CS410 Technology Review
A Quick Survey of LDA Topic Model Testing and Evaluation

John Hanratty
johnh7@illinois.edu

Abstract

Topic discovery and modeling with Latent Dirichlet Allocation (LDA) has gained popularity as a data mining tool but presents some challenges compared to other machine learning tasks. Since the primary task is discovery and definition of topics within a corpus, a tagged, ground truth data set for comparing results does not exist. Validating model output requires some creative thinking.[1] This paper explores some options to help improve model predictive capability, aid topic discovery and qualify the final model. The genesis for this discussion was the desire to expand and provide better context than existing trade press articles [4, 5] about topic model evaluation to support my final project.

Introduction

Autonomous topic model discovery and validation is not an option today for a variety of reasons[1,6]. First, a topic definition and relevance closely tie to a unique target application. For example, a topic definition that is appropriate for categorizing biology papers will not work for machine learning papers. Second, since the goal is to discover and define topics, an authoritative data set with known topic results is not available for testing. Third, independently determining topics in a document test set is a labor intensive, application specific, subjective process performed by humans. Humans will often disagree on topics covered by a document and miss relevant or hidden topics. Fourth, popular automated metrics (e.g., perplexity, coherence), although useful, have shown negative correlation to topic relevance.[5, 6] This paper provides a brief overview of options.

"While the models make different assumptions, inference algorithms for all of these topic models build the same type of latent space: a collection of topics for the corpus and a collection of topic proportions for each of its documents. While this common latent space has explored for over two decades, its interpretability remains unmeasured."[1]

Model Development Phases

Topic development entails three phases.

(1) Model predictive characterization,
(2) Topic discovery and optimization,
(3) Final performance validation.

As a project progresses through these phases, the process becomes less automated and requires more human judgement. These phases overlap and are iterative depending on validation data and project requirements.

Phase 1: Model Predictive Characterization

The initial model development phase requires a "smoke test" for the base model and data set, which drives initial data set preprocessing techniques (e.g., word stemming or bigram extraction) and model hyper parameter settings. The goal is to determine how well the model represents the raw data input and extracts plausible topics from the corpus. Since 'ground truth' data doesn't exist, the usual metrics for accuracy, precision and recall are not calculable. Most practitioners use perplexity and coherence calculations for this process. This initial stage produces the model's initial configuration and parameters.

Perplexity

The consensus in literature is that "*Optimizing for perplexity may not yield human interpretable topics.*"[4] Although the metric is automated and readily available in machine learning toolkits, the value is questionable.[6]

Coherence

Coherence provides a better option for initial testing but has limitations.[6] Coherence calculates a score by measuring the degree of semantic similarity between high value words in a topic.[4] Coherence enables an initial sanity check for a topic model and data set preprocessing but has some limitations:[6] (a) the training and test corpora must match closely, (b) the results do not provide variance information, so outliers might influence results, (c) models with a different number of topics are not comparable, and (d) it does not account for the number junk topics produced. Coherence metrics come in several flavors that address various strengths and shortcomings. These include[7]: C_v, C_p, C_uci, C_umass, Cmpmi, and C_a.

Phase 2: Topic Discovery and Optimization

Topic Discovery entails a manual and iterative process for topic investigation, refinement, and modeling. This analysis studies the topic model's output and tunes parameters to improve the topic relevance for the target application. LDA optimization parameters include number of topics (K), topic density per document (alpha), prior word-to-topic probabilities (eta), stop words, and document preprocessing. For example, a higher alpha value results in a more specific topic distribution per document, and a higher beta results in a more specific word distribution per topic. The bottom line is that "discovering topics is an unsupervised process. There is no gold-standard list of topics to compare against for every corpus".[4] So, model analysis and tuning is a manual and subjective process.

Model output analysis drives the tuning process. The Gensim LDA model[11] provides the following outputs:

- Topics with probability and coherence,
- Terms for each topic with probability

- Matrix of terms to topics with probability
- Documents for each topic with probability
- Topics within each document with probability
- Topics that for a particular word with probability
- Overall model coherence and perplexity

Visualization helps the discovery process. High dimensional data presents a challenge for graphing. Nevertheless, graphing the data can expose useful information. Bar charts, heatmaps and word clouds are routinely used. PyLDAvis and Termite are two off-the-shelf visualization tools for topic model development. They provide developers with innovative topic analyses and visual representations.

pyLDAvis

The pyLDAvis application (web site) provides "*a novel method for choosing which terms to present to a user to aid in the task of topic interpretation, in which we define the relevance of a term to a topic*".[10] Rather than ranking terms purely by probability they allow user to explore topic-term relationships using their innovative 'relevance' metric to interpret model performance. This metric "*accounts for the degree to which a term appears in that particular topic to the exclusion of others."* The GUI enables the data scientist to view topic and term relationships.

