ_{22}	θ_{23}	...	θ_{2I}
Shopping Fanatic	θ_{31}	θ_{32}	θ_{33}	...	θ_{3I}
⋮	⋮	⋮	⋮	⋱	⋮
Family Vacationer	θ_{U1}	θ_{U2}	θ_{U3}	...	θ_{UI}

Table 1: User-Item Representation.

document. This gives additional advantages over the unsupervised LDA, such as the possibility to analyze the similarity between two labels and the potential to infer the distribution over words given a new set of labels. Here, given a pair of multi-labeled traveler and attraction spot, we seek to personalize the display of attraction review by making use of the Author Topic Model framework.

Here we describe how extraneous user tags and attraction tags can be leveraged. Take for example a user tagged as a *Luxury Traveler* who visited an attraction spot with the tag *Scenic Railroads*. It is our intuition to discriminate between the traveler's proclivity for comfort or elegance and the nature of rail mass transit that usually compensates coziness for reasonable prices. It would thus not surprise us if the traveler "did consider taking a **private taxi** for the day" and as a consequence described themselves as "**sardines** in a can for an hour" on the train. This motivated us to treat each pair of attraction tag and user tag as a label, and model the distribution on semantics associated with one of the multiple labels of a review to allocate higher probability to some word tokens such as the ones in boldface.

Given the number of all possible attraction tags I (I for item) and that of all user tags U, 237 and 19 in our case, respectively, we redefine the topic matrix Θ as a K-by-U-by-I semantic tensor. Each component projected onto the second and third dimensional space, see Table (1), corresponds to a length K vector $\theta_{:,u,i}$ that describes the semantic density characterized by the matching of the u's user tag with the i's attraction tag. The mixed density with x marginalized out can be interpreted as the general idea expressed by the user about the item, which is simply the arithmetic mean of all semantic vectors of the exhaustive pairwise combinations of user and attraction tags.

5.1. Model evaluation and selection

Following our terminology, when given a new set of user tags and attraction tags, we want to know how our model performs in terms of prediction accuracy. We can do this by looking at the log-likelihood of the posterior predictive distribution. We take the one-dimensional label (as in author topic model) for instance, and this can be generalized

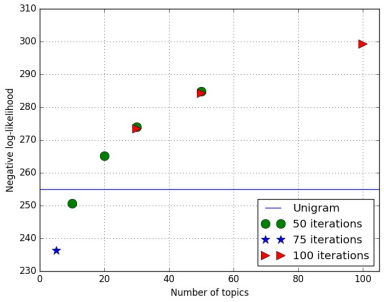


Figure 2: Negative log-likelihood: model selection.

to the two-dimensional case:

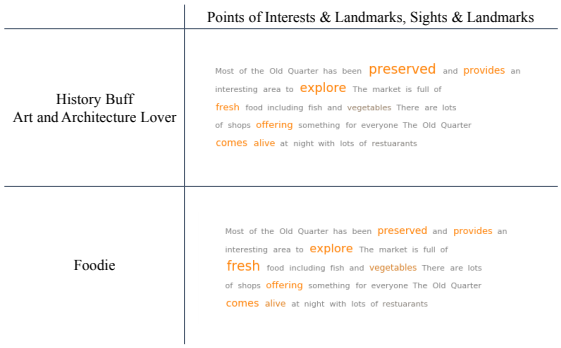
$$w|L_d, Data_{train} \sim \text{Multi}(\frac{\bar{\Phi}\bar{\Theta}L_d}{L'_d L_d}) \tag{2}$$

where $\bar{\Phi}$ and $\bar{\Theta}$ are the posterior means for vocabulary and topics. Derivation can be found in Appendix B. We then can calculate the negative log-likelihood of the multinomial (which strictly speaking is multinoulli instead), or cross-entropy, on the validating set. We refer to cross-entropy because to compare the result with unigram we actually do not assume a predictive distribution for the latter case. We add N_{val}, the number of documents in the held-out set, as a scaling factor.

$$\begin{aligned} &\frac{1}{N_{val}} \log p(Doc_{val}|L_{val}, Data_{train}) \\ &= \frac{1}{N_{val}} \sum_{d=1}^{D_{val}} \sum_{i=1}^{N_d} w_{d,i} \log \phi_{d,i}^{post} \end{aligned} \tag{3}$$

where ϕ_d^{post} is the posterior predictive parameter conditioned on L_d and the training data, as in Eq. (2).

We see that the topic distribution takes the arithmetic mean of the label-specific posterior probabilities. We then treat the probability assigned to a topic as a mixing proportion to have a weighted average of the posterior over the words.



(a) Review on Hoi An Ancient Town

(b) Review on Ben Thanh Market

Figure 3: Examples of Highlighted Reviews.

