Topic modeling (LSA with SVD approach)

Ching-wen Wang (cwang553@wisc.edu) 9077508928
Mo Xiao (mxiao28@wisc.edu) 9077625565

As technology advances, tons of information are generated everyday. It would cost a lot of time and money to read through every single document to find the information we needed. Therefore, if we can efficiently associate each document with relevant topics it contains, we can improve the process of information retrieval. This goal can be achieved through topic modeling, which discovers abstract topics embedded in a collection of documents. In this project, we will explore a renowned method of topic modeling, latent semantic analysis (LSA), that uses a singular value decomposition (SVD) approach (Gong & Liu, 2001). Although LSA can enhance information retrieval, there are also limitations such as data sparsity or large storage space problems. Therefore, we are also keen to find out how LSA can be improved by incorporating syntactic information (Kanejiya, Kumar & Prasad, 2003), part of speech (Kakkonen, Myller & Sutinen, 2006), or implementing regularization methods of lasso and ridge regression (Wang, Q., Xu, J., Li, H., & Craswell, N., 2013).

Sources:

Gong, Y., & Liu, X. (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *SIGIR'01.* https://www.cs.bham.ac.uk/~pxt/IDA/text_summary.pdf

Kakkonen, T., Myller N., & Sutinen E. (2006). Applying Part-of-Speech Enhanced LSA to Automatic Essay Grading. *Proceedings of the 4th IEEE International Conference on Information Technology: Research and Education.* https://arxiv.org/abs/cs/0610118

Kanejiya, D., Kumar, A., & Prasad, S. (2003). Automatic evaluation of students' answers using syntactically enhanced LSA. *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, *2*, 53–60. https://doi.org/10.3115/1118894.1118902

Wang, Q., Xu, J., Li, H., & Craswell, N. (2013). Regularized latent semantic indexing: A new approach to large-scale topic modeling. ACM Transactions on Information Systems (TOIS), 31(1), 1-44. http://www.hangli-hl.com/uploads/3/1/6/8/3168008/rlsi-tois-revision.pdf