<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *2016 IEEE International Conference on Cloud Computing and Big Data Analysis*.

Citation for the original published paper:

Song, W., Zou, L. (2016)

LDA-TM: A Two-Step Approach to Twitter Topic Data Clustering.

In: *Proceedings of the 2016 IEEE International Conference on Cloud Computing and Big Data Analysis* (pp. 342-347). IEEE CS

IEEE CS

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:du-22827>

LDA-TM: A Two-Step Approach to Twitter Topic Data Clustering

Luyi Zou, William Wei Song*

Information Systems & Business Intelligence
Dalarna University
Borlänge, Sweden
e-mail: wso@du.se

Abstract—The Twitter System is the biggest social network in the world, and everyday millions of tweets are posted and talked about, expressing various views and opinions. A large variety of research activities have been conducted to study how the opinions can be clustered and analyzed, so that some tendencies can be uncovered. Due to the inherent weaknesses of the tweets – very short texts and very informal styles of writing – it is rather hard to make an investigation of tweet data analysis giving results with good performance and accuracy. In this paper, we intend to attack the problem from another aspect – using a two-layer structure to analyze the twitter data: LDA with topic map modelling. The experimental results demonstrate that this approach shows a progress in twitter data analysis. However, more experiments with this method are expected in order to ensure that the accurate analytic results can be maintained.

Keywords—big data; twitter data; data analytics; LDA; topics model

I. INTRODUCTION

The Twitter System is an online social network service that enables users to post and read messages called "tweets". As more and more people enjoy posting tweets and sharing information with others, Twitter System gains its popularity world-wide. As of May 2015, Twitter has had more than 500 million users, among which about 302 million are active users. Millions of daily information is posted on Twitter, reflecting various aspects of people's work and life. Through analyzing Twitter data to understand what is popular on Twitter becomes possible, which can provide users with real-time hot topics focused by the users on Twitter rather than the only news provided by followers. It can also help businesses to make accurate decisions from serve their customers based on the hot topics discovered.

It is hard to directly find out patterns for tweets and what topics these tweets focus from billions of twitter texts, given the features of a tweet being short-texted, very informal, strong temporal, and mix-language based. Moreover, using a short period (days or weeks) of twitter data samples for topic analysis cannot provide a convincing result while considering a long term twitter data leads to a large number of twitter data and messy topic patterns. In this research we consider a two-step approach to twitter data analysis. The first is to generate a topic model and the second to cluster tweets into topic-based categories.

Latent Dirichlet Allocation (LDA) is a commonly used topic model for text data mining [1]. A topic model is a type of statistical models for discovering the abstract "topics" that occur in a collection of documents. However, the major challenge of Twitter data analysis for LDA is that a tweet text is too short. The theories of those models require a reasonable large size of the texts in order for the topic model analysis to work properly. However, messages in Twitter are restricted to maximum 140 characters. The average length of Twitter posts is less than 20 words. After removing stops words, only about 3-7 words contain useful information. The performances of those models are vulnerable with such little information. The second challenge of the models is that Twitter data is too informal and filled with colloquialisms messy and colloquial to be accurately classified with useful results. Most of the tweets do not follow grammar, and as a result, noun, adjective and verbs do not provide very useful information for the topic models.

Considering the above challenges, we aim to identify an efficient pipeline for analyzing and visualizing Twitter hot topics. We use topic models to cluster Twitter messages into different topics and then use the topic's words combined with the tweets' influence to graphically visualize the trends of the topics. In this paper, by looking into the twitter syntax, a pooling schema is proposed to enhance the performance of topic model on Twitter. Meanwhile, in order to identify the accuracy of the proposed method, we make a comparative study during each step of the entire data analysis process. We look at the performance results before and after the pooling schema for data sampling is introduced, the performance before and after the use of the TF*IDF method for data processing, and the performance between LDA and the LDA derived model – the CTM for topic model. The experiments' results show that the LDA with TF*IDF and pooling has the most reliable pipeline, that meet the expected requirements.

The rest of the paper is organized as follows. Section II provides a review of related work. Section III shows the implementation of the methods. Section IV discusses the procedure of the data collection and preparation and the results of comparison of the methods proposed. The conclusion and discussion are presented in the final section, Section V.

