Topic Model analysis and their application

Zhe Hu
Northeastern University
hu.zhe@husky.neu.edu

Introduction

With the development of Internet and large amount of data, people want to express their opinions and acquire information according to their demand more easily. Topic models has gained a large amount of attention and been widely applied in many problems such as text data mining, text classification and topic discovery.

In this project, I analyze three topic models, LDA, Sentiment LDA and Labeled LDA. Also, I test them on different corpus to show their respective advantages and how to be applied in real problems.

Latent Dirichlet Allocation

LDA

LDA is one of the most popular topic models [Blei, Ng, and Jordan, 2003]. LDA is a generative probabilistic model to explain the process that each document is a mixture of a small number of topics and that every word of the document is generated based on the topics that assigned to the document. Different from previous models such as Unigram model or probabilistic latent semantic indexing (pLSI), in LDA we assume each document is assigned with several topics and each word is chosen from the "word bag" by the associated topics of this document. Until now, LDA is still widely used in natural language processing (NLP).

Figure 1 shows the graphical model representation of LDA. Here, the boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document.

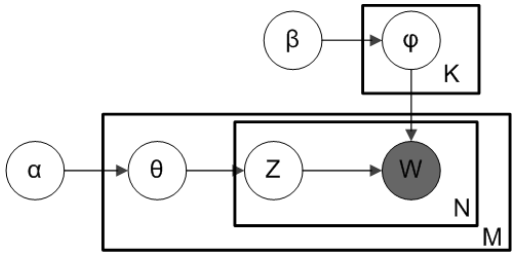

Figure 1. Graphic model of LDA

The generative process of LDA is as follows [Blei, Ng, and Jordan, 2003]:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Now there are some well-developed LDA implementation packages, such as Sklearn, Gensim, etc. In the following experiments, I choose these two python libraries for implementation.

LDA tested on 20newsgroups data

In this experiment, I trained LDA topic model on 20newsgroups data (http://qwone.com/~jason/20Newsgroups). This corpus is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The categorizations of the text data in this corpus are shown in Figure2:

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Figure 2. Categorization of 20newsgroups data

I set the number of topics for LDA training to be 20, and the experiment results are shown in Figure 3:

Figure 3. Topic-word distributed with 20 topics on 20newsgroups data

I select 4 of the topics from total 20 topics with the 20 top words in each topic, shown as figure 4:

Topic 1	Topic 2	Topic 3	Topic 4
armenian , israel , russian , president , armenians , ed , armenia , university , turkish , national , today ,* population , government, turkey , press , arab , soviet, state, republic , history	windows , drive , nt , disk, conference , file, copy , microsoft , win , mouse , help , driver, ms , bit , idea , tell , tried , software , installed , hd	game , team , card , games,drivers, season, players, play, speed, performance, teams , hockey , got , player , win , won , mhz , period , run , memory	image , file , bit , gif , color , available , images , format , files , version , free , ftp , software , data , quality , graphics , programs , display , program , pub

Figure 4. Examples of the results.

From figure 4, we can easily get the conclusion that topic 1 is related to nation, topic 2 is related to computer, topic 3 is related to sports and topic 4 is related to image. The result clearly shows the topics we extracted from the corpus and the corresponded words in each topic.

However, in real life task, we want to extract the topic more detailed with different text data. For example, in 20newsgroups corpus, data in both hockey and baseball are related to sports, and in this case we want to distinguish these 2 topics. Nevertheless, we cannot make it with LDA, even enlarge the number of topics. This is a weakness of LDA model.

Sentiment LDA

Model Introduction

Sentiment analysis is a popular research area with the development of Internet and large text data. In most of work, people focus on the sentiment function of word appearing in the corpus, but the relation between word and topics is seldom considered in the sentiment analysis task. However, in real-life problems, words may have different sentiment under different topics.

Sentiment LDA is an extension topic model which has a latent sentiment label. In general, the sentiment of words is dependent on the domain and topics, and we can consider the sentiment and topic simultaneously [Li, Fangtao, Minlie Huang, and Xiaoyan Zhu, 2010]. In sentiment LDA, there is an extra sentiment layer associated with topic layer, and each word is generated by both topic and sentiment. The graphic model of sentiment LDA is shown in figure 5[[Li, Fangtao, Minlie Huang, and Xiaoyan Zhu, 2010]].