This has a similar effect as the other augmented LDA models, such as the transforming matrix T^y used in DiscLDA (Lacoste-julien et al.). Here, one can see clearly that the advantage of author topic model is that one can store the label-specific parameters ($\bar{\Theta}$) and that the model can generalize to other combinations of labels unseen in the training set; for example, when a new item or user unseen in the training set is given.

We start with training the model using a set of training corpus of size 449827, with a vocabulary set of length 78078. After iterations of sampling, we compute the negative log-likelihood using the held-out validation set of size 112445. We present the results for different models in Figure 2. The first advantage of having a latent variable model over a simple model such as unigram is flexibility. One sees that we can alter the complexity of the model by changing the number of topics, for example, making it possible to explore the parameter space to account for more complicated tasks such as language modeling.

We also see that the author topic model outperforms the unigram model in terms of generalization ability when we choose a smaller number of topics. This is different from the findings of (Blei et al., 2003) and (Griffiths & Steyvers, 2004) using the LDA model, yet similar to the result of Author topic model (Rosen-Zvi et al., 2004).

We attribute the smaller number of topics to the incorporation of side information and the nature of reviews. Introducing the author-specific topic density as a parameter to be learned increases the complexity of the model. The perplexity for a held-out document tends to be higher when the prediction is conditioned on the author information, which in our case can be thought of as an input feature vector of length 4503 (237×19). Also, take the supervised versions of PCA for example, in which case the model will find the covariate (the latent variable) that best explains the covariance of two conditionally independent observed variables. More specifically, the canonical correlation analysis (Bach & Jordan, 2005) model allows some of the variance in both of the input space and output space to be explained by a separate subspace. In a similar sense, we can consider the author information and the documents have some covariance to be explained, which when given a fixed finite number of data points is limited. As a result, when we expand the latent space of topics the model is prone to overfitting.

Second, our corpus is believed to have a smaller amount of semantic information as compared to the common corpora in topic modeling such as scientific documents and news articles. This becomes obvious if we do not filter out the most common words in the preprocessing stage. Some words appear in most of the documents and do not possess discriminative power. For example, we found that our model built on this larger corpus would allocate much higher probability to words such as “temple", “experience" and “visit". Different from the documents in a scientific corpus usually having one or multiple recognizable topics, the travelers’ reviews are believed to be generated from an inherently smaller number of “sub-topics" as traveling itself is already considered as a super-topic.

In the next section, we choose the author topic model with 5 topics and give an applicative example as to how to actively highlight a part of a review.

5.2. Personalized display of review: an application

In this section we see how we can achieve personalization of review display using the model described above.

We have seen how to predict words conditioned on a new set of user and attraction tags. The way to highlight part of a review regarding an attraction spot based on a user’s attributes is to scale the printed words in terms of size and possibly change the color to a degree proportional to the conditional probability. For illustrative purpose, we pick up two reviews (for two different attraction spots) out of a hidden set of test data. Since a review is about an attraction

spot, we are naturally given a set of attraction tags on one side. Now pretend a new user bearing an arbitrary set of attributes, say *History Buff* and *Art and Architecture Lover* is browsing through the webpage of the reviews about this attraction spot. Now we have a set of new authors/labels to condition on. First, averaging out the labels and marginalizing out the topics yield the probability distribution over the word tokens. We look only at the set of words in the intersection of the words in the review and words in the vocabulary (the set over which we have the probability density). This approach is intuitive since the set of words not in the vocabulary are either too common or too rare. It is also simplifying the problem since otherwise we cannot highlight in the same way the words of which we do not have the quantitative description (i.e. probability mass). Second, we set the default color to be grayish, font size to be minimum. We change the RGB proportion of a word by the amount of rescaled probability mass it possesses to make to distinguishable. We also scale the font size according to $size^{1+\frac{p_w}{s}}$, where p_w is the rescaled probability mass of the word token and s is a scaling factor. Note that we picked up this formula and some arbitrary parameters for exemplification, and these are to be fine-tuned in operation.

We present the illustration in Figure 3. The column name is the attraction tags of the place the review is about. The index name of each row is the user tag(s) of a potential user. Look at the left hand side where we have a review on an attraction spot called Hoi An Ancient Town in Quang Nam Province, Vietnam. We see that some words such as “preserved" and “explore" are highlighted with greater intensity for a *History Buff* and *Art and Architecture Lover*, whereas some others such as “fresh" and “vegetables" are emphasized for a *Foodie*. On the right hand side, we have another review on Ben Thanh Market, a place in Ho Chi Minh City, also in Vietnam. We see that words “shoes", “accessories" and “souvenirs" are all much more perceivably highlighted for a *Shopping Fanatic* than a user considered as a *Thrill Seeker* and *Nature Lover*.

