Syllabus: CSI 796 Probabilistic Topic Models

Thomas W. Jones

5/23/2018

Schedule

Week	Date	Topic	Assignment
1	05-24-2018	RcppParallel	Read: RcppParallel
2	05-31-2018	RcppParallel	Read: RcppParallel
3	06-07-2018	RcppParallel	Do: Program DTM, TCM
4	06-14-2018	LDA	Read: Latent Dirichlet Allocation
5	06-21-2018	LDA	Read: Consistency in Latent Allocation Models
6	06-28-2018	LDA	Read: Finding Scientific Topics, Fast Collapsed Gibbs Sampling
7	07-05-2018	LDA	Read: Distributed Inference for Latent Dirichlet Allocation
8	07-12-2018	LDA	Read: PLDA, PLDA+
9	07-19-2018	LDA	Do: Plan LDA program, Program LDA
10	07-26-2018	LDA	Do: Program LDA
11	08-02-2018	pLSA	Read: Probabilistic Latent Semantic Analysis
12	08-09-2018	pLSA	Do: Program PLSA

Objectives

Broadly, I see several learning objectives, described below. I also list deliverables which should demonstrate the understandings gained.

1. Learn the foundational computational tools. Deliverable: functions to create core data structures for LDA and pLSA accessible through R.
2. Learn the mathematical foundations of LDA. Deliverable: Summaries of several core papers deriving the LDA model.
3. Learn the computational foundations of LDA. Deliverable: A parallel Gibbs sampler for training LDA models from R.
4. Learn the mathematical foundations of pLSA. Deliverable: A summary of the core paper deriving the pLSA model.
5. Learn the computational foundations of pLSA. Deliverable: A function for training a pLSA model using the EM algorithm from R.

Tentative reading list

RcppParallel

1. RcppParallel's web book, including examples

Latent Dirichlet Allocation

1. Latent Dirichlet Allocation, Blei et al., JMLR (3), 2003.

2. Womack, A., Moreno, E., & Casella, G. (2013). Consistency in Latent Allocation Models. Submitted to Annals of Statistics.
3. Finding scientific topics, Griffiths and Steyvers, PNAS (101), 2004.
4. Fast collapsed gibbs sampling for latent dirichlet allocation, Porteous et al., KDD 2008.
5. Distributed Inference for Latent Dirichlet Allocation, Newman et al., NIPS 2007.
6. PLDA: Parallel Latent Dirichlet Allocation for Large-scale Applications. Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang. AAIM 2009.
7. PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing. Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, Maosong Sun. ACM Transactions on Intelligent Systems and Technology, special issue on Large Scale Machine Learning. 2011.

Probabilistic Latent Semantic Analysis

1. Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc.

Additional Topic models

1. Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In Advances in neural information processing systems (pp. 121-128).
2. Blei, D. M., & Lafferty, J. D. (2005, December). Correlated topic models. In Proceedings of the 18th International Conference on Neural Information Processing Systems (pp. 147-154). MIT Press.
3. Wallach, H. M. (2008). Structured topic models for language (Doctoral dissertation, University of Cambridge).
4. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
5. Goldberg, Y., & Levy, O. (2014). word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.