Data Analysis Data Science List Question

What are good ways of evaluating the topics generated by running LDA on a corpus?

Latent Dirichlet Allocation

7 Answers

William Chen, Data Scientist at Quora
Written Dec 12, 2013 · Upvoted by Melissa Dalis, Data Scientist at Uber

One reason LDA is easier to explain and teach than other hierarchical ML models is that **you can usually tell a story about the generated topics.**

You can do a hard assignment of each word to its most likely topic and examine the words in each topic. If your topics are good, you should be able to tell those stories.

For example, in the original Biel, Ng, and Jordan paper, the authors were able to attach labels to their topics that corresponded well with the words in the topic.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

With the stories, you have not only confidence that your topics at capturing real

Source - http://wv

Brian Feeny, Harvard Grad Student
Written Dec 13, 2013 · Upvoted by William Chen

I just went through this exercise. We looked at almost 1M reviews and used LDA to build a model with 75 topics. I found no better way to truly evaluate the topics, rather than having humans look at them and see if they made sense.

There is a lot of work that needs to go into getting to the point where your topic's make sense. It starts at the beginning, with data quality:

1. Are you allowing words to go into LDA that would not belong in describing topics?

In our model we only used nouns (treebank NN and NNS). Part of speech tagging is important to building most topic models. At a minimum you should stop common stop words (the, a, it, etc). You should also make sure that you don't allow very high frequency words to overpower the rest of the corpus. You likely don't want very infrequent words either. In our model we removed any words that were in 5 or less documents and any word that appeared in more than 60% of documents.

We used techniques like spell correction and lemmatization to further aggregate words and reduce the dimension of the data. This may or may not be appropriate for your particular model.

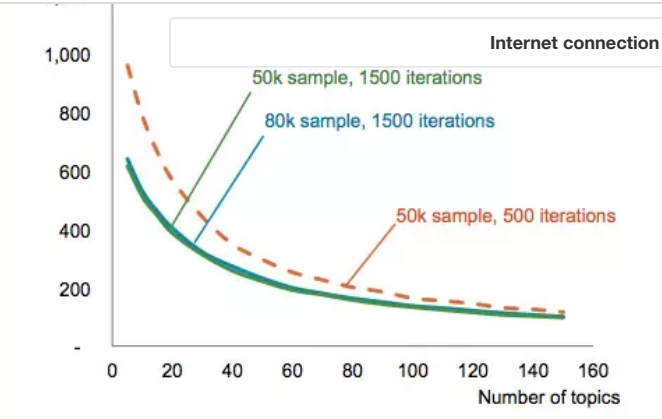
2. Are you running LDA long enough?

There are may ways to evaluate if LDA is running long enough. You can look at things like topic differences or topic drift. This can help you understand the convergence. These types of methods require time to settle in, so you will get poor topics unless you allow proper convergence. Depending on the size of the data, and the amount of resources you have, this could take minutes, hours, days, or even longer!

3. Did you select the proper number of topics?

It's very easy to pick too few or too many topics. One method for evaluating the number of topics is to use a grid search and evaluate perplexity of the model on hold out data. We did this, but there is no substitute for actually running a few different models with different number of topics and actually have humans evaluate it. You could use Mechanical Turk if you have a lot of data/topics. In my situation we settled on 75 topics and it took each of the four people in my group a few hours each to look at the top 10 words in each topic.

Because this is unsupervised learning, you can't just arbitrarily set a number of topics. Its similar to K-means clustering, you have to find the ideal number of centroids. And in similar ways, there are methods to help you do this, but its not perfect.



4. Do the density of words in the topics make sense?

If you have topics with weak/low densities for its constituent words, then its likely a weak topic. Once you classify documents, and you give the topics names, pull up the documents that strongly correlate to that topic. Read them, does the topic make sense?

5. Does the word have multiple meanings? Do you need to add some context?

Many words can have multiple meanings or contexts. In some ways you can "help" LDA by using bigrams in addition to normal unigrams. You may use adjective/noun pairs, or just nouns, verb/nouns. You can use chunking to look at verb phrases or noun phrases only.

6. Are your expectations too high?

There are likely to be some hard to identify topics. The frequency distributions insist they are a topic, but it may be weak. You can classify the topic name as "Ambiguous" or "Unknown". I think with our 75 topics on almost 1M reviews, we had maybe 7 or maybe 10 at most topics that were very weak.

Some things that can really screw up topics are words of sentiment. Some people like to use the same words over and over in their writing, such as the word "great". I have seen techniques that limit the use of some words by almost building a "set" of words per document. I personally don't like this approach. I have also seen TF-IDF transform done on the corpus to give different weights to words. I experimented with this myself, but it does not hold true to traditional LDA to use this approach (I am not saying its not worth looking into).

