Implementing and evaluating an LDA model

Sandeep Avula

asandeep@unc.edu

What is Topic Modeling?

- Topic modeling looks to automatically discover the hidden thematic structure in a large corpus of text documents.

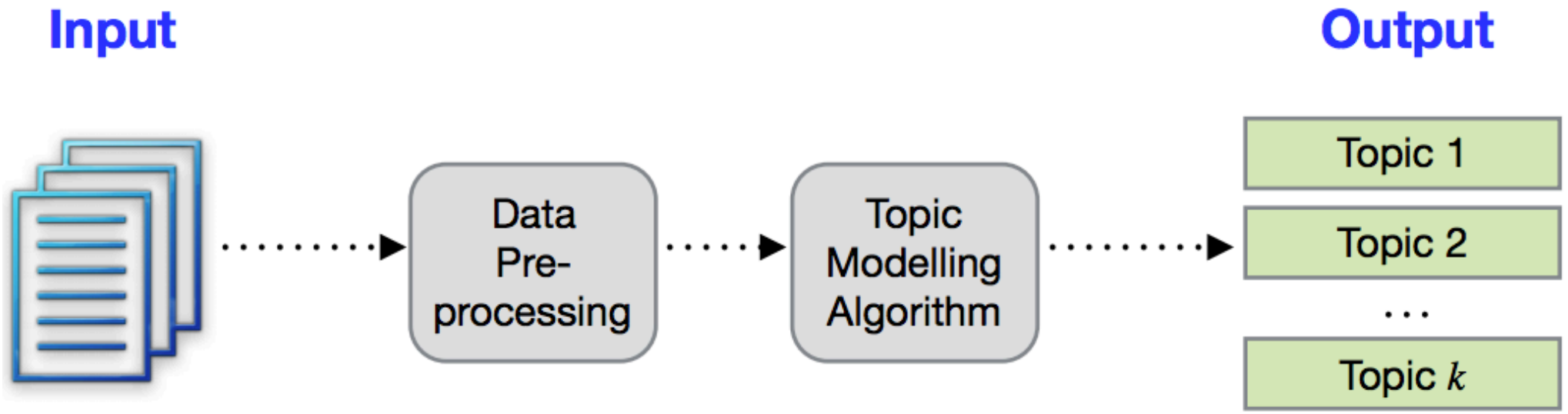
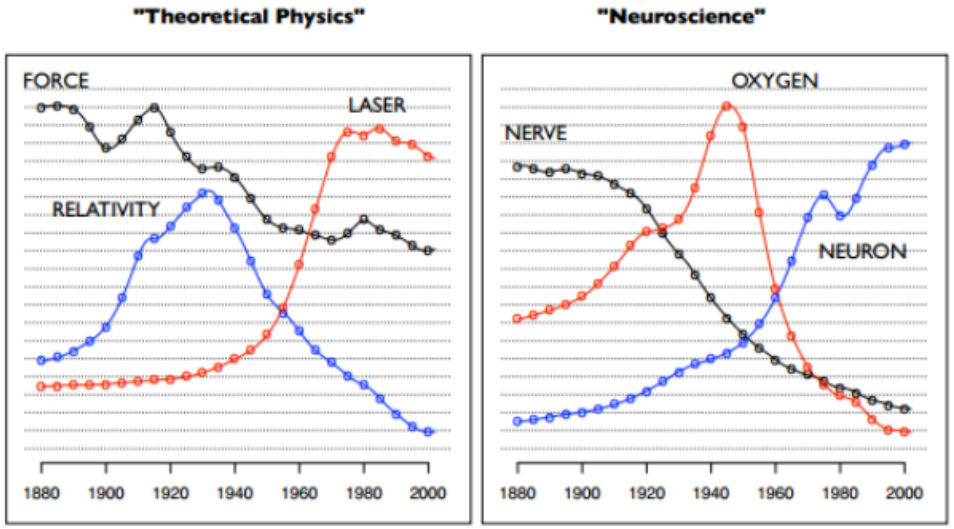