II. LITERATURE REVIEW

In recent years, many new research methods have been introduced in short text classification and clustering

applications. It has become a hot topic in the field of natural language and culture text mining. Traditional methods of text feature extraction and representation in the classification often encounter many problems.

The topic model is an unsupervised or supervised probabilistic model aiming to discover the topics from a collection of documents. The topic structure is a hidden structure that the topics, per-document topic distributions, and the per-document per-word topic assignments. A topic model method enables us to organize and summarize electronic archives at a scale that would be impossible to be done through human annotation [2].

The most popular topic model is the Latent Dirichlet Allocation (LDA) [3] model, which is a Bayesian model for discrete data where topics are assumed to be uncorrelated. The Simple LDA is a powerful tool for discovering and exploiting the hidden thematic structure in a large archive of articles. However a number of derivative models from LDA have been proposed for more complicated situations and LDA has been extended in several ways:

Relaxing assumptions. One objective of topic modelling research is to extend the assumptions to uncover more sophisticated structure in the texts. For example, Griffiths, et al developed a topic model that switches between LDA and Hidden Markov Model (HMM) [4]. Wallach developed a topic model that assumes that the topics generated words conditionally on the previous word [5]. These models extend the parameter space and show a performance improvement. Another assumption is the extension of the order of documents. For example, Blei developed a dynamic topic model involving the order of the documents, which gives a richer posterior topic structure than LDA [6]. A third extension of assumption is to remove the fixed number of topics. Teh proposed a Bayesian non-parametric topic model learning the number of topics during the posterior inferences [7]. Correlated Topics Model (CTM) is also an extended version of LDA, which allows the occurrence of topics to exhibit correlation. The spherical topic model considered the situations where words are unlikely to appear in a topic [8].

Incorporating metadata. In many cases, the articles will contain additional information, such as author, title, geographic locations, and links as well as what we may want to consider. The author-topic model [9] is an early success story under this category. In this model the topic proportions are attached to authors. The author-topic model allows for inferences about authors as well as the original documents. Several topic models were proposed to consider links, such as citations in scientific papers and hyperlinks in webpages. The related topic model [10] assumes that each document is modelled as in LDA and the link between the documents depends on the distance between their topic proportions. The general purpose methods for incorporating metadata into topic models include Dirichlet-multinomial regression models [11] and supervised topic models.

In this paper we only focus on the LDA topic model and its extension, the CTM topic model, because LDA is the most popular model currently applied to short text documents in Twitter currently [12]. Moreover, the R package topic models provide an implementation of CTM

algorithm developed by Blei and an implementation of the LDA topic model with Gibbs sampling developed by Phan and co-authors [13], [14]. Another popular topic model is Labeled-LDA. The Labeled-LDA [15] constraining the normal LDA topic model by defining a one-to-one correspondence between LDA's latent topics and user tags. This allows the Labeled-LDA model to directly learn word-tag correspondences. For example, a news story about a highway transportation bill is naturally labelled as transportation and politics. Microsoft Research Center conducted a project called *Twahpic* which is based on Labeled-LDA. And research targeting on Twitter topic model has been done. It has several directions.

Polling Schema Improvement. In this type of research, researchers investigate the methods to improve topics learned from Twitter content without modifying the basic machinery of LDA [16]. The researchers try to improve the performance by aggregating tweets in the data pre-processing step of LDA. Liangjie and Brian tested several pooling schema, including associated information from user's profiles, and precision of different schema ranging from 42% to 61%. Rishabh M, et al tried several schemas, including **Author-wise pooling**- Pooling tweets according to authors, **Temporal Pooling**- pooling gathering all tweets posted within the same hour to capture the case when a major event occurs. A large number of users often start tweeting about an event within a short period of time. **Hashtag-based Pooling** – pooling by hashtag is another pooling schema, which identifies specific topics or events. They find that the TF similarity-based hashtag assignment to Hashtag-based pooling outperforms all other pooling strategies [16].

Considering the inability of LDA of modelling correlation between topics, Blei and Lafferty [12] developed the Correlated Topic Model (CTM). They supposed the topics exhibit correlation via the logistic normal distribution and modify the parameters of CTM so that the divergence between approximation and true distribution is small. They applied CTM and LDA on articles from journal Science and found CTM has a better fit than LDA as well as providing a visualizing way of the dataset [12].

Huang & Malisiewicz [17] consider that Expectation-Maximization provides a method for LDA and CTM, which are, unsupervised training. However, in some cases, if one can observe the latent variables during the training process, then it makes more sense to take advantage of them. For CTM, the way to compute the parameter β is equal to LDA, but estimating μ and Σ is more involved since the Logistic-Normal distribution is not a member of the exponential family. The distribution of z_i given Logistic Normal parameters is called a Hierarchical Logistic Normal distribution [17].

III. TWITTER TOPIC MODEL IMPLEMENTATION

In this section, we investigate the feasibility of an application to graphically present trends uncovered from Twitter topics using the topic models. We discuss two difficulties in Twitter LDA model. We will first discuss the converging issue with the Gibbs Sampling method and then

examine how to improve the keywords-relevance to topics method.

A. Converge Issue in Gibbs Sampling

The difficulty for the Twitter dataset is that, different from the articles which contain hundreds of words, the useful words in a tweet after cleaning up URL and stop words contain only 3-7 words, which makes it hard to converge at the Gibbs Sampling step.

For LDA, Gibbs sampling aims to determine document-topic portions θ_d , and the topic-word distribution $\phi^{(z)}$. The Gibbs sampler for LDA needs to compute the probability of topic z being assigned to a word w_i , given all other topic assignments to all other words. In other words, we are interested in computing the following posterior:

$$P(z_i | Z_{-i}, \alpha, \beta, w)$$

The LDA Gibbs sampler randomly initializes the topics for each word in every document, and then loops over iterations where in each loop a topic is re-sampled from the posterior mentioned above, and the for specific word assigned to all topics can be calculated as:

$$P(z = k | \bullet) = (n_{d,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$$

here $n_{d,k}$ is the number of words assigned to topic k in **document** d ; $n_{k,w}$ is the number of times the word w is assigned to topic k ; and n_k is the total number of words assigned to topic k .

While for long-text documents, the probability of topic z being assigned to a word w_i , as for p ($z=k|.$), will bias towards certain topic as for $n_{k,w}$ will dramatically bias towards the assigned topic after iterations. The reason for this is that words tend to being highly repeated in long documents. However, in Twitter, words appear only once, and after iterations, the posterior probability described above is almost even each time when it resampled. Thus for Twitter, Gibbs sampling will not converge or bias towards certain topic for a specific document after whatever iterations it might take. A simple solution for this problem is just to replicate the sentence in Twitter, which will mandatorily increase the replicates for a word; this solution turns out to be fairly efficient to separate topics for two tweets. For example, two tweets (Tweets 1: "PICTURES: Arsenal get ready for their final Premier League game of the season" and Tweet2: "Jailed Washingtonpost reporter in Iran, Jason Rezaian, set to go on trial May 26 for espionage.") need to be classified to two topics since they focus on different topics. Without applying the pooling schema, the probabilities of classifying the word "PICTURES" to two topics are 0.032 and 0.044 after iterated for 100 times. It performs like randomly assign a word to a topic. However, after adding the pooling schema and the iteration, the two probabilities of classifying the word to two topics become 0.008 and 3.92, which shows the word has a higher probability belongs to one of the topics. Meanwhile, after applying the method to the two tweets, the words in the Tweets 1 are all classified to

one topic and the words in the Tweets 2 are all classified to another topic.

