CCS Thesis: Topic Model Space

Alfiuddin R. Hadiat
Rijksuniversiteit Groningen
Computional Cognitive Science MSc
S2863685

Research Papers

Dataset

- arXiv metadata json (Clement et al., 2019)

Good Practice

- No need to remove stop words (schofield et al. 2017)

Technology Used

- Grobid for PDF extraction (*GROBID*, 2008–2021)
- BigARTM for regularized topic modeling implementation (Bulatov et al., 2020)
- Gensim (Rehurek & Sojka, 2011)
- sklearn (Pedregosa et al., 2011)

Topic Modeling

- Original LDA paper (Blei et al., 2003)
- LDA using online variation bayes (M. D. Hoffman et al., 2010; M. Hoffman et al., 2012)
- Oversight in topic model research (Blei, 2012)
- Survey of Topic Modeling Techniques (Sharma et al., 2017)
- Additive regularization topic modelling (i.e. regularized pLSA) (Vorontsov & Potapenko, 2015)

Applied Topic Modeling

- Finding scientific topics (Griffiths & Steyvers, 2004)
- Trend analysis with LDA (Alga et al., 2020)
- Binary classification using topic models (Sarioglu et al., 2012)
- Topic modeling on historical newspapers (Yang et al., 2011)
- Using topic modeling to measure history of ideas (Hall et al., 2008)

Topic Modeling in Recommender Systems

- LDA for tag recommendation (Krestel et al., 2009)
- Recommendation using cosine similarity of topic distribution (Chang et al., 2017)
- Replacing item latent vector with topic distributions in user-item recommender (C. Wang & Blei, 2011)
- LDA improves content-based recommendation systems that use Naive-Bayes, K-Nearest Neighbors, and Regression (Luostarinen & Kohonen, 2013)

Topic Modeling with Word Embeddings

- Use word embeddings for term-space modelling (Sahlgren, 2020)
- LDA using Gaussian mixtures over word embedding spaces (Das et al., 2015)
- LDA using von Mises-Fisher distribution modeling over word embedding spaces (Batmanghelich et al., 2016)
- LDA, but assign words to topics using embedding agreements and variational inference (Dieng et al., 2020)
- Topic modeling by K-Means clustering on word embeddings (Sia et al., 2020)

Neural Topic Modeling

- BERTopic paper – used in experiment (Grootendorst, 2022)
- Neural networks topics are good, but probabilistic topic modeling still better for document representation (Doan & Hoang, 2021)
- Neural variational document model (Miao et al., 2016, 2017) using variational auto-encoders (Kingma & Welling, 2014) of Gaussian softmax distributions
- ProdLDA, a neural network based topic modeling technique using a variational autoencoder and product of experts (Srivastava & Charles, 2017; Hinton, 2017)
- Combine ProdLDA with SBERT word embeddings to create CombinedTM, which achieves solid coherence (Bianchi et al., 2020)
- Neural topic modeling using bidirectional adverserial training of word embeddings (R. Wang et al., 2020)

Combined Topic Modelling

- Use topic modeling (LDA or GDSMM) as layers in BERT layers for semantic similarity (Peinelt et al., 2020)
- Also topic modeling with BERT layering, but for abusive speech (Bose et al., 2021)

Topic Evaluations

- Comprehensive examination of different coherency measures; used by Gensim coherence model (Röder et al., 2015)
- Pointwise mutual information for coherence (Newman et al., 2010)
- log-conditional probability of document-word frequency as coherence (Mimno et al., 2011)
- Coherence as topic evaluation correlates with human judgment; also, normalized PMI (Lau et al., 2014)
- Inclusion of overlap, coverage, uniqueness, and separation as topic evaluators (Sahlgren, 2020)

Metaresearch

- P-value use in biomedical research (Chavalarias et al., 2016)
- Science mapping on biases in biomedical research (Chavalarias & Ioannidis, 2010)
- Using a digital archive as a historical archive for psychology (Burman, 2018a)
- Network analysis of citations for historical analysis (Burman, 2018b)

References

Alga, A., Eriksson, O., & Nordberg, M. (2020). Analysis of scientific publications during the early phase of the covid-19 pandemic: Topic modeling study. *Medical Internet Research*, *22*, 1-11.

Batmanghelich, K., Saeedi, A., Narasimhan, K., & Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the conferenece. association for computational linguistics. meeting.* (Vol. 2016, pp. 530–539).

Bianchi, F., Terragni, S., & Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *ArXiv preprint arXiv:2004.03974*.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Machine Learning Research*, *3*, 993-1022.