6. Discussion and conclusion

The simplest way to model text corpora without taking account of the contextual information is to use the unigram model in an unsupervised manner. On the one hand, our naive approach to average out probabilities of words associated with the given labels was shown to be empirically acceptable. Yet a probabilistically formalized way to do this is to use the conditional mixture model (McCallum, 1999; Rosen-Zvi et al., 2004). On the other hand, the LDA model makes use of the latent topic variable to represent the semantics of words, allowing one to take into consideration the correlated nature of word tokens. The author topic model can be thought of as a hybrid of the two cases: an admixture model that has two layers of nodes associated with each word in a document. An author is tied with a topic distribution, and a topic is tied with a distribution on words.

It is also possible to use other augmented LDAs to this end. For example, using DiscLDA (Lacoste-julien et al.) allows us to infer the document-specific topic density, which is supposed to be more accurately representative of the review when compared to the author topic model, as the latter uses the arithmetic mean of multiple distributions instead. Throughout the report, though, we have put much stress on author topic model’s advantage in learning "author"-wise topic distribution. This is useful in the online setting where reviews are generated all the time and thus it will be less efficient to do approximate inference on the document-specific density. The faster way is to use the extraneous attributes of the item and user to have a rough representation of the semantics of a review.

In addition, it is important as well to examine some of the assumptions we made to simplify the problem in using the model. In the first place, we assumed conditional independence for the word tokens, so that we can use the bag-of-words representation in our model. Disregarding the order of a sequence can simplify the model, and is useful in topic modeling. However, in natural language processing it is not always the best document representation. The use of Markov chain serves to be more meaningful especially when we seek to model the sentiments (which will be explained later), since the syntax plays an indispensable role in encoding the sentiments expressed by a sentence. Also, we make use of the term frequency per document representation, simply because our sampling scheme relies on integer input. As shown in (Wilson & Chew, 2010), it is possible to incorporate a term weight in the Gibbs sampling algorithm, so that a full conditional distribution will depend on a weighted proportion of the counts of samples, which will help address the problem of too common and less discriminative words. Thus it can be used to replace the preprocessing task of filtering out such words.

One last thing worth noting is that on review platforms such as TripAdvisor users are usually encouraged to rate and and comment on items. It is thus interesting to consider modeling ratings and reviews altogether, which enables one to account for the sentiment of a review. In traditional collaborative filtering context, for instance, people are usually concerned with predicting rating toward an item made by a user. It is possible to use probabilistic models to include side information in matrix factorization (PMF). In (Shan & Banerjee, 2010), LDA and PMF model are coupled by sharing the topic distribution, where, for example, side information such as movie genre is modeled as words. However, in our setting, labels are placed higher in the hierarchy, and

we can accordingly devise a model where x is marginalized out and replaced by an auxiliary node of a mixture of multinouillis which can be considered as a confounding variable that generates both reviews and rates. This allows us to extend our multilabel model framework to include rating information to be co-trained with textual information as a potential future work.

References

Bach, Francis R. and Jordan, Michael I. A probabilistic interpretation of canonical correlation analysis. Technical report, 2005.

Blei, David M. and Lafferty, John D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 113–120, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143859.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

Du, Lan, Buntine, Wray, Jin, Huidong, and Chen, Changyou. Sequential latent dirichlet allocation. *Knowledge and Information Systems*, 31(3):475–503, 2012. ISSN 0219-3116. doi: 10.1007/s10115-011-0425-1.

Griffiths, Thomas L. and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004. doi: 10.1073/pnas.0307752101.

Griffiths, Thomas L., Steyvers, Mark, Blei, David M., and Tenenbaum, Joshua B. Integrating topics and syntax. In Saul, L. K., Weiss, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 17*, pp. 537–544. MIT Press, 2005.

Gruber, Amit, Rosen-zvi, Michal, and Weiss, Yair. Hidden topic markov models. In *In Proceedings of Artificial Intelligence and Statistics*, 2007.

Lacoste-julien, Simon, Sha, Fei, and Jordan, Michael I. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification.

Mcauliffe, Jon D. and Blei, David M. Supervised topic models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 121–128. Curran Associates, Inc., 2008.

McCallum, Andrew Kachites. Multi-label text classification with a mixture model trained by em. In *AAAI 99 Workshop on Text Learning*, 1999.

Minka, Thomas and Lafferty, John. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pp. 352–359, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1-55860-897-4.

Ramage, Daniel, Hall, David, Nallapati, Ramesh, and Manning, Christopher D. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pp. 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6.

Rosen-Zvi, Michal, Griffiths, Thomas, Steyvers, Mark, and Smyth, Padhraic. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pp. 487–494, Arlington, Virginia, United States, 2004. AUAI Press. ISBN 0-9749039-0-6.

