Topic modeling

Literature overview

The input

- **22.287** abstracts - 15676 journal, 3086 papers, 2418 conference, 241 book chapters
- Average document length ranging from **140** to **200** words

Potential issues:

- Heavy tailed word distribution
- Sparse document-topic distribution (i.e. 1 topic per document)
- Correlated documents and topics (e.g. follow-up works,...)
- Evolving topics over time evolving vocabulary
- Considering abstracts as short texts?

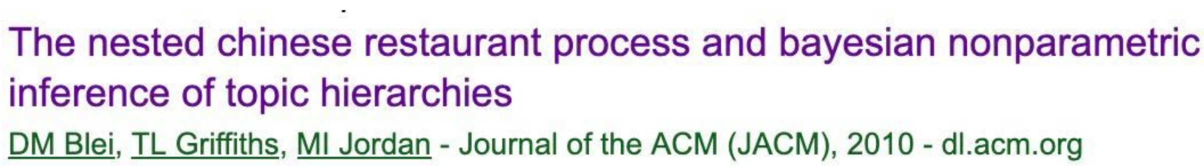
Our goals

Desiderata of the final topic model (in order of priority):

- High performance (topic coherence, perplexity, qualitative assessment)
- Hierarchical model
- Automatic topic name inference
- Lifelong learning model, i.e. the ability to incorporate new knowledge into the model while retaining the old knowledge
- Exploiting the information contained in the metadata
- Modeling temporal evolution

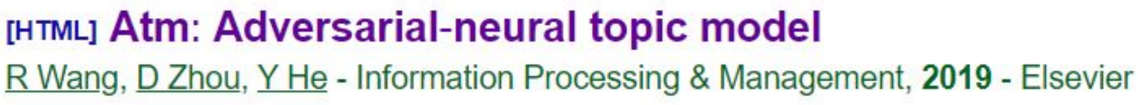
The papers

Topic modeling literature



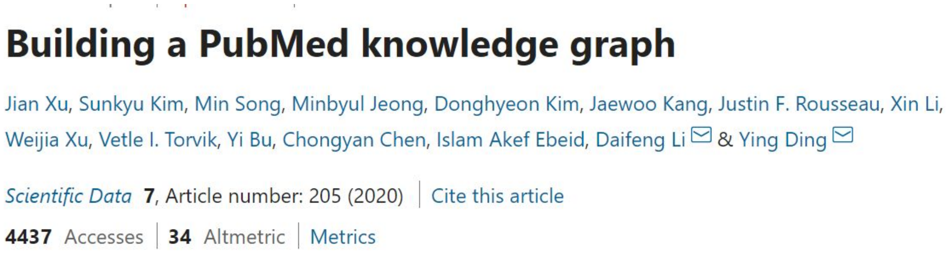
The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies

DM Blei, TL Griffiths, MI Jordan - Journal of the ACM (JACM), 2010 - dl.acm.org



[HTML] Atm: Adversarial-neural topic model

R Wang, D Zhou, Y He - Information Processing & Management, **2019** - Elsevier

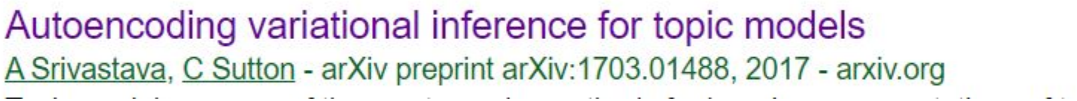


Building a PubMed knowledge graph

Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F. Rousseau, Xin Li, Weijia Xu, Vetle I. Torvik, Yi Bu, Chongyan Chen, Islam Akef Ebeid, Daifeng Li ✉ & Ying Ding ✉

Scientific Data **7**, Article number: 205 (2020) | Cite this article

4437 Accesses | **34** Altmetric | Metrics



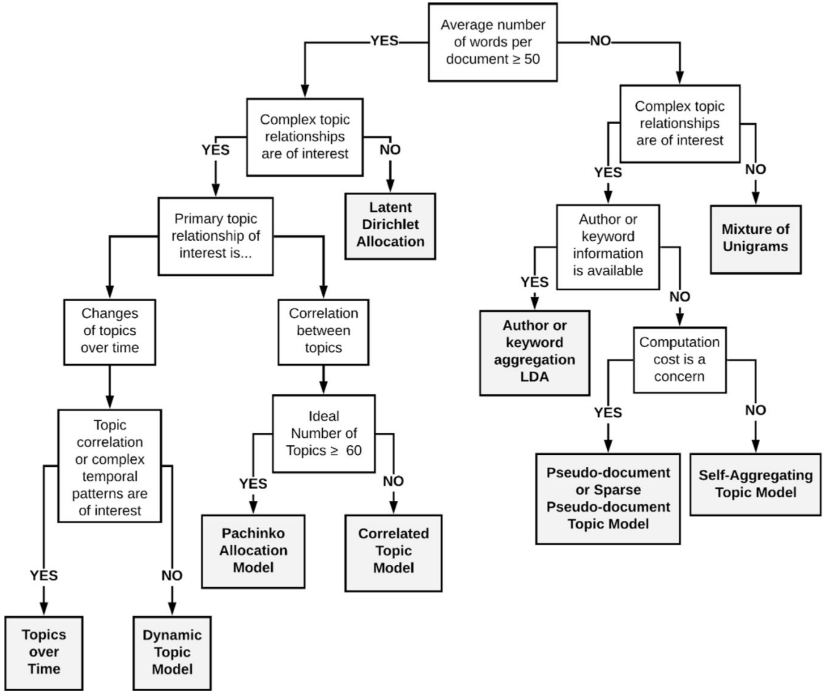
Autoencoding variational inference for topic models

A Srivastava, C Sutton - arXiv preprint arXiv:1703.01488, 2017 - arxiv.org

...

The baseline

- Latent Dirichlet Allocation (LDA)
- Correlated Topic Model (CTM)
- Pachinko Allocation Model


Average number of words per document ≥ 50
YES
NO
Complex topic relationships are of interest
YES
NO
Primary topic relationship of interest is...
Latent Dirichlet Allocation
Changes of topics over time
Correlation between topics
Topic correlation or complex temporal patterns are of interest
YES
NO
Topics over Time
Dynamic Topic Model
Ideal Number of Topics ≥ 60
YES
NO
Pachinko Allocation Model
Correlated Topic Model
Complex topic relationships are of interest
YES
NO
Author or keyword information is available
Mixture of Unigrams
YES
NO
Author or keyword aggregation LDA
Computation cost is a concern
YES
NO
Pseudo-document or Sparse Pseudo-document Topic Model
Self-Aggregating Topic Model

Adversarial-neural Topic Model (ATM)

- 🟢 Outperforms common topic modelling approaches
- 🔴 No hierarchy/lifelong learning/metadata

Hierarchical LDA (hLDA)

- 🟢 Hierarchical model
- 🟢 Flexible number of topics
- 🟢 Outperforms baseline
- 🔴 Cannot model interdisciplinary texts

Lifelong Neural Topic Modeling (LNTM)

- 🟢 Modular: can be applied to other topic models
- 🟢 Lifelong learning
- 🔴 Memory/computation requirements
- 🔴 Static number of topics

Embedded Topic Model (ETM)

- 🟢 Automatic topic name inference
- 🟢 Modular approach
- 🔴 No hierarchy/lifelong learning/metadata

Questions?