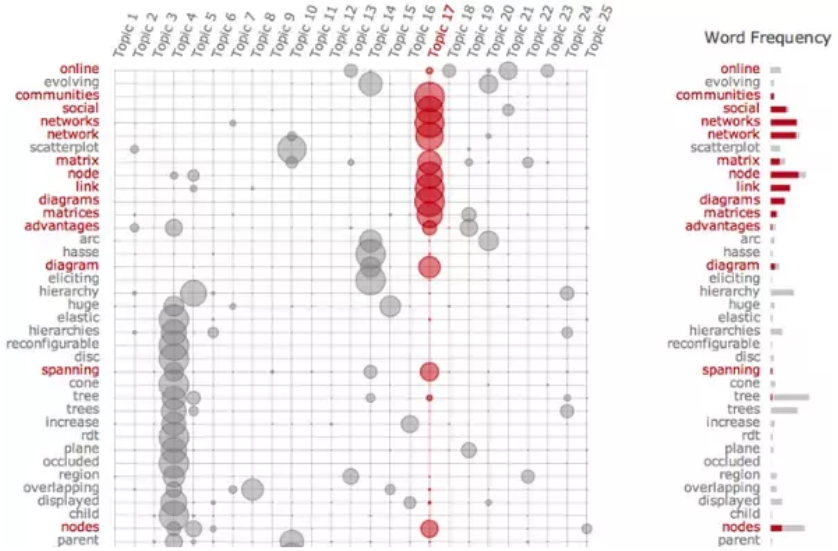
David Mimno has done some interesting work in LDA, here is an interesting approach to building models where you have fixed topic-word probability distributions: Page on Cornell

Joel Chan, Doing it full-time
Written Dec 12, 2013 · Upvoted by William Chen, Data Scientist at Quora

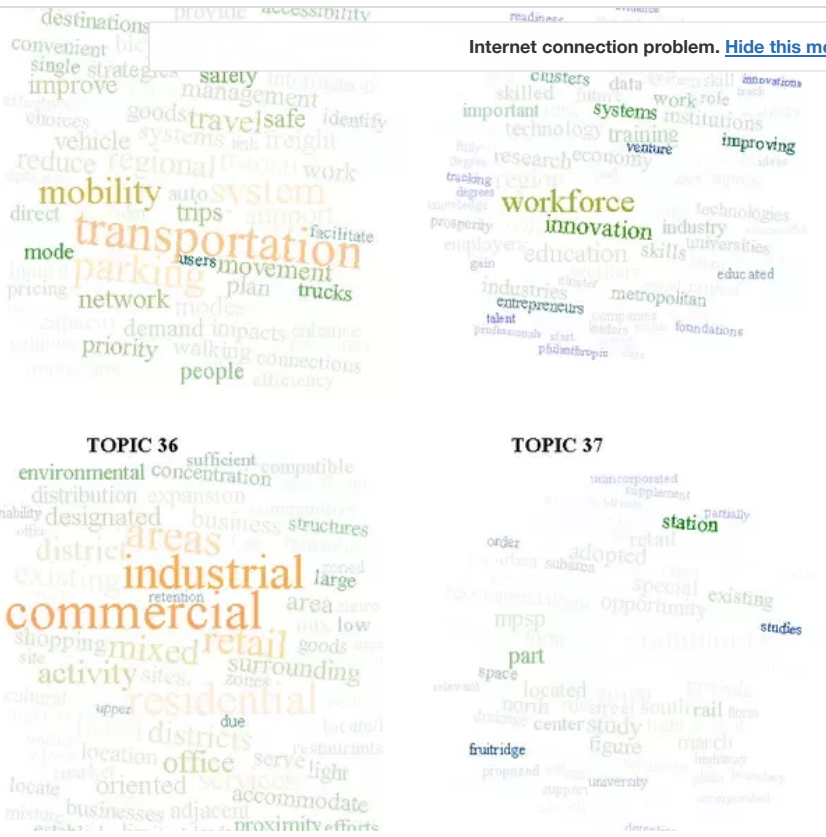
This is an important question. Topic models are often evaluated using internal metrics, such as held-out probability. But, as people in David Blei's group have shown, this can be a poor indicator of the quality of the topics, at least as measured by human judges: Page on Wustl

So, as William Chen notes, it's important to look at the topics themselves, and see if they tell a coherent and sensible story. I'll add that it's important to not just consider the top N words, but also consider the distributions of words within and across topics, as Ben Schmidt argues in this excellent paper on using topic modeling as a discovery/research tool: Words Alone: Dismantling Topic Models in the Humanities

But looking at lots of words in a list can be hard. So, Jason Chuang (originally at Stanford, now at U Washington) co-developed this to help: Termite: Visualization Techniques for Assessing Textual Topic Models



You can also use word clouds to visualize the top N words, and represent words that are more or less informative (e.g., words that are uniquely associated with certain topics) in an intuitive fashion: Using Word Clouds for Topic Modeling Results



Finally, David Mimmo, one of Blei's former postdocs, has also developed an alternative internal metric (taking advantage of co-occurence information that isn't used by the model) that is able to predict whether you are likely to produce "bad" topics (e.g., incoherent topics, "fused" topics, topic with "intrusions"): Page on Umass

4.5k Views · View Upvotes

Upvote 14 Downvote

Add a comment... Recommended All

Alex Blocker, Statistician at Google, PhD in statistics from Harvard, computational biologist much of the time.
Written Mar 22, 2014

I've found predictive and large-scale human evaluations persuasive for LDA and related models. Predictive evaluations test how well the topic scores predict another feature of interest. Classification is the most common task, like the Reuters comparisons from the original 2003 paper. I find these evaluations persuasive for the task of interest but narrow in their applicability.