This tool provides a valuable view of model performance. A couple hints were mentioned in a talk by Matti Lyra[6]. The circle sizes are used to indicate the relevance magnitude are misleading since circle size grows as square root of magnitude. So, smaller differences in magnitude are hard to detect visually. The second is that comparison graphs use PCoA to reduce the dimensionality so are difficult to interpret. These limitations are not unique to pyLDAvis nor do they detract from many truly useful features.

Termite

The description from their paper[2] (web site): "Termite, a visual analysis tool for assessing topic model quality. Termite uses a tabular layout to promote comparison of terms both within and across latent topics." "Termite allows analysts to identify coherent and significant themes." The team presents two innovative metrics, "saliency" and "seriation," that help reveal clustering structure and highlight related terms.

One warning on this tool is the installation difficulty. It requires some work to set up the web server and tool.

Phase 3: Final Performance Validation

Topic model validation requires both model unit testing and system testing as a part of the target application. In the end, performance of the target application is the most important and relevant measure. Unfortunately, the ability to modify and measure a production application is often expensive, complex, or unavailable. A plan is required to collect and analyze the model's

impact on application performance. This activity measures the ultimate payback for the development project.

Unit testing verifies that the model reliably identifies topics contained in documents outside the training data set. This task requires human judgement. Testing includes a combination of "eyeballing" the results and checking results against a "gold standard" data set. The first would enlist a team to manually validate topics predicted by the model against source documents. The second would enlist a team to tag documents for automated validation of the model. This has the advantage of reuse but requires dilligent maintenance as topics are discovered or morph.

The challenge is that humans will classify the same document differently based on expertise, point-of-view, fatigue, or other factors. Chang, et. al.[1] proposes techniques to improve human model evaluation to judge relevance: *Topic Intrusion* and *Word Intrusion*. The team is presented examples with "extra" topics or words and must choose the 'intruder.' Their research shows improvement in matching topic use cases. This technique could improve results of both eyeballing and tagging activities.

Conclusion

This paper provides a survey of the topic model development, testing and validation practices. Topic definition and evaluation are manual, subjective processes during discovery, test, and production phases of model development. Unlike other machine learning developments, topic models do not have tagged data sets for validation. The fact that topics are unique for each application also introduces challenges. Users will have different views of a topic based on knowledge, application experience or other factors. There is plenty of room for new ideas for making the process more efficient and productive.

References

[1] **Reading Tea Leaves: How Humans Interpret Topic Models**
Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, David M. Blei,
https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

[2] **Termite: Visualization Techniques for Assessing Textual Topic Models**
Jason Chuang, Christopher D. Manning, Jeffrey Heer
Advanced Visual Interfaces, 2012
PDF (2.3 MB) | Website | Software]

[3] **Causal Support: Modeling Causal Inferences with Visualizations**
Alex Kale, Yifan Wu, Jessica Hullman
IEEE Trans. Visualization & Comp. Graphics (Proc. VIS), 2022
PDF | Supplement | Preregistration (1) | Preregistration (2) | Best Paper Honorable Mention

[4] **Evaluate Topic Models: Latent Dirichlet Allocation (LDA)**
Towards Data Science, Aug 19, 2019
Shashank Kapadia
https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[5] **Topic Model Evaluation**
Giri
Natural Language Processing High Demand Skills, updated June 22, 2021
https://highdemandskills.com/topic-model-evaluation/

[6] **Evaluating Topic Models**
Matti Lyra
Pydata Conference 2019 Berlin Presentation, Video
https://www.youtube.com/watch?v=UkmIljRIG_M

[7] **Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus.**
Mifrah, Sara & Benlahmar, EL Habib. (2020). International Journal of Advanced Trends in Computer Science and Engineering. 10.30534/ijatcse/2020/231942020.

[8] **Exploring the Space of Topic Coherence Measures**
Michael Röder, Andreas Both, Alexander Hinneburg
WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining
February 2015

[9] **LDA Alpha and Beta Parameters - The Intuition**
Blog Vector the Thought Vector blog, October 22, 2015
https://www.thoughtvector.io/blog/lda-alpha-and-beta-parameters-the-intuition/

[10] **LDAvis: A method for visualizing and interpreting topics**
Carson Sievert, Kenneth Shirley
Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces
June 2014
https://aclanthology.org/W14-3110/

[11] **models.ldamodel – Latent Dirichlet Allocation**
Datasheet
https://radimrehurek.com/gensim/models/ldamodel.html