Image source: http://derekgreene.com/slides/topic-modelling-with-scikitlearn.pdf

Applications

Annotate images



Understand the evolution of topics

Image source: http://derekgreene.com/slides/topic-modelling-with-scikitlearn.pdf

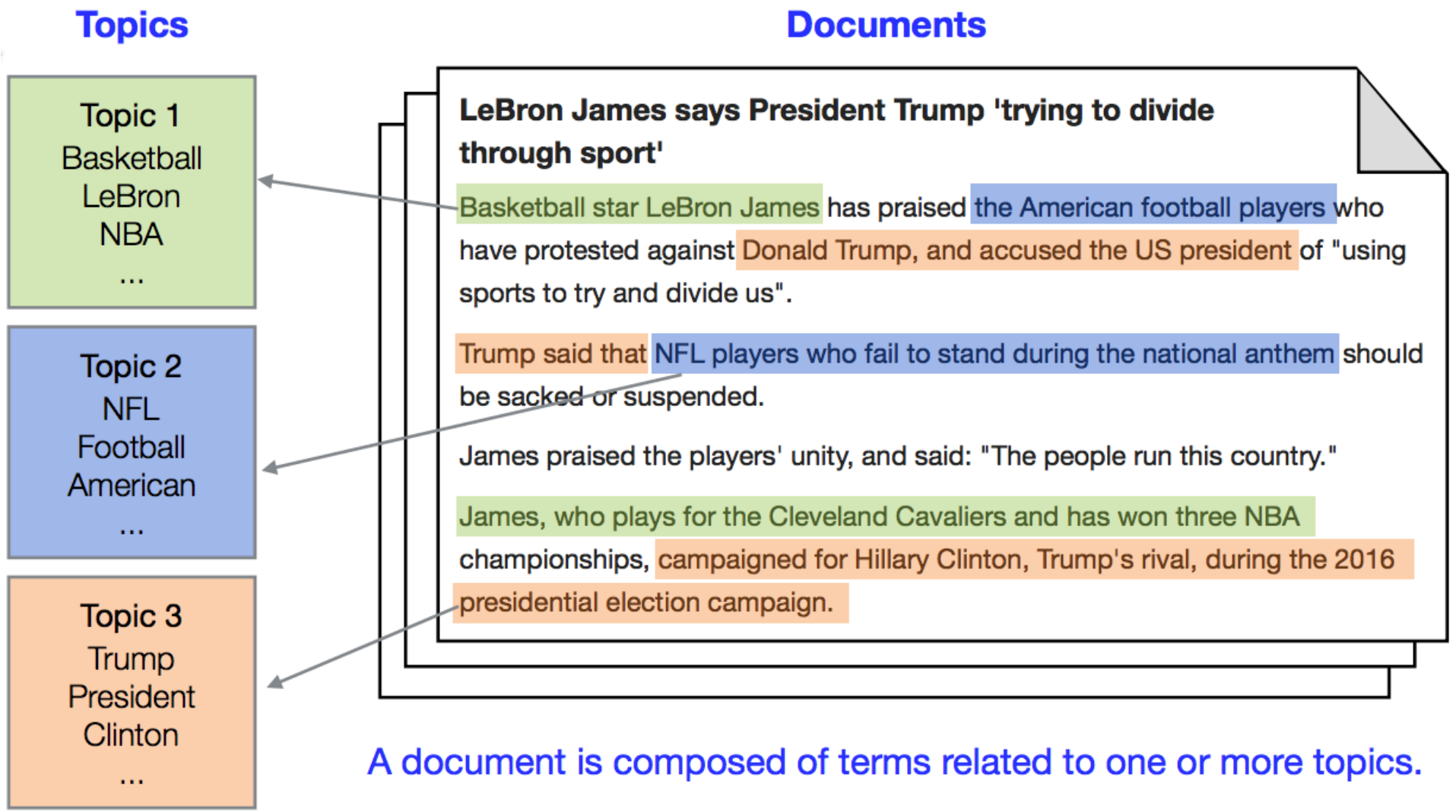
But what are these topics?

- They are ***latent*** and learned by the model.

- Latent variables are also referred to as “hidden” variables and they are inferred rather than being given.

- How can you think about this? Have you ever watched a movie and found it difficult to find the right vocabulary to explain your preference?
  - Think of latent topics in a similar manner.

Latent topics in a document



A document is composed of terms related to one or more topics.