B. Improve Keywords Relevant to Topics

The difficulty for the Twitter Topic Model is that a regular noun, adjective, and verb are not very meaningful in the Twitter short word context. For example, "RT to win 2 passes to The Charlatans gig on the roof of Hotel Football. I'll pick a winner at midnight on Saturday". In this twitter, football is actually not meaning football the soccer sport, but a hotel's name. Another example is "Bill Paxton and Lance Henriksen are the only actors to have been killed by a Terminator, an Alien, and a Predator", you won't know that this tweet is not about politics, unless you know that Bill Paxton is an American actor and film director, and Lance Henriksen is an American actor in science fiction movies. Since "kills" and "Predator" are words appearing in the social and political news quite often; this tweet will be easily put inside a political category by topic model. An easy way to correct this is to associate relative information with every tweet and give a tweet a more meaningful context environment. We download more tweets whenever encountering '#', which is used to identify certain events or topics; or capitalized names, which are normally a person's name or a place. Therefore the content in messages is limited, but can contain more meaningful information for the topic model to work. Results turn out to be significantly improved. We test the Gibbs LDA model with around 80 tweet's from "Twitter Top 1000", including BBC News, CNN News, New York Times, USA Today, Wired, Peoples, and so on. After associate with context, 'sandra bullock' (person's name), 'top gear' (movie's name) appear in the top frequently words under certain topics, otherwise, people's name or place's name will be treated as separate and meaningless word and will never show up on the list.

IV. EVALUATION

We use *precision* to evaluate the methods. To identify the *precision*, we manually determine which topic a tweet belongs to as the topic model result. For example, from the tweet "Don't miss @Xtina's last night on @Nashville ABC as Jade-St. John! Tune in 10pm ET on @ABC! See a sneak peek here", where "Jade-St. John" is a female pop star, this tweet originally belongs to the entertainment topic. The precision rate p is defined as the number of tweets correctly identified by the model to a specific topic divided by the total number of tweets classified to this topic.

A. Topic Models with Pooling Improvement

In this section, we explain the fundamental reason that LDA-based methods does not work well on Twitter short text environment. So we first compare topic models with pooling schema. In order to identify which pipeline is more suitable for Twitter data classification, we compare the precision of CTM, LDA, CTM with TF*IDF and LDA with TF*IDF. The average precision of each category during the sampling days shows in Table III. We use standard deviation to measure the error of the precision.

TABLE I. COMPARISON PRECISIONS BETWEEN TOPIC MODELS WITH POOLING IMPROVEMENT

topic models	Precision (Std Dev)
CTM	0.35 (0.03)
LDA	0.33 (0.09)
CTM-TF*IDF	0.47 (0.06)
LDA-TF*IDF	0.57 (0.03)

Figure 1. Comparison precisions between topic models with pooling improvement

From the above Fig. 1, we know that the precisions for LDA and CTM both increase after adding TF*IDF during data processing step. This demonstrate that TF*IDF strongly supports eliminating the words with high frequency but less importance, which has the same effect as removing the stop words manually. From the values of standard deviation, we observe that the LDA has the highest standard deviation, which means the variation of using the LDA method for classifying the four days data is the largest. However, CTM and LDA with the TF*IDF methods have the smallest variation. Therefore, the LDA-TF*IDF with pooling schema is more suitable for Twitter data clustering and classification. The further discussion on the result is as follows.

B. LDA-TF*IDF with and without Pooling Improvement

In Section 4.1, we explain the fundamental reason that the LDA-based method does not work well on Twitter short texts. To demonstrate how the pooling methods as we proposed in Section 4.1 and 4.2 improve the LDA topic model, we compare the average topic distribution of the original topics we sampled through Twitter API and which after LDA in the case with and without pooling improvement. Since LDA-TF*IDF is the most accurate method so far, we just compare the results of the LDA-TF*IDF with and without pooling improvement we described in Section 4.1 and 4.2. Results of classification by using LDA-TF*IDF without and with pooling are shown in the following two tables, Table II and Table III.

TABLE II. RESULT OF CLASSIFICATION BY USING LDA-TF*IDF WITHOUT POOLING

LDA-TF*IDF	Social	En'mt	Sport	Tech	Others
Social	<u>0.44^H</u>	0.28	0.07	0.17	0.04
En'mt	0.21	0.32	0.19	0.14	0.14
Sport ^L	0.45	0.18	<u>0.19^L</u>	0.15	0.03
Tech	0.36	0.27	0.10	0.21	0.06
Others	0.30	0.11	0.13	0.06	0.40

TABLE III. RESULT OF CLASSIFICATION BY USING LDA-TF*IDF WITH POOLING

LDA-TF*IDF	Social	Enmt	Sport	Tech	Others
Social	<u>0.81^H</u>	0.04	0.08	0.04	0.03
En'mt	0.21	0.38	0.15	0.12	0.14
Sport ^L	0.28	0.14	<u>0.51^L</u>	0.05	0.02
Tech	0.31	0.27	0.10	0.23	0.09
Others	0.33	0.12	0.13	0.12	0.30

We typically show the result from **highest** precision topics and **lowest** precision topics and see what may cause the difference. The highest precision topic category learnt by the model is denoted as superscript H, while lowest precision topic category is denoted as superscript L. We summarize the results in Table IV-VII.