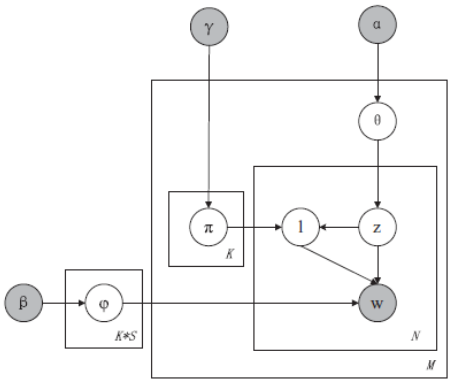

Figure 5. Sentiment LDA.

The generated process of Sentiment LDA is as shown in the following figure:

1. For each document d, choose a distribution θ from $Dir(\alpha)$
2. For each topic z, under document d, choose a distribution $\pi_{d,z}$ from $Beta(\gamma)$
3. For each word w_i in document d
   3.1. Choose a topic z_i from $Multinomial(\theta)$
   3.2. Choose a sentiment label l_i from $Bernoulli(\pi_{d,z_i})$
   3.3. Choose a word w_i from the distribution over words defined by the topic z_i and sentiment label l_i, φ_{z_i,l_i}

Figure 6. Sentiment LDA generation.

Sentiment LDA tested on reviews data

In this section, I did experiment on Sentiment LDA on two corpuses. One is users' reviews from Amazon, IMDB and Yelp from UCI dataset. In the dataset we do not have any information on topics, and sentiment labels are positive and negative. The other one is users' reviews from Amazon, under two topics: "book" and "electronics". The sentiment labels are positive and negative.

In the first experiment, I test sentiment LDA on Amazon, Yelp and IMDB data. I simply set the number of topics to be 1 and the number of sentiment is 2. The result is shown in figure 7.

Topic 1 sentiment 0	Topic 1 Sentiment 1
'movi:0.0294159645866', 'film:0.0269884335285', 'bad:0.0147079822933', 'disappoint:0.00871055262031', 'time:0.00871055262031', 'excel:0.00813936884192', 'charact:0.00771098100814', 'watch:0.00756818506354', 'act:0.00685420534057', 'wast:0.00671140939597', 'make:0.00628302156219', 'worst:0.00571183778381', 'thing:0.00485506211624', 'play:0.00485506211624', 'scene:0.00485506211624', 'problem:0.00456947022705', 'stori:0.00456947022705', 'money:0.00442667428245', 'star:0.00442667428245', 'real:0.00428387833786'	'good:0.0257066547963', 'great:0.0234809703984', 'phone:0.0198085911418', 'work:0.0161362118852', 'like:0.0160249276653', 'place:0.0140218117071', 'food:0.0140218117071', 'servic:0.0122412641887', 'love:0.0103494324505', 'use:0.0103494324505', 'realli:0.00968172713109', 'best:0.00868016915201', 'time:0.00845760071222', 'qualiti:0.00745604273314', 'recommend:0.00712219007345', 'price:0.00623191631427', 'sound:0.00623191631427', 'headset:0.00612063209437', 'product:0.00612063209437', 'nice:0.00600934787447']

Figure 7. Result of experiment 1.

For sentiment 0, there are some words like “bad”, “disappoint”, “problem”. For sentiment 1, there are words like “good”, “great”, “like”. It is not hard to see that sentiment 0 is for negative and sentiment 1 is for positive.

In the second experiment, I tested sentiment LDA on Amazon reviews under “books” and “electronics”. I set the number of topics to be 2, the same as the true number of topics. I set the number of sentiment to be 2. The result is shown in figure 8.