Approaches to finding Latent Topics

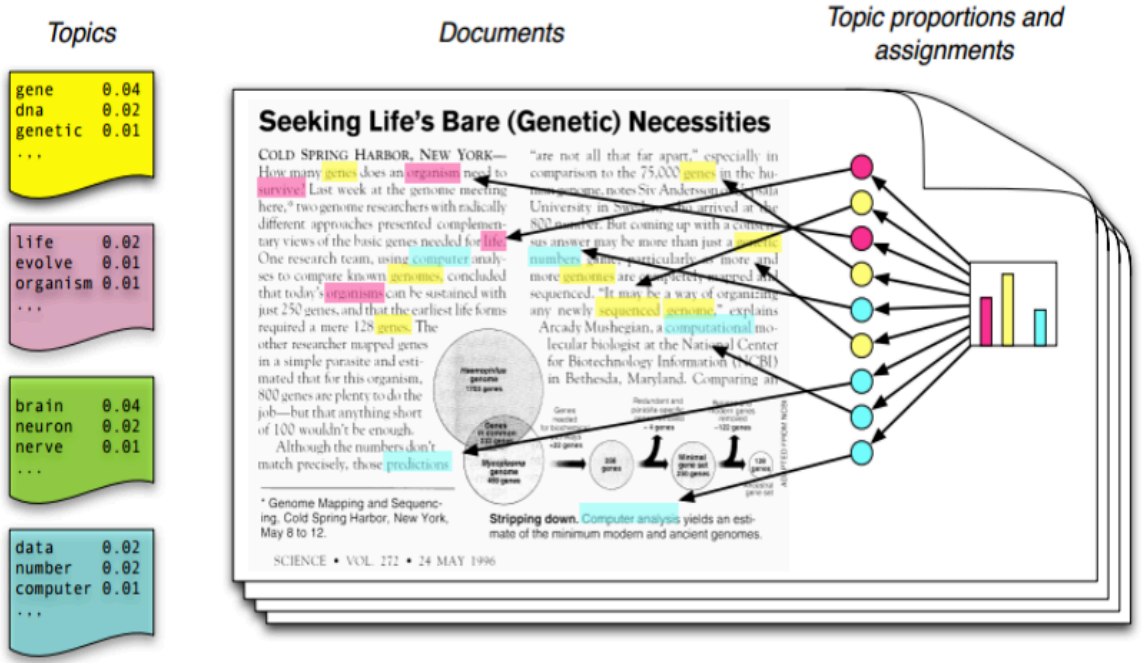
- Non-Negative Matrix factorization.
- Latent Dirichelt Allocation (LDA)
- Probabilistic Latent Semantic Indexing (pLSI)
- Correlated Topic Model (CTM)

Approaches to finding Latent Topics

- Non-Negative Matrix factorization.
- **Latent Dirichelt Allocation (LDA)**
- Probabilistic Latent Semantic Indexing (pLSI)
- Correlated Topic Model (CTM)

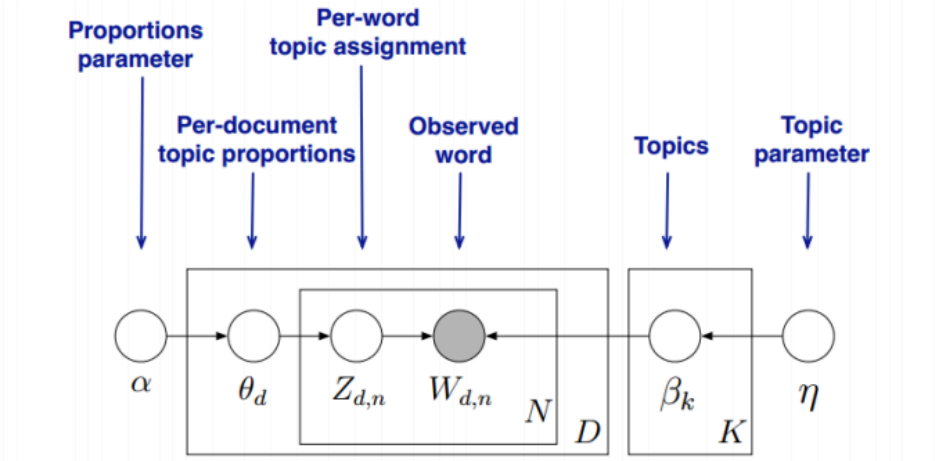
LDA

- What does it do?
  - Detect latent topics.
  - Provide a distribution of a topic over words.
  - Provide a distribution of topics over documents.



LDA

- How does it learn?
  - Statiscal inference.
  - Each topic is picked from a dirichlet distribution.
  - For each document: the topic is first first based on a proportionality parameter.
  - Each words with the document and topic is picked from a multinomial distribution.