TABLE IV. TOPIC DISTRIBUTION OF LDA-TF*IDF WITH POOLING AND WITHOUT POOLING

LDA-TF*IDF	Social	En'mt	Sport	Tech	Others
Social ^H	<u>0.44</u>	0.28	0.07	0.17	0.04
pooling-social ^H	<u>0.81</u>	0.04	0.08	0.04	0.03
Sport ^L	0.44	0.18	<u>0.19</u>	0.18	0.01
pooling-sport ^L	0.28	0.14	<u>0.51</u>	0.05	0.02

TABLE V. TOPIC DISTRIBUTION OF LDA MODEL WITH POOLING AND WITHOUT POOLING

LDA-TF*IDF	Social	En'mt	Sport	Tech	Others
Social ^H	<u>0.33</u>	0.15	0.23	0.05	0.24
pooling-social ^H	<u>0.70</u>	0.15	0.08	0.06	0.01
Sport ^L	0.33	0.28	<u>0.19</u>	0.10	0.1
pooling-sport ^L	0.18	0.11	<u>0.36</u>	0.05	0.3

From Table IV, the topic of social has the highest precision (0.44) among all the five topics and the lowest precision (0.19) lies in classifying sport words in topic of sport when pooling is not added. After adding the pooling method, the precisions of corresponding topics increased largely. It shows that the precision greatly improved by using the pooling method. The results of other three days are shown in the following Table V, Table VI and Table VII in next page.

TABLE VI. TOPIC DISTRIBUTION OF LDA MODEL WITH POOLING AND WITHOUT POOLING

LDA-TF*IDF	Social	En'mt	Sport	Tech	Others
En'mt ^H	0	<u>0.52</u>	0.20	0.16	0.12
pooling-en'mt ^H	0	<u>0.71</u>	0.07	0.15	0.07
Social ^L	-	-	-	-	-
pooling-social ^L	-	-	-	-	-

LDA-TF*IDF	Social	En'mt	Sport	Tech	Others
Social ^H	<u>0.53</u>	0.26	0.16	0	0.12
Pooling-social ^H	<u>0.72</u>	0.13	0.08	0	0.07
En'mt ^L	0.28	<u>0.48</u>	0.12	0	0.12
Pooling-en'mt ^L	0.12	<u>0.6</u>	0.2	0.02	0.06

Other interesting phenomena we can see from the data are: 1) Among the popular tweets, the technology tweets compose the least proportion, and then comes the sport and entertainment categories, and the social category has most influenced tweets. 2) During the weekdays, the social category has most tweets, but on Friday, the social category has least tweets that have been retweeted; so people tend to concern social news less during weekend.

To identify the hottest topics under each category, we use the following methodology. We use the retweet number donated f_{dj} as a measurement of the influence of a tweet. Assume we have a collection of m tweets under specific topic. For each word in a tweet, we use $P(T_i)$ represent the probability of a word representing the topic and m represents the number of words in a tweet, then we can represent the influence of the tweets by:

To generate the cloud map, we replicate the word for $P(Ti) * f_{dj}$ times to magnify the importance of the word in the topic. The cloud map generated from April 27, 2015 tweets for all topics are shown in Fig. 2.

Retweet #	Tweets content
531	📍ABC: Baltimore-Police: 15 police officers are injured, 2 are hospitalized after violent protests.
153	Baltimore police won't release report on Freddie-Gray's death #BaltimoreRiots
615	So <i>sad</i> about the <i>earthquake</i> in <i>Nepal</i>
317	📍ABC producers report what appears to be tear gas deployed in Baltimore
94	Penn 3 blocks from White-House. protest in DC for #Freddiegray.

