需要学习或注意的内容有以下：

Python；

PathFinder算法；

Spark；

Word 分词的安装与使用；

χ 2 统计的实现；

基于χ 2 统计特征选择；

基于贝叶斯(Bayes)方法的文本分类；

VSM描述文本；

知识抽取（这是很大的一部分，项目的重中之重！！！）；

知识图谱的RDF 存储方式；

可视化映射怎么实现；

D3.js；

UI设计。