$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta)\right)\left(\prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n})\right)$$

Practical exercise

- Switch on your jupyter notebook’s or Anaconda.

- Link to code: <after_class>

How to evaluate LDA topics?

- Automatic vs. Manual.

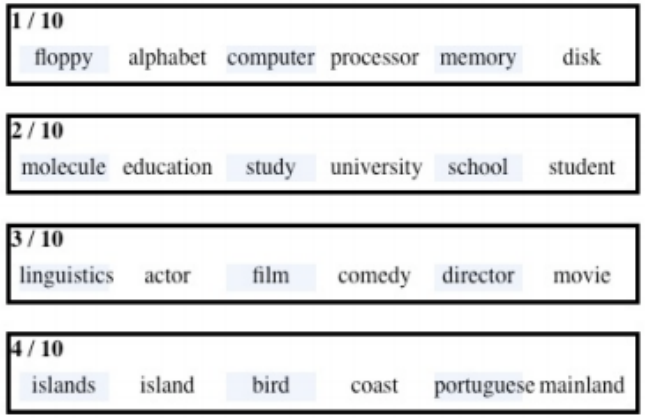
- Automatic:
  - Log-likelihood ratio as a score (This explains how well the parameters we have chosen can explain the data)
  - Perplexity

- Manual: (Chang et al.(2009))
  - Word intrusion
  - Topic intrusion

Our focus: Manual

- Word intrusion.

- What to do?
  - Pick a topic and the top 5 words in that topic.
  - Next, pick another word in the bottom list which is a top word in another topic.
  - We now have 6 words and the task is to present these 6 words to a person and ask them to pick the odd one.

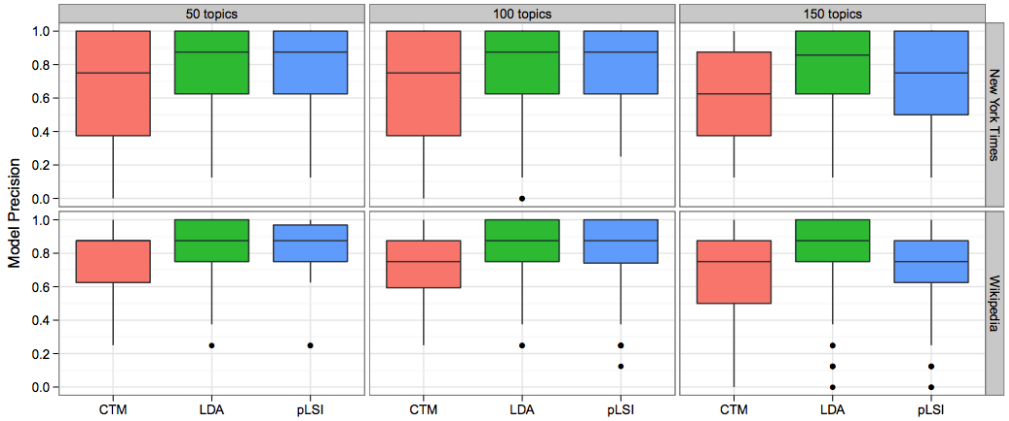


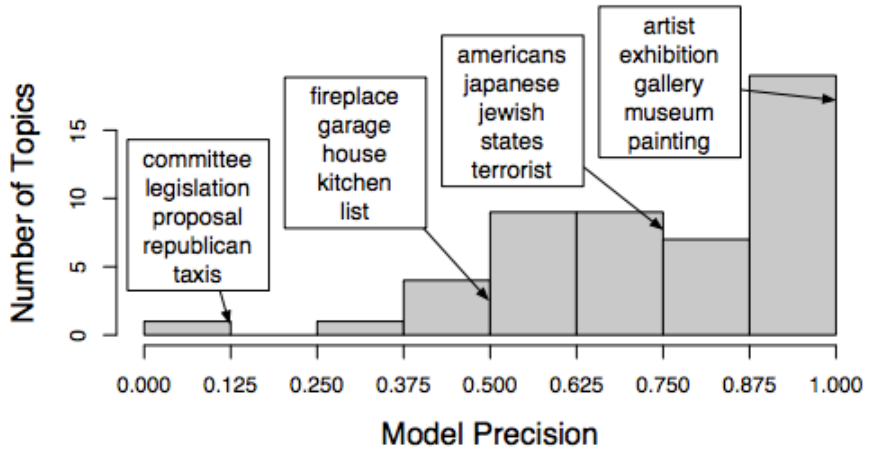
Guess the intruder?

