# 主题模型学习

从LSA到pLSA，再到LDA。都好好弄弄吧。

## LSA学习感悟

本来计划着是学习好主题模型的，从PLSA到LDA，但是发现LSA是PLSA的前辈，所以好好的学习了一下。发现LSA其实就是将word-by\_doc这个矩阵进行了一下svd分解，然后将分解后的U，V矩阵进行特征查询，保留了前面一些特征值较高的列，来发现数据内部的特征。将LSA应用到文本检索里面就是常说的LSI，所以LSA和LSI其实是一个东东。如果想要对LSA有个大致的了解，可以参考下[这里](http://www.puffinwarellc.com/index.php/news-and-articles/articles/33-latent-semantic-analysis-tutorial.html?showall=1)。

目录下的 lsa.txt 是word-by-doc矩阵，lsa\_tutorial.m 和 lsa\_tutorial.py 分别是对应的matlab和python代码，可以运行了解下效果。

LSA的有点是可以解决同义词检索的问题，但是对于多义词，效果并不好，而且svd分解的计算时间也是非常大的。