Topic Model: Development, Evaluation and Application

Pengyang Zhou
Baskin School of Engineering, University of California, Santa Cruz

Abstract

Topic models can help people find general topics behind articles and text based on probabilistic knowledge and machine learning technology. Among topic modeling methodology, the Latent Dirichlet Allocation (LDA) is the most effective one to analyze discrete data from a large volume of collections. LDA uses a three-level hierarchical Bayesian model to find the hidden themes of texts and abstracts the main topics for an article without human involvement. It can be easily embedded in other models to solve complicated problems and has been applied to a wide range of fields, such as bioinformatics, recommendation systems, text classification, and social network analysis.

Introduction

Recently, we find it more difficult to filter useful data due to the rapid increase of information. When searching online, keywords are the main method we use when searching online. However, human interventions are still needed to narrow the scope in a large volume of results. Therefore, researchers have made significant findings considering the problem of abstracting latent topics from texts and other collections of data. In this poster, we investigated the development of topic models and evaluated Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and LDA. We also explored the application of LDA and its future directions.

Development

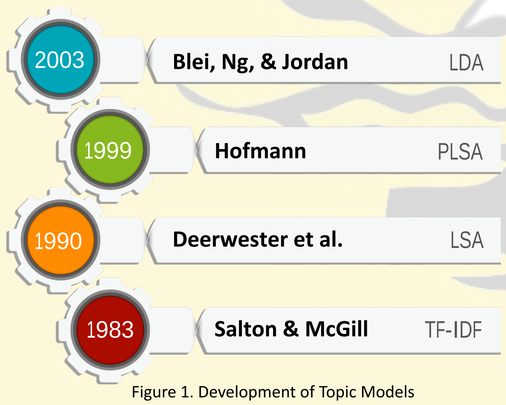

Figure 1. Development of Topic Models

Evaluation

LSA

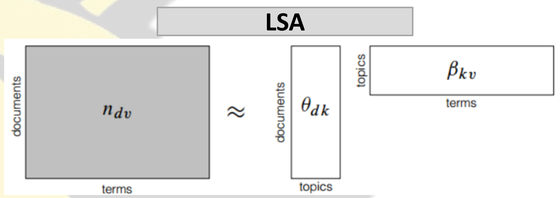

Figure 2. Graphic model of LSA

- Use a singular value decomposition of the X matrix to identify a linear subspace in the space of TFIDF scores
- Achieve significant compression in large collections
- **Limitations**: LSA cannot capture polysemy

PLSA

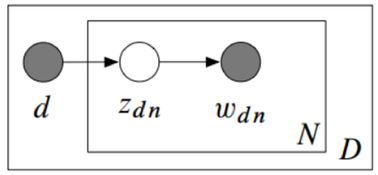

Figure 3. Graphic model of PLSA

- A probabilistic model based on the main idea of LSA
- Define a topic as a distribution over terms
- Describe each document as a distribution over topics
- Learn these two sets of parameters with EM
- **Limitations:**

– Parameters grows linearly with size of the corpus, resulting in overfitting

– PLSA cannot perform well for documents outside of the training set

LDA

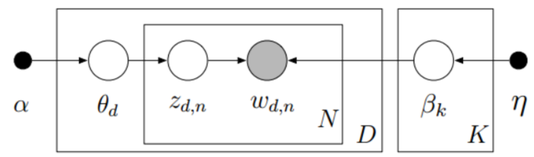

Figure 4. Graphic model of LDA

- A generalization of PLSA(the Bayesian version)
- Three levels: Corpus-level, Document-level, Word-level.
- Two goals:
  - In each document, allocate its words to few topics.
  - In each topic, assign high probability to few terms.
- **Advantages:**
  - Generalize easily to new documents
  - Parameters do not grow with the size of the training corpus, avoiding overfitting issues.

Application

LDA can be applied to various fields, like finding find patterns in genetic data, images, and social networks，document modeling, text classification, recommendation system, Bioinformatics, content-based image retrieval.

Application 1: Analysis of "Reuters News" using LDA model in Python:

the most probable topic of a document

```
for n in range(10):
    topic_most_pr = doc_topic[n].argmax()
    print('doc: {} topic: {}\n{}...'.format(n,
                                            topic_most_pr,
                                            titles[n][:100]))
```

```
doc: 0 topic: 8
0 UK: Prince Charles spearheads British royal revolution. LONDON 1996-0
doc: 1 topic: 13
1 GERMANY: Historic Dresden church rising from WW2 ashes. DRESDEN, Germ
doc: 2 topic: 14
2 INDIA: Mother Teresa's condition said still unstable. CALCUTTA 1996-0
doc: 3 topic: 8
3 UK: Palace warns British weekly over Charles pictures. LONDON 1996-08
doc: 4 topic: 14
4 INDIA: Mother Teresa, slightly stronger, blesses nuns. CALCUTTA 1996-
doc: 5 topic: 14
5 INDIA: Mother Teresa's condition unchanged, thousands pray. CALCUTTA
doc: 6 topic: 14
6 INDIA: Mother Teresa shows signs of strength, blesses nuns. CALCUTTA
doc: 7 topic: 14
7 INDIA: Mother Teresa's condition improves, many pray. CALCUTTA, India
doc: 8 topic: 14
8 INDIA: Mother Teresa improves, nuns pray for "miracle". CALCUTTA 1996
doc: 9 topic: 8
9 UK: Charles under fire over prospect of Queen Camilla. LONDON 1996-08
```

the top 5 words of each topic

```
n = 5
for i, topic_dist in enumerate(topic_word):
    topic_words = np.array(vocab)[np.argsort(topic_dist)][:-(n+1):-1]
    print('Topic {}\n- {}'.format(i, ' '.join(topic_words)))
```

```
Topic 0
- british churchill sale million major
Topic 1
- church government political country state
Topic 2
- elvis king fans presley life
Topic 3
- yeltsin russian russia president kremlin
Topic 4
- pope vatican paul john surgery
Topic 5
- family funeral police miami versace
Topic 6
- simpson former years court president
Topic 7
- order mother successor election nuns
Topic 8
- charles prince diana royal king
Topic 9
- film french france against bardot
```

