Author: guomianzhuang

Title: StandardLDA and SparseLDA

StandardGibbs Sampling Complexity is $O(MN_mK)$(N_m is the average document length). The standard implementation is too slow to apply to the project. So we need a more faster implementation. Yao presents an algorithm and data structure for evaluating Gibbs sampling distribution, SparseLDA. This method can reduce the time complexity to O(MNm|NoneZero(Nkt)|), if the topic num is large, the Nkt will be sparse, this method can improve the speed of model inference.

We have implemented this two LDA model with Gibbs Sampling. (See LdaModel.py and LdaModel_SparseLDA.py)

Exprimental environment:

System: Mac OS X

CPU: 2.7GHz intel Core i5

Memory: 8GB

Result:

I set the alpha=50.0/K, beta = 0.01, iteration=1000 for all cases and use a corpus including 1000 articles.

The result show in Figure 1, SparseLDA has a lower overall time than

StandardLDA, and the time increases slowly as we increase the topics.

As shown in Figure 2, the perplexity of two algorithm is almost the same.

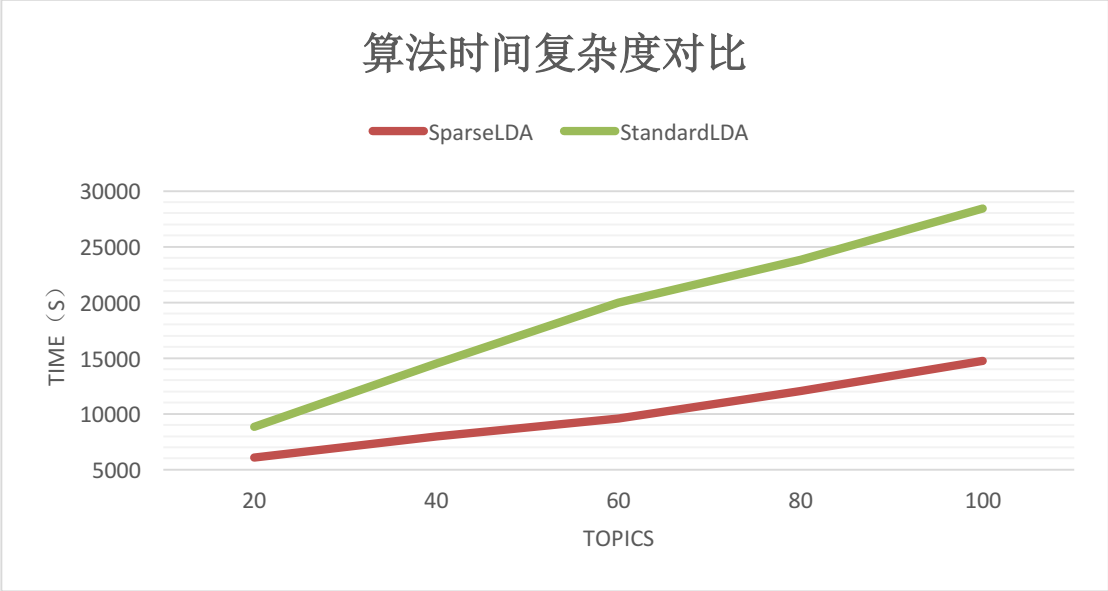

Figure 1 Comparison of time complexity

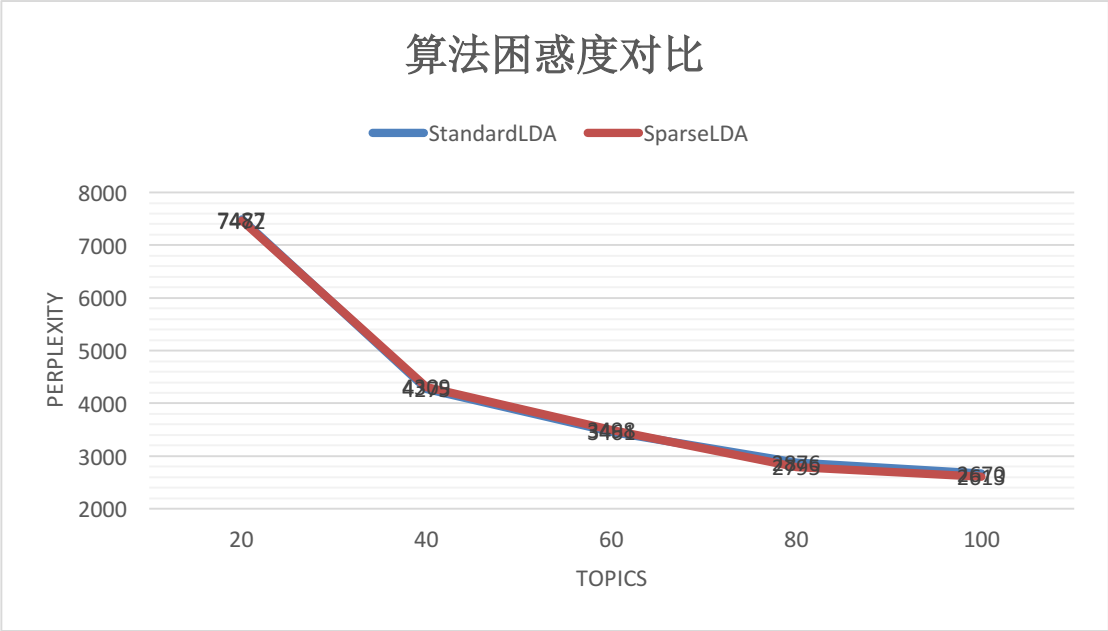

Figure 2 Comparison of Perplexity

Discussion:

Though the SparseLDA reduces the time cost, but it still is

time-consuming. Next, I will try to parallelize the algorithm.

The implementation is just for study, we should use more mature tools, such as Gensim.

参考文献

[1].[LDA 工程实践之算法篇-2] SparseLDA 算法. http://www.flickering.cn/nlp/2014/10/lda%E5%B7%A5%E7%A8%8B%E5%AE%9E%E8%B7%B5%E4%B9%8B%E7%AE%97%E6%B3%95%E7%AF%87-2-sparselda%E7%AE%97%E6%B3%95/

[2].Griffiths T L, Steyvers M. Finding scientific topics.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 Suppl 1(1):5228.

[3].Yao L, Mimno D, Mccallum A. Efficient methods for topic model inference on streaming document collections[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July. DBLP, 2009:937-946.

[4].Li A Q, Ahmed A, Ravi S, et al. Reducing the sampling complexity of topic models[J]. 2014:891-900.