In this paper, we aim to find out a pipeline to semantically identify and graphically visualize the hottest topics discussed on Twitter during a certain period of time. The application can provide a hint for users to notice undergoing big events, popular topics that others might be

interested in. It also provides information for business to make targeted decisions from the obtained popular topics.

In order to achieve this goal, we use a studied pipeline – we sampled Twitter message from “Twitter Top 1000 followers” using Twitter API with result type as “popular” and cluster tweets using unsupervised topic models to identify the topic as well as topic words, finally we use topic words combined with tweet’s influence to graphically visualize the hottest topics during the sampling time.

Regarding on the challenges for the topic model on Twitter, we propose a pooling schema associated to meaningful contexts for a tweet to make the LDA based methods run properly. As a result, this strategy proves to improve the accuracy of the LDA based model for short texts, such as twitter data.

Meanwhile, to identify a reliable pipeline for the purpose, we also conduct comparative studies at every step of the pipeline, compare performance with and without proposed pooling schema in data sampling step, performance with and without TF*IDF in data processing step and performance of two LDA derivative method: Latent Dirichlet allocation (LDA) with Correlated Topic Models (CTM). From the result, LDA –TF*IDF with pooling schema is identified an efficient pipeline in finding hot topics on Twitter data as it shows the best performance in our research.

For the future work, investigations could be conducted from several aspects: First, the method could be tested on larger datasets because the large-scale of data is the key feature of social dataset. Second, the vocabulary-topic parameter of LDA could be trained so that more accurate parameter could be provided at the Gibbs Sampler step. Third, other topic model is worth investigating, for example Label-LDA, where each category is divided into a few related topics, which probably will provide more specific and accurate topic divisions for large-scale tweets dataset.

REFERENCES

- [1] Blei, D. M. and Lafferty, J. D., 2009. Topic models. Text mining: classification, clustering, and application.
- [2] Blei, D. M., 2012. Probabilistic topic models. *Communications of the ACM*, vol. 55, no. 4, pp. 77-84.
- [3] Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- [4] Griffiths, T., Steyvers, M., Blei, D., Tenenbaum, J., 2005. Integrating topics and syntax. *Advances in Neural Information Processing Systems 17*. L. K. Saul, Y. Weiss, and L. Bottou (eds.) MIT Press, Cambridge, pp. 537-544.
- [5] Wallach, H., 2006. Topic modeling: Beyond bag of words. In *Proceedings of the 23rd International Conference on Machine Learning*.
- [6] Blei, D. and Lafferty, J., 2006. Dynamic topic models. In *International Conference on Machine Learning*, ACM, New York, NY, USA, pp. 113-120.
- [7] Teh, Y., Jordan, M., Beal, M., Blei, D., 2006. Hierarchical Dirichlet processes. *By J. Am. Stat. Assoc.* 101, 476, pp. 1566-1581.
- [8] Reisinger, J., Waters, A., Silverthorn, B., Mooney, R., 2010. Spherical topic models. In *International Conference on Machine Learning*.
- [9] Rosen, M., Griffiths, T., Steyvers, M., Smith, P., 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 487-494.
- [10] Chang, J. and Blei, D., 2010. Hierarchical relational models for document networks. In *Ann. Appl. Stat.* 4, 1, pp. 7-10.
- [11] Mimno, D. and McCallum, A., 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*.
- [12] Blei, D. M. and Lafferty, J. D., 2007. A Correlated Topic Model of Science. *The Annals of Applied Statistics*, 1(1), 17-35.
- [13] Grun, D. and Hornik, K., 2011. An R Package for Fitting Topic Models. published by the American Statistical Association.
- [14] Darling, W. D., 2011. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. School of Computer Science University of Guelph, pp. 2-10.
- [15] Daniel, R., et al. 2009. Empirical Methods in Natural Language Processing: EMNLP '09 Proceeding, Volume 1, pp. 248-256.
- [16] Rishabh, M., et al., 2013. Research and development in information retrieval. *Proceedings of the 36th International ACM SIGIR*, pp. 889-892.
- [17] Huang, J. and Malisiewicz, T. 2006 Correlated topic model details. Technical Report, Carnegie Mellon University, http://people.csail.mit.edu/tomasz/papers/huang_ctm_tech_report_2006.pdf