Shan, H. and Banerjee, A. Generalized probabilistic matrix factorizations for collaborative filtering. In *2010 IEEE International Conference on Data Mining*, pp. 1025–1030, Dec 2010. doi: 10.1109/ICDM.2010.116.

Wilson, Andrew T. and Chew, Peter A. Term weighting schemes for latent dirichlet allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 465–473, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL http://dl.acm.org/citation.cfm?id=1857999.1858069.

A. Full conditional joint for Gibb's sampling.

We now turn to the simple case of Dirichlet-Multinomial distribution. We have $\theta \sim Dir(\alpha)$ and $Z \sim Multi(N, K; \theta)$. The posterior is given as

$$\begin{aligned} p(Z|\alpha) &= \frac{Beta(N+\alpha)}{Beta(\alpha)} \\ &= \frac{\Gamma(\sum_j \alpha_j)}{\Gamma(N + \sum_j \alpha_j)} \prod_j \frac{\Gamma(N_j + \alpha_j)}{\Gamma(\alpha_j)} \end{aligned} \tag{4}$$

From here, we can calculate the full conditional

$$\begin{aligned} p(z_t = k|z_{\neg t}, \alpha) &\propto p(z_t, z_{\neg t}|\alpha) \\ &= \frac{\Gamma(A)}{\Gamma(N+A)} \prod_j \frac{\Gamma(N_j + \alpha_j)}{\Gamma(\alpha_j)} \\ &\propto \prod_j \Gamma(N_j + \alpha_j) \\ &= \Gamma(N_k + \alpha_k) \prod_{j \neq k} \Gamma(N_j + \alpha_j) \\ &= (N_{k,\neg t} + \alpha_k)\Gamma(N_{k,\neg t} + \alpha_k) \prod_{j \neq k} \Gamma(N_{j,\neg t} + \alpha_j) \\ &= (N_{k,\neg t} + \alpha_k) \prod_j \Gamma(N_{j,\neg t} + \alpha_j) \\ &\propto N_{k,\neg t} + \alpha_k \end{aligned} \tag{5}$$

Now we can decompose the full conditional joint distribution of the author topic model for Gibb's sampling into two Dirichlet posterior distributions:

$$\begin{aligned} p(z_{i,j}, x_{i,l}|w_{i,v}, w_{\neg i}, z_{\neg i}, x_{\neg i}, L_d) &\propto p(z_{i,j}, z_{\neg i}, x_{i,l}, x_{\neg i}, w_{i,v}, w_{\neg i}|L_d) \\ &= p(w_{i,v}, w_{\neg i}|z; \beta) p(\{z_{i,j}, x_{i,l}\}, z_{\neg i}, x_{\neg i}|L_d; \alpha) \end{aligned} \tag{6}$$

The first term is the full conditional of Dirichlet-Multinomial conditioned on the topic variable, which is proportional to the normalized count of times topic j has been assigned to word i plus the prior pseudo-count. The second term is two-dimensional, as we're sampling a topic and an author for word i, and is proportional to the normalized count of times author l has been associated with topic j, conditioned on the set of authors for this document. The uniform probability of will be incorporated into the normalization term and can be neglected. As a result, we are sampling topic and author jointly from a $K \times \sum_l L_{d,l}$ non-zero probability matrix with other authors not the in set L_d being assigned zero probability.

B. Posterior predictive distribution.

Given a new set of labels, we seek to predict the most probable set of words that are related to the labels based on the training data, i.e. $p(w|L_{new}, Data_{train})$. Here we marginalize out the latent variables representative of topic z and author x for a given words:

$$\begin{aligned} p(w|L_{new}, Data_{train}) &= \sum_j \sum_l p(w|z_j, Data_{train}) p(z_j|x_l, Data_{train}) p(x_l|L_{new}) \\ &= \sum_j \int_{\phi_j} p(w|z_j, \phi_j) p(\phi_j|Data_{train}) d\phi_j \sum_l \int_{\theta_l} p(z_j|x_l, \theta_{j,l}) p(\theta_i|Data_{train}) d\theta_{j,l} \frac{L_{new,l}}{\sum_{l'} L_{new,l'}} \\ &= \frac{1}{\sum_{l'} L_{new,l'}} \sum_j^K \bar{\phi_j} \sum_l^{|L|} \bar{\theta}_{i,l} L_{new,l} \\ &= \frac{1}{{L_{new}}' L_{new}} \bar{\Phi} \bar{\Theta} L_{new} \end{aligned} \tag{7}$$

which is just composed of the posterior means of ϕ and θ. Since we assume Dirichlet priors, these can be easily computed by counting the samples of topics assigned to each word and label. Here the last line is the vector form that corresponds Eq. (2).