# 【关于 主题模型】那些你不知道的事



## 一、 trick

1. 利用优质“少量”数据学习模型，缓解单机速度和内存问题，然后对剩余/新文档做推导（可数据并行）。比如用微博训练LDA时，先把长度短的微博过滤掉（有工作得出长度为7的短文本适合LDA进行学习），过滤相似微博（转发会造成很多近乎相同的微博）。当训练数据量大并且单机环境中可试一下GraphLab Create，该工具还支持采样比较快的alias LDA。如果不仅是为了学习当前语料中的主题分布，并且也用于预测新数据，则数据量越大越好。
2. 去除一些TF/DF较低/高的词，较低的词在拟合的过程中会被平滑掉，较高的词没有区分力，标点，助词，语气词也可以去掉（中文常用词60万左右）。在中文中应考虑全角变半角，去乱码，繁转简，英文中考虑大小写转换。实际处理数据时会发现分词后不同词个数很容易达到百万级别，这里很多词是没有意义的，数字词，长度过长的词，乱码词。此外，分词过程中如果两个词在一起的频率比较高，那么分词结果会把两个词合并，那么合并与否对LDA的训练是否有影响呢？有的词应该合并，比如”北京 大学“，也有的词分开会好一些，比如”阶级 斗争“。
3. 根据上下文合并短文本，比如合并用户所有的微博作为一个文档，合并相似的微博作为一个文档，把微博当做一个查询，利用伪反馈来补充微博内容（中文微博比twitter字数更多一些，长微博不用扩展已经可以正确分类，短微博本身可能就是歧义的，扩展效果也不一定好），把微博及其评论作为一个文档。在一定程度上可缓解短文本问题。
4. Topic Model的训练是一个数据拟合过程，找出latent topic最大训练语料库的似然概率，当不同类的数据不平衡时，数量量少的主题可能会被数据量多的主题主导。LDA本来就倾向于拟合高频的topic。LDA很多奇怪的结果大多都是因为词的共现导致的。
5. 训练过程中，迭代次数一般可设为1000 – 2000次，可根据时间要求，机器配置选择。迭代次数达到一定值后，会在最小值处来回跳转。LDA的运行时间和文档数，不同词个数，文档长度，topic个数有关。
6. K的选择，对每个K跑一个LDA，肉眼观察每个topic的情况最靠谱。当训练数据量大时不可行。此时可以根据不同的topic的相似度来调整K。假设不同topic之间的相似性小为佳（Perplexity，GraphLab Create直接输出这个结果）。一个经验设置是K × 词典的大小 约等于 语料库中词的总数。
7. 挖掘优质的词典很重要，一方面有助于分词，也有助于明确潜在的主题。
8. 数据量大后，LDA和PLSA的效果差不多，但是PLSA更容易并行化。LDA和PLSA的最大区别在于LDA对于Doc的Topic分布加上了一层先验，Doc-topic分布是当作模型变量，而LDA则只有一个超参数，Doc-Topic分布则是隐藏变量。在预测的时候，plsa是求一个似然概率，lda则是有两项，先验乘以似然。
9. LDA在文本领域中，把word抽象成topic。类似，LDA也可以用在其它任务中，我们在信用评估中，直接把每个用户当成一个文档，文档中的词是每个关注的人，得到的topic相当于是一个用户group，相当于对用户进行聚类。还有，把微博中的@/rt的人当作word。http://www.machinedlearnings.com/2011/03/lda-on-social-graph.html
10. 超参数\alpha \beta对训练的影响？\alpha越大，先验起的作用就越大，推导的topic分布就越倾向于在每个topic上的概率都差不多。\alpha的经验选择为50/k, 其中k是topic数目，beta一般为0.01
11. the color of a word tend to be similar to other words in the same document.
12. the color of a word tend to be similar to its major color in the whole corpus.
13. 用大的数据集训练一个general的model，还是根据垂直领域训练一个specific的model呢？应该看是想得到一些小众的topic，还是比较热门的topic。

14. 为什么LDA的最大似然难求？含有两个连续的隐藏变量，需要积分掉，对于一个word，需要考虑每个topic生成这个word的概率，因此也有个求和项。因为这个条件分布很难求，导致求解带隐变量优化问题的EM算法也不行，因此EM算法往往都是用一个近似分布来代替。Gibbs Sampling则是生成p(z|...)的几个样本来近似这个条件分布。经过多次迭代（一次迭代对于一篇文章中的一个词只采样一次），一开始随机产生的 topic-word 矩阵 和 doc-topic 会处于稳定，真实的分布。对于一个Doc，根据词之间的可交换性，取不同词对应的topic的过程也是独立的。
15. 短文本可以尝试TwitterLDA（假设一个短文本只关于一个话题），https://github.com/smutahoang/ttm