Topic 1 sentiment 0	Topic 1 Sentiment 1	Topic 2 Sentiment 0	Topic 2 Sentiment 1
'book:0.0225031605563', 'stori:0.012389380531', 'read:0.0108000722413', 'charact:0.0101499006682', 'world:0.00646559508759', 'time:0.0063211125158', 'reader:0.00577930287159', 'author:0.00552645837096', 'make:0.00440671843959', 'peopl:0.00408163265306', 'man:0.00368430558064', 'end:0.00325085786527', 'human:0.00317861657938', 'year:0.00310637529348', 'dog:0.00310637529348', 'seri:0.00303413400759', 'life:0.00296189272169', 'view:0.0028896514358', 'american:0.0028896514358', 'histori:0.00285353079285'	'book:0.0346064728306', 'read:0.0152976057005', 'like:0.014874056357', 'good:0.00837132820091', 'work:0.00739965617759', 'time:0.00707576550315', 'know:0.00702593616862', 'realli:0.00695119216683', 'new:0.00677678949598', 'great:0.00627849615068', 'mani:0.0061290081471', 'think:0.00600443481077', 'way:0.00590477614171', 'novel:0.00585494680719', 'make:0.00565562946907', 'want:0.00550614146548', 'say:0.00543139746369', 'love:0.00538156812916', 'look:0.00535665346189', 'peopl:0.00530682412736'	'use:0.026195967122', 'work:0.0215244515404', 'problem:0.0109100585418', 'product:0.0109100585418', 'need:0.0104961267814', 'unit:0.00904736562001', 'card:0.00765773756726', 'time:0.00745077168707', 'batteri:0.00700727337236', 'connect:0.00612027674295', 'purchas:0.00576547809118', 'mous:0.00570634498256', 'comput:0.00511501389628', 'softwar:0.00490804801608', 'drive:0.00484891490746', 'instal:0.00478978179883', 'devic:0.00478978179883', 'month:0.0047306486902', 'button:0.00470108213589', 'tri:0.00470108213589'	'sound:0.0200899439693', 'good:0.0173621350634', 'great:0.0148554998526', 'phone:0.013233559422', 'use:0.0113167207313', 'qualiti:0.0110586847538', 'speaker:0.010284576821', 'like:0.0099896785609', 'ipod:0.0098422943085', 'bought:0.00980536714833', 'cabl:0.0093998820407', 'buy:0.0082940356532', 'better:0.0080359775873', 'headphon:0.00759363019758', 'price:0.00704069595989', 'player:0.00704069595989', 'look:0.00641403715718', 'case:0.00600855204954', 'music:0.00586110291949', 'listen:0.00538189324683']

Figure 8. Result of experiment 2.

The first two columns in figure 8 are for topic 1, and the rest are for topic 2. We can see that in topic 1, it has some words related to books and for topic 2, it has some words related to electronics. Under each topic, there are some positive sentiment words appear in sentiment 1 such as “like”, “great”, “good”. However, there are less negative sentiment words show in sentiment 0. This may happen because in real life, people tend to express their negative feeling with some negative sentence. For example, people would like to use “I don’t like it” than “I dislike it”.

Labeled LDA

Model Introduction

In real life, we want to do some tasks like document tag problem with some known topic labels or word tag problem to associate each word with appropriated tags. Traditional LDA is an unsupervised model which we cannot control the generated topics from the known tags. Labeled LDA is a supervised topic model to constrain the topics to the original topic labels of each text data.

In LLDA, each document is modeled under mixture underlying topics and each word is generated from the related topic. Also, the labeled topics of each document constrains the topic model to use only those topics corresponding to the document observed label.
The graphic model of LLDA is shown in figure 9[Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009, August).].

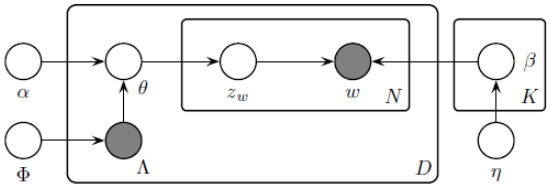

Figure 9. Graphic model of LLDA

The generated process of labeled LDA is as follows:

For each topic $k \in \{1, \ldots, K\}$:
  Generate $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,V})^T \sim \text{Dir}(\cdot|\eta)$
For each document d:
  For each topic $k \in \{1, \ldots, K\}$
    Generate $\Lambda_k^{(d)} \in \{0, 1\} \sim \text{Bernoulli}(\cdot|\Phi_k)$
  Generate $\alpha^{(d)} = L^{(d)} \times \alpha$
  Generate $\theta^{(d)} = (\theta_{l_1}, \ldots, \theta_{l_{M_d}})^T \sim \text{Dir}(\cdot|\alpha^{(d)})$
  For each i in $\{1, \ldots, N_d\}$:
    Generate $z_i \in \{\lambda_1^{(d)}, \ldots, \lambda_{M_d}^{(d)}\} \sim \text{Mult}(\cdot|\theta^{(d)})$
    Generate $w_i \in \{1, \ldots, V\} \sim \text{Mult}(\cdot|\beta_{z_i})$

LLDA test on 20newsgroups data

Here I still use 20newsgroups data for experiment. The number of topics is set to be 20, the same as the true number of topics. I select 6 of the 20 topics in the figure 10:

