EVERYTHING I KNOW ABOUT LATENT DIRICHLET ALLOCATION

I. A beginner’s guide to LDA

Source: <https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2>

Author: Ria Kulshreshtha (twitter, Software Engineer at Google)

- A statistical model for discovering abstract topics. Aka topic modelling

1. What is topic modeling?

- A method of unsupervised learning, multiple classification of documents

- Similar to clustering on numeric data

2. Why topic modeling?

- Automatically organize, understand, search, and summarize large electronic archives (or corpus)

- It can help with the following:

+ Discover hidden themes in the collection

+ Classify documents into the discovered themes

+ Using classification to organize/summarize/search documents

3. LDA

- One of the most popular topic modeling methods

- The goal of LDA: To find topics a document belongs to, based on the words in it.

- Background:

+ Each document is made up of various words

+ Each topic has multiple words belonging to it.

4. Finding representative words for a topic:

- We can sort the words with respect to probability score, then take the top n words with the highest appearance rate.

- We can set a threshold. Any word that passes this threshold will be a representative for its topic

5. Assumptions:

- Order of words and grammatical role are not considered

- Eliminate stop words

- Number of topics, k, is pre-decided

- All topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how documents are generated (Q: I don’t quite understand this)

6. How does LDA work?

- There are 2 parts in LDA:

+ The words that belong to a document (what we know)

+ The words that belong to a topic or the probability of words belonging into a topic (what we need to calculate)

7. The algorithm to find the above

- Go through each document and randomly assign each word in the document to one of k topics.

- For each document d, go through each word w and compute:

+ p(topic t | document d): The proportion of words in document d that are assigned to topic t. Excluding current word. If this value is large, it is more probable that word w belongs to t.

+ p(word w | topic t): Proportion of assignments to topic t across all documents that come from word w. Tries to capture how many documents are in topic t because of word w.

Q: Is the p() here probability or proportion?

- LDA represents documents as a mixture of topics. A topic is a mixture of words. If a word has high probability of being in a topic, all the documents having *w*will be more strongly associated with *t*as well.

- Update the probability for the word w belonging to topic t as:

p(word w with topic t) = p(topic t | document d) \* p(word w | topic t)

Q: Needs clarification on why this is indeed correct

II. Introduction to Latent Dirichlet Allocation (another blog)

Link: <http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/>

Author: Edwin Chen – Founder of Surge AI, the world’s most powerful data labeling platform and workforce for NLP

III. Latent Dirichlet Allocation (more technical)

Link: <https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation>

Notes:

- There is another thing to do in the pre-process part: Need to convert terms, words to their “root” lexical forms – e.g., “barks”, “barking”, “barked” will be converted to “bark”.

POSSIBLE PACKAGES TO USE AND EXAMPLES

I. topicmodels, tidytext, stringr, scales, and other common ones (ggplot2, dplyr, tidyr)

Link: <https://www.tidytextmining.com/topicmodeling.html>

- A detailed example for reference.

II. tidytext + textmineR

Link: <https://medium.com/swlh/topic-modeling-in-r-with-tidytext-and-textminer-package-latent-dirichlet-allocation-764f4483be73>

- Example not detailed, only throws in the code for it to work

III. lda

Link: <https://www.rdocumentation.org/packages/lda/versions/1.4.2/topics/lda-package>

<https://cran.r-project.org/web/packages/lda/lda.pdf>

<https://rdrr.io/cran/topicmodels/man/lda.html>

- Little information, no example to work with

Conclusion: I propose we proceed with topicmodels, as there is a detailed example for use to look at. Furthermore, the package was last updated in 2021, while lda’s latest update was from 2015.

REFERENCES

FOR “lda” PACKAGE:

1. Blei D.M., Ng A.Y., Jordan M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Link: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>

2. Phan X.H., Nguyen L.M., Horguchi S. (2008). Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In Proceedings of the 17th International World Wide Web Conference (WWW 2008), pages 91–100, Beijing, China.

3. Lu, B., Ott, M., Cardie, C., Tsou, B.K. (2011). Multi-aspect Sentiment Analysis with Topic Models. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, pages 81–88.

FOR TOPICMODELS PACKAGE:

4. topicmodels: An R Package for Fitting Topic Models

Link: <https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf>