山东大学 计算机科学与技术 学院

信息检索与数据挖掘 课程实验报告

学号：201600301148	姓名： 周阳	班级： 16 人工智能
实验题目：VectorSpaceModel		
实验内容： 构建向量空间（VectorSpace），使用特征向量表示文档（Document）。		

实验过程中遇到和解决的问题：

（记录实验过程中遇到的问题，以及解决过程和实验结果。可以适当配以关键代码辅助说明，但不要大段贴代码。）

数据集：20 Newsgroups dataset

过程简述：

• 数据预处理

Stemming（提取词干），使用 nltk 库 SnowballStemmer 进行词干提取

去大小写（部分大小写）

去特殊符号与数字

去停用词（stopwords）

• Tokenization

构建词典，对处理结果 split 后统计词频，得到词：词频的 dict 对应，最终词典大小：133015 词

• 构建 tf-idf 向量代表文档特征

$$Tf\ (w,\ d) = count(w,d)/size(d)$$

Tf 代表文档词频，代表词 w 在文档 d 中出现的次数

$$Idf\ (w,\ D)\ = log\ (n/docs(w,D))$$

IDF 代表逆文档词频，主要代表词 w 出现在的文档数的倒数的对数值

$$Tf\text{-}idf = Tf \times idf$$

TF-IDF 的值表示词对于某个文档的重要程度（key word）：TF-IDF 值高的词根据上述公式具有特点：1，在该文档中出现次数大。2，在全体文档中出现的并不频繁。

从而具有这样特征的词，可以更好地表示文档，这一样的词，可以理解为这一文档的 keyword。

计算向量使用 numpy array 进行快速运算。

• 存储向量

存储向量和 list 对象使用 numpy.save () 进行序列化操作，最终储存在.bin 二进制文件中。并且使用一个文本文件记录文件路径与表示对应的向量

结论分析与体会：

本次实验中，主要任务在对英文数据的预处理，和构建 vocab 的工作上，让我对自然语言处理，语言部分的信息处理，信息挖掘有了新的认识。

并且在最后部分的向量存储，向量计算上，由于每个向量的维度过大，存储的时需要一个一个向量的存储进文本文件，最终的文件大小超过 20GB，也让我对大数据有了进一步的认识。