Our focus: Manual

- Word intrusion
  - Metric: Model precision (MP).
  - For the k^{th} topic the Model Precision (MP_k) generated by S subjects for an intruder variable i is as follows:

$$MP_k = (\sum_s \mathbb{1}(i_{k,s} = w_k))/S$$

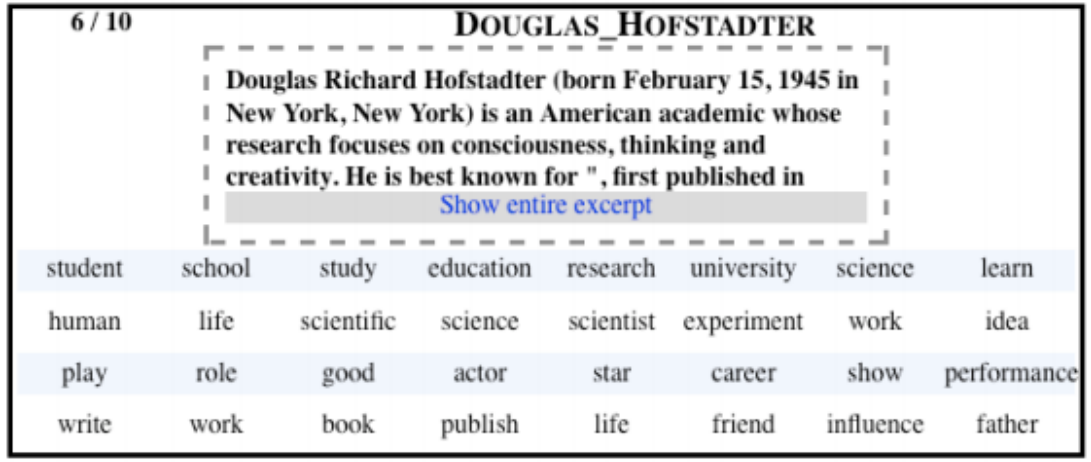




Our focus: Manual

- Topic intrusion.

- What to do?
  - First, pick the top 3 most probable topics.
  - Next, from the low probability topics, randomly pick one.
  - Show subjects a small snippet of the document and ask them to pick the intruder.



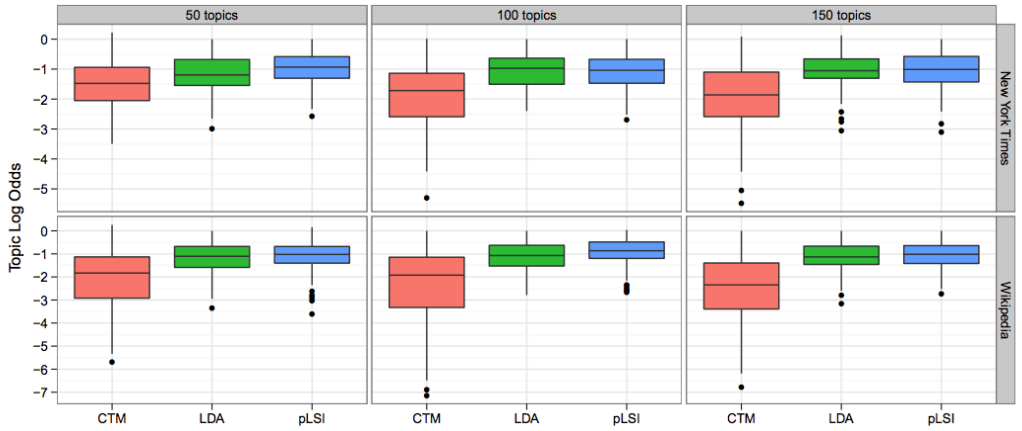
Guess the intruder?