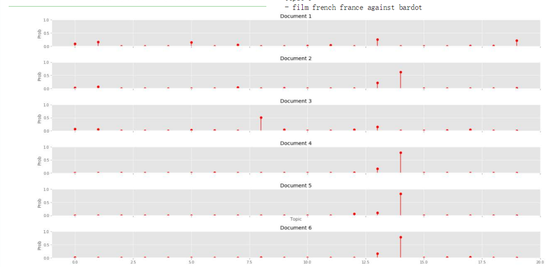

Figure 5,6,7. Some analysis of words, topic and probability distributions

Application 2： Visualizing data in Bioinformatics

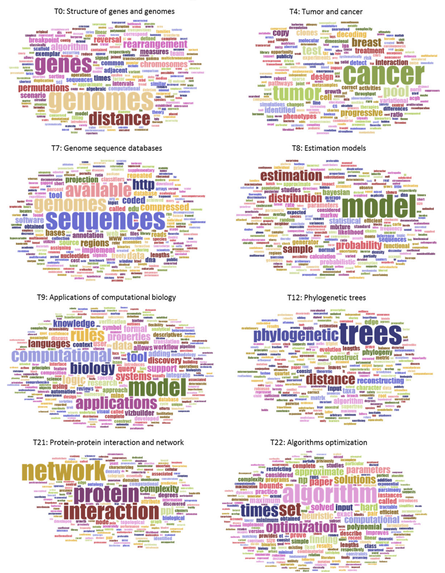

Figure 8. "Eight example topics obtained by LDA modeling with 40 topics on TCBB dataset"(Zhao et al., 2015)

Discussion

LDA is a simple and appealing model, but it still contains some assumptions and limitations.

LIMITATIONS

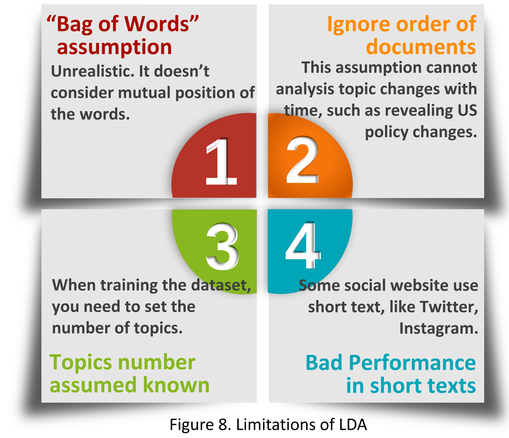

Figure 8. Limitations of LDA

Conclusion

Topic models provide us methods to uncover the hidden patterns of a collection of documents, to annotate these documents with discovered patterns and then helps us summarize and understand the texts.

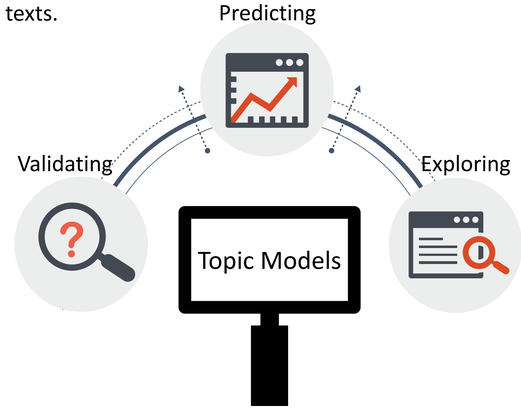

Figure 9. Topic Models usages

Future Direction

Advanced outcomes in machine learning provide us with new research directions to further develop topic models, like model performance checking, visualization and data discovery.

Contact Information

Pengyang Zhou
Baskin School of Engineering, UCSC
Email: pzhou15@ucsc.edu
Website: https://www.linkedin.com/in/pengyangzhoua1/
Phone: (650)238-7912

References

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
2. Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.
3. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.
4. Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc.
5. lda: Topic modeling with latent Dirichlet Allocation — lda documentation. (n.d.). Retrieved from https://lda.readthedocs.io/en/latest/index.html
6. Salton, G., & McGill, J. M. (1983). Introduction to modern information retrieval. New York: McGraw Hill Book Co.
7. Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015, December). A heuristic approach to determine an appropriate number of topics in topic modeling. In BMC bioinformatics (Vol. 16, No. 13, p. S8). BioMed Central.

Acknowledgements

I would like to thank Keri Toma and all other advisors for helping us in the Graduate Preparation Program and providing us this poster fair to reflect what we learned. I am grateful to Sarah Tomlinson, Shannon Sinclair and Shannon Cummings for their dedicated work and excellent teaching skills.