LDA Metrics:

Bug localization using latent Dirichlet allocation -- No LDA Eval -- Use Case Study

------------------------------------------------------------------

Improving Reliability of Latent Dirichlet Allocation by

Assessing Its Stability Using Clustering Techniques on

Replicated Runs

**5 stars**

Cluster analysis and similarity calculation

The importance measure is intuitive, because it scores words high which occur often

in the present topic, but less often in average in all other topics.

stability measure S-CLOP

We introduce a novel algorithm for assessing the stability of LDA by calculating pairwise

similarities of replicated runs and quantifying similarity of sets of runs with our new

measure S-CLOP

------------------------------------------------------------------

A guided latent Dirichlet allocation

approach to investigate real-time latent

topics of Twitter data during Hurricane

Laura

human

experts observed the automatically generated topics and manually assigned and organised category labels for them

------------------------------------------------------------------

Leveraging Latent Dirichlet Allocation in processing free-text personal

goals among patients undergoing bladder cancer surgery

Overview of LDA analytic plan

# of topics,

Not very related.

------------------------------------------------------------------

Understanding Text Pre-Processing for Latent Dirichlet Allocation

quantitative comparisons

of different pre-processing treatments that

might not be directly comparable.

Section:

Bad Models vs. Bad Representations

------------------------------------------------------------------

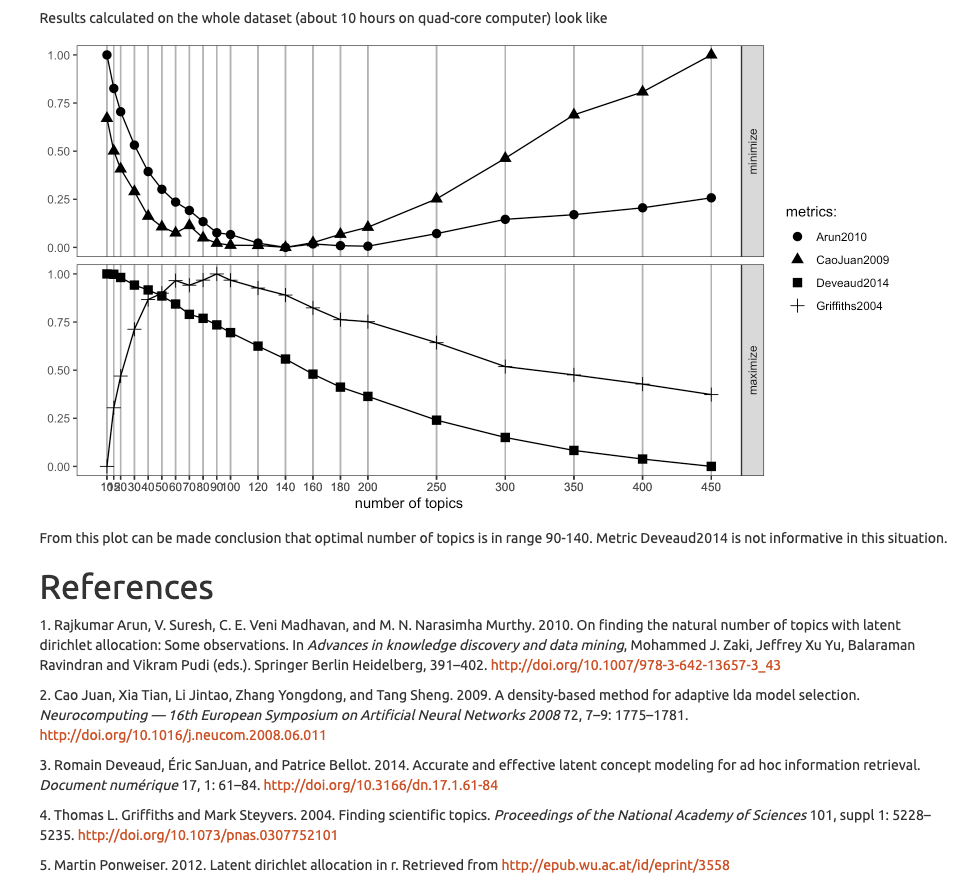
Lda tuning

4 metrics

<https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html>

<https://github.com/nikita-moor/ldatuning>

<https://zhuanlan.zhihu.com/p/105226228>



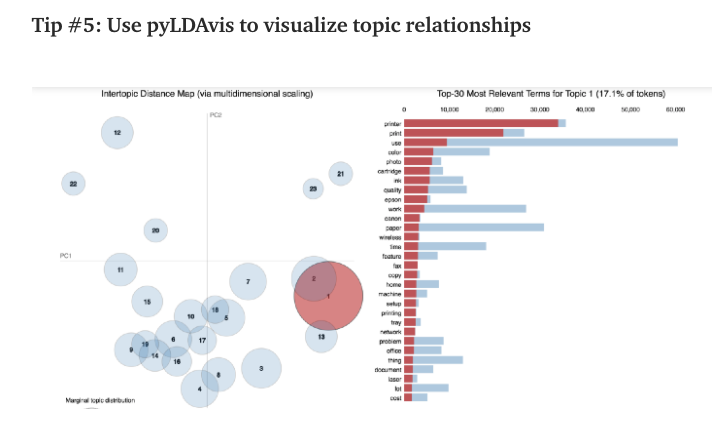
LDA is an unsupervised topic modeling algorithm that attempts to describe a set of observations (documents) as a mixture of different categories (topics). The “per-word log-likelihood” (PWLL) metric measures the likelihood that a learned set of topics (an LDA model) accurately describes a test document dataset. Larger values of PWLL indicate that the test data is more likely to be described by the LDA model.

<https://docs.aws.amazon.com/sagemaker/latest/dg/lda-tuning.html>

LDA requires specifying the number of topics. We can tune this through optimization of measures such as predictive likelihood, perplexity, and **coherence**

<https://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf>

[pyLDAvis](https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf)



Package to visualize the distance between topics

<https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2>

<https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2>

<https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html>

<https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/>

* Most important hyperparameters?
  + Preprocessing: max\_df, min\_df, min\_tf, max\_tf
  + # of topics k： based on prior knowledge
  + Alpha
  + Beta
  + Max\_iteration affects??? Ask kexin on LDA

Log Likelihood

Perplexity On a different note, perplexity might not be the best measure to evaluate topic models because it doesn’t consider the context and semantic associations between words.This can be captured using topic coherence measure, an example of this is described in the gensim tutorial I mentioned earlier.

**Hyperparameter optimization**

Hyperband

<https://jmlr.org/papers/volume18/16-558/16-558.pdf>

Github: <https://github.com/liamcli/hyperopt>

<https://github.com/zygmuntz/hyperband>

<https://github.com/thuijskens/scikit-hyperband>

<https://gist.github.com/PetrochukM/2c5fae9daf0529ed589018c6353c9f7b>

<https://github.com/thegaussians/hyperband-for-any-model/blob/master/hyperband_demo.ipynb>

<http://hyperopt.github.io/hyperopt/>

<https://github.com/liamcli/hyperopt>

LDA also (semi-secretly) takes the parameters alpha and beta. Think of alpha as the parameter that tells LDA how many topics each document should be generated from. beta is the parameter that tells LDA how many topics each word should be in. You can play with these and you may get better results.

However, LDA is an unsupervised model, and even the perfect settings for k, alpha, and beta will result in some incorrectly assigned documents. If your data isn't preprocessed well, it almost doesn't matter what you assign the parameters, it will always produce poor results.

Online LDA

<https://zhuanlan.zhihu.com/p/105226228>