The core paper on human evaluation is Chang et al. (2009) (Page on papers.nips.cc). The authors used Amazon Mechanical Turk to evaluate the coherence of their topics. They use very clever but straightforward word and topic "intrusion" tasks for these, asking raters to identify discordant words or sets of words. These ratings seem less persuasive than predictive checks to me, but they can capture abstract topic quality in a way other evaluations struggle to.

Perhaps Sean Gerrish can comment on this, as he was an author the paper I mentioned?

Related Questions

In latent dirichlet allocation (LDA), how can one map real topic meanings to algorithm topics?

How does Quora generate its topic taxonomy?

Topic Models: How do I transform the LDA model to see the topic evolution in chat content?

What are good ways to determine user's age and gender?

What are good master thesis topics for applying text mining/information retrieval on scientific literature corpus?

Natural Language Processing: When computing similarity between documents using LDA, (how) can I give more "weight" to certain topics in the co...

What are the pros and cons of LDA and NMF in topic modeling?

What is a good way of classifying text documents against an arbitrary topic model?

Sean Gerri
What's unclear from the original poster's question is *what* the topics are being evalu...

Trent Niemeyer, Currently working on analysis of narrative travel blogs
Written Sep 18, 2015

To add to Joel Chan's answer there is a helpful library called LDAVis that carries on the work from the Stanford Research (Page on stanford.edu). I've been using the python port pyLDAvis (https://pyldavis.readthedocs.org...) within a Jupyter project along with gensim LDA (https://radimrehurek.com/gensim/...).

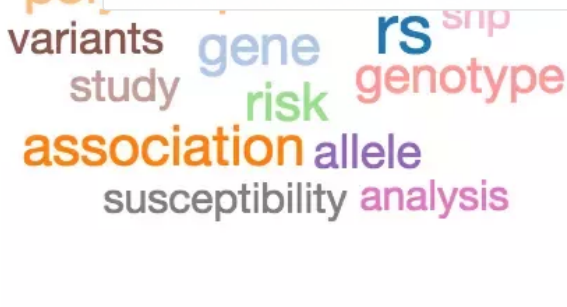
I believe the original LDAVis was written for R.

Here a sample Jupyter notebook for pyLDAVis
http://nbviewer.ipython.org/gith...

Priya Desai, data scientist, love learning new technologies
Written Oct 10, 2016

At Stanford we build ran LDA on 2 years worth of Pubmed data- (ie titles and abstracts) and settled on 150 topics (http://scireader.org). Since we were 'categorizing' biomedical research, and all our documents were paper abstracts, we used a couple of methods to 'evaluate' our topics. see my blog (Topic Modeling: LDA... LSI.. Oh My!

- find journals that are very category specific eg: 'Stroke', 'Hypertension', 'Thrombosis' etc, and feed articles from those journals and check to see if the topic probability distribution is what you expect.
- adding to the above: sum up the probabilities from all the articles in that specific journal and see if the topic distribution for that journal is consistent with your expectations
- pick well known scientists in specific areas- create the 'topic distribution' for their published papers - does it make sense?
- Since the topic names were human curated, we created word clouds to help our users understand what we meant by that topic. for eg: for the topic we called 'Genetic associations" we had this word cloud (font size proportional to word probability)



and “blood cancers” is :

1.1k Views · View Upvotes

Upvote 1 Downvote

Add a comment... Recommended All

Charles H Martin, Calculation Consulting; we predict things
Written Dec 13, 2013

easy. label the documents by the topics learned in LDA, and run a multi-class SVM classifier on the data set. the SVM training accuracy is your evaluation

2.4k Views

Upvote Downvote

Add a comment... Recommended All

Top Stories from Your Feed

Gene Spafford wrote this · 2016

What is it like to be rejected by Harvard, Yale, Stanford, MIT, Caltech, Princeton, Cornell, and UPenn simultaneously?

Gene Spafford, Prof for 30 years, and student for another 8.
Written Feb 20, 2016 · Upvoted by Alan Wu, studied at Massachusetts Institute of Technology and Jake Januzelli, B.A

Anne Hunter wrote this · 8h

Does the fact that I am from Azerbaijan increase my chances of being accepted to MIT?

Anne Hunter, Undergraduate Academic Administrator for >30 years in EECS.
Written 8h ago

I disagree with Chris Lee that you need to make any case that Azerbaijiani

Mary Lou Jepsen wrote this · Mar 1

What are the lessons learned from one laptop per child (OLPC) in terms of innovation & development?

Mary Lou Jepsen, former CTO and co-founder at One Laptop Per Child (2005-2008)
Written Mar 1

There are a variety of ways to feel. You can

good fit for MIT is all you need

DONE! I was the co-founder (heavy