第十四周读书报告 5.25-5.31

61518426 周之遥

·读书进度

学习统计学习方法第20章：潜在狄利克雷分布。

隐含狄利克雷分布在文本挖掘领域包括文本主题识别、文本分类以及文本相似度计算方面都有应用。

LDA是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

它是一种主题模型，它可以将文档集中每篇文档的主题按照概率分布的形式给出。

同时是一种无监督学习算法，在训练时不需要手工标注的训练集，需要的仅仅是文档集以及指定主题的数量k即可；

·下周计划

完成21章PageRank算法。