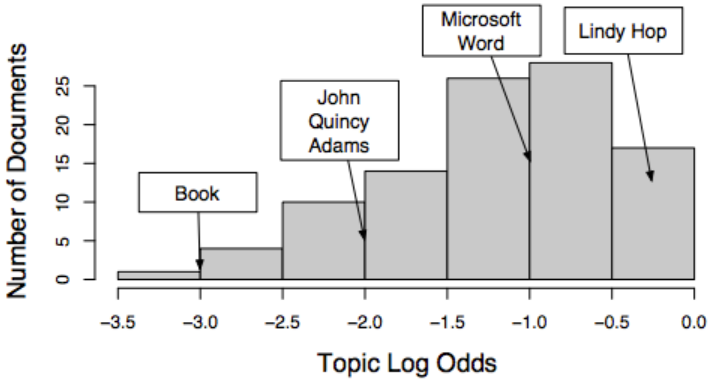
Our focus: Manual

- Topic intrusion.
  - Metric: Topic log odds (TLO)
  - For a topic d, TLO is measured as the difference in log odds ration between the intruder topics estimate for that topic and the one selected by a subject *s*.

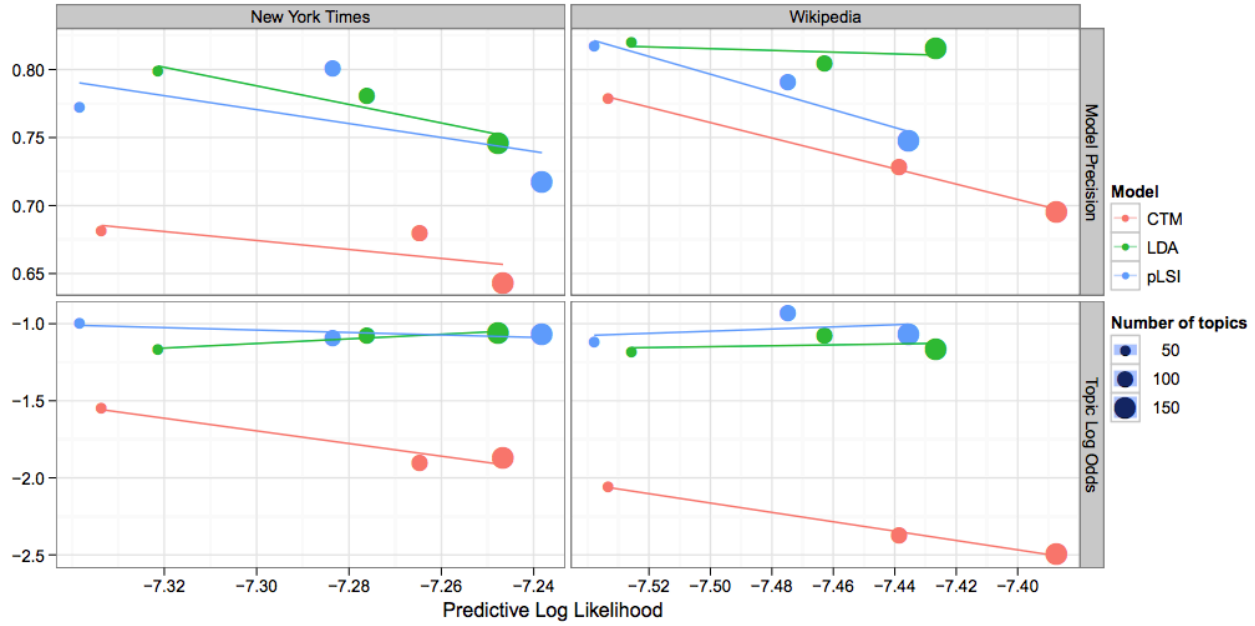
$$TLO_d = (\sum_s (log\hat{\theta}_{d j_{d,*}} - log\hat{\theta}_{d j_{d,s}}))/S$$

Intruder Selected by subject





Which method of evaluation do you chose?



Model Precision and Topic Log Odds are not directly correlated with log-likelihood ratio values.
Suggestion: For interpretability use Model Precision and Total Log odds ratio.

Relevant readings:

- http://derekgreene.com/slides/topic-modelling-with-scikitlearn.pdf
- https://nateaff.com/2017/09/11/lego-topic-models/
- https://cfss.uchicago.edu/fall2016/text02.html
- http://chrisstrelioff.ws/sandbox/2014/11/13/getting_started_with_latent_dirichlet_allocation_in_python.html
- https://dzone.com/articles/python-scikit-learnlda
- https://nateaff.com/2017/09/11/lego-topic-models/
- https://ece.umd.edu/~smiran/LDA.pdf
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*(pp. 288-296).