Bose, T., Illina, I., & Forh, D. (2021). Generalisability of topic models in cross-corpora abusive language. In *2021 workshop on nlp4if: Censorship, disinformation, and propaganda, naacl.*

Bulatov, V., Alekseev, V., Vorontso, K., Polyudova, D., Veselova, E., Goncharov, A., & Egorov, E. (2020). Topicnet: Making additive regularization for topic modelling accessible. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6745–6752).

Burman, J. T. (2018a). Through the looking-glass: Psychinfo as an historical archive of trends in psychology. *History of Psychology*, *21(4)*, 302-333.

Burman, J. T. (2018b). What is history of psychology? network analysis of journal citation reports, 2009-2015. *SAGE Open*.

Chang, T.-M., Hsiao, W.-F., & Hsu, M.-F. (2017). Hidden topic analysis for personalized document recommendation.

Chavalarias, D., & Ioannidis, J. P. A. (2010). Science mapping analysis characterizes 235 biases in biomedical research. *Clinical Epidemiology*, *63*, 1205-1215.

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting p values in the biomedical literature, 1990-2015. *American Medical Association*, *315(11)*, 1141-1148.

Clement, C. B., Bierbaum, M., O'Keeffe, K. P., & Alemi, A. A. (2019). *On the use of arxiv as a dataset.*

Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian lda for topics models with word embeddings. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Vol. 1: Long Papers, pp. 795–804).

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453.

Doan, T. N., & Hoang, T. A. (2021). Benchmarking neural topic models: An empirical study. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 4563–4368).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *PNAS*, *101*, 5228-5235.

Grobid. (2008–2021). https://github.com/kermitt2/grobid. GitHub.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hall, D., Jurafsky, D., & Manning, C. D. (2008, October). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 363–371). Honolulu, Hawaii: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D08-1038`

Hinton, G. E. (2017). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*(8).

Hoffman, M., Blei, D. M., Wang, C., & Paisley, J. (2012). *Stochastic variational inference.* arXiv. Retrieved from `https://arxiv.org/abs/1206.7051` doi: 10.48550/ARXIV.1206.7051

Hoffman, M. D., Blei, D. M., & Bach, F. (2010). Online learning for latent dirichlet allocation. In *Proceedings of the 23rd international conference on neural information processing systems - volume 1* (p. 856–864). Red Hook, NY, USA: Curran Associates Inc.

Kingma, D., & Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of iclr.*

Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. *RecSys'09 - Proceedings of the 3rd ACM Conference on Recommender Systems*, 61-68.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics* (pp. 530–539).

Luostarinen, T., & Kohonen, O. (2013). Using topic models in content-based news recommender systems. In *Proceedings of the 19th nordic conference of computational linguistics (nodalida 2013)* (pp. 239–251).

Miao, Y., Grefenstette, E., & Blunsom, P. (2017). Discovering discrete latent topics with neural variational inference. In *34th international conference on machine learning.*

Miao, Y., Yu, L., & Blunsom, P. (2016). Neural variational inference and learning in belief. In *Proceedings of icml.*

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).

Newman, D., La, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 100–108).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peinelt, N., Nguyen, D., & Liakata, M. (2020). tbert: Topic models and bert joining forces for semantic similarity detecting. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7047–7055).

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (p. 399–408). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2684822.2685324` doi: 10.1145/2684822.2685324

Sahlgren, M. (2020). Rethinking topic modeling: From document-space to topic-space. In *Proceedings of the 2020 conference on empirical methods in natural language processing: Findings* (pp. 2250–2259).

Sarioglu, E., Choi, H. A., & Yadav, K. (2012). Clinical report classification using natural language processing and topic modeling. In *2012 11th international conference on machine learning and applications* (Vol. 2, pp. 204–209).

Sharma, D., Kumar, B., & Chand, S. (2017). A survey on journey of topic modeling techniques from svd to deep learning. *I. J. Modern Education and Computer Science*, *7*, 50-62.

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topics models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*.

Srivastava, A., & Charles, S. (2017). Autoencoding variational inference for topic models. *ArXiv preprint arXiv:1703.01488*.

Vorontsov, K., & Potapenko, A. (2015). Additive regularization of topic models. *Machine Learning*, *101*(1), 303–323.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 448–456).

Wang, R., Hu, X., Zhou, D., He, Y., Xiong, Y., Ye, C., & Xu, H. (2020, July). Neural topic modeling with bidirectional adversarial training. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 340–350). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.acl-main.32` doi: 10.18653/v1/2020.acl-main.32

Yang, T.-I., Torget, A., & Mihalcea, R. (2011, June). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 96–104). Portland, OR, USA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W11-1513`