Label 0 'alt.atheism'	Label 1 'comp.graphics'	Label 3 comp.sys.ibm. pc.hardware	Label 4 comp.sys.mac. hardware	Label 9 'rec.sport.baseball'	Label 10 'rec.sport.hockey'
jesus: 0.0253 god: 0.0230 matthew: 0.0134 people: 0.0127 believe: 0.0102 bible: 0.0083 said: 0.0077 prophecy: 0.0075 evidence: 0.0065 exist: 0.0061 messiah: 0.0058 theism: 0.0056 faith: 0.0054 true: 0.0052 belief: 0.0052 isaiah: 0.0050 david: 0.0050 islam: 0.0050 psalm: 0.0050 king: 0.0046	jpeg: 0.0378 image: 0.0221 file: 0.0179 gif: 0.0163 images: 0.0119 color: 0.0114 bit: 0.0114 format: 0.0109 available: 0.0097 files: 0.0093 graphics: 0.0091 data: 0.0087 version: 0.0084 software: 0.0080 ftp: 0.0079 quality: 0.0073 edu: 0.0071 free: 0.0066 programs: 0.0065 use: 0.0064	os: 0.0256 edu: 0.0252 scsi: 0.0196 card: 0.0147 drive: 0.0147 mb: 0.0147 com: 0.0147 mhz: 0.0138 bus: 0.0118 dx: 0.0102 comp: 0.0100 controller: 0.0098 irq: 0.0098 ide: 0.0087 dos: 0.0085 isa: 0.0078 use: 0.0071 chip: 0.0065 port: 0.0060 board: 0.0058	mac: 0.0233 mhz: 0.0230 drive: 0.0203 apple: 0.0193 card: 0.0107 cpu: 0.0090 operational: 0.0090 problem: 0.0090 iisi: 0.0087 mb: 0.0087 simms: 0.0087 thanks: 0.0083 software: 0.0080 hardware: 0.0077 memory: 0.0077 monitor: 0.0077 cable: 0.0073 bit: 0.0073 cache: 0.0073 fpu: 0.0073	year: 0.0230 game: 0.0178 team: 0.0125 baseball: 0.0125 hit: 0.0117 better: 0.0117 won: 0.0100 players: 0.0089 think: 0.0086 league: 0.0086 player: 0.0080 morris: 0.0075 stats: 0.0072 games: 0.0069 good: 0.0069 lost: 0.0067 season: 0.0067 win: 0.0061 hitter: 0.0061 average: 0.0058	game: 0.0145 team: 0.0129 hockey: 0.0115 det: 0.0112 tor: 0.0087 nyr: 0.0083 games: 0.0081 chi: 0.0076 nhl: 0.0073 period: 0.0071 la: 0.0071 bos: 0.0069 shots: 0.0069 stl: 0.0064 mtl: 0.0062 season: 0.0062 van: 0.0060 blues: 0.0058 players: 0.0055 play: 0.0053

Figure 10. 6 topics and word distribution with LLDA.

Since LLDA is a supervised topic model, it will constrain the topic to the original topic set. From the results we can see it has the better results than LDA since the topics keep the same as the

label. This could be applied in many applications such as document tag or credit attribution.

Conclusion and Future Work

In this project, I compared three different topic models, LDA, Sentiment LDA and Labeled LDA, and compare their performance through experiments. LDA is a powerful topic model which has been widely used in many domains such as natural language processing and computer vision. Sentiment LDA adds a latent sentiment label on LDA model and can extract the sentiment information. LLDA is a supervised model which can constrain the topics to the original labels, and has been widely used in document tag and credit attribution.

REFERENCES

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[2] Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009.

[3] Li, Wenbo, Le Sun, and Da-Kun Zhang. "Text classification based on labeled-LDA model." CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION- 31.4 (2008): 620.

[4] Li, Fangtao, Minlie Huang, and Xiaoyan Zhu. "Sentiment Analysis with Global Topics and Local Dependency." AAAI. Vol. 10. 2010.

[5] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.

[6] Lin, Chenghua, and Yulan He. "Joint sentiment/topic model for sentiment analysis." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[7] Mcauliffe, Jon D., and David M. Blei. "Supervised topic models." Advances in neural information processing systems. 2008.

[8] Gensim Document. https://radimrehurek.com/gensim/models/ldamodel.html

[9] Krestel, Ralf, Peter Fankhauser, and Wolfgang Nejdl. "Latent dirichlet allocation for tag recommendation." Proceedings of the third ACM conference on Recommender systems. ACM, 2009.

[10] Wikipedia LDA. https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation