作业 1

1、 题目：索引词的选择

2、 目的：建立索引是信息检索系统中不可缺少的组成部分，它的质量直接影响到检索的效果。这个作业的目的是使大家充分理解索引词的选择过程及其重要性，进一步发现这一过程中所遇到的重点、难点问题及当前的解决方法。

3、 资源

分词软件、文档集、停用词表（如需要，自己创建）

4、 任务：

1、 切词及词频统计：利用已提供的海量智能分词软件对文档进行切词处理，并进行词频统计，形成 DocIndex 文件，结构为：文档号、频率、词，具体例子参见 reference/docIndex。注意保留中间结果，建立合理的数据结构来存储。

2、 分配词权重： 采用词频标准化（$tf_i = tf_i/Max(tf)$）和 tf*idf 两种方式分配词的权重。由 DocIndex 文件生成 DocIndex(tf) 和 DocIndex(tf*idf)文件。注意阈值的确定，词的取舍。

3、 形成倒置文档：将 DocIndex(tf) 和 DocIndex(tf*idf)文件转换为 DocInvert(tf) 和 DocInvert (tf*idf)文件。

5、 要求：

程序：结构清楚，代码优化，注释清晰

报告：见项目报告格式及要求

编程环境：任选

6、 期限、提交方式：将于 1 月 5 日以答辩的形式检查。提交源程序、索引结果及报告。

7、 参考资料：

英文文档的处理结果：Example-DocIndex，Example-DocIndex(tf)，Example-DocIndex(tfidf) Example-DocInvert(tf)，Example-DocInvert(tf*idf),

8、 文件格式：

Files	Formats		
DocIndex	DocID	Frequency Number	Term/Word
DocIndex(tf)	DocID	Weight(tf)	Term/Word
DocIndex(tfidf)	DocID	Weight(tfidf)	Term/Word
DocInvert(tf)	Term/Word	Weight(tf)	DocID
DocInvert(tfidf)	Term/Word	Weight(tfidf)	DocID