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：201600130053 | 姓名：王斌 | 班级： 16人工智能 |
| 实验题目：预处理文本数据集，并且得到每个文本的VSM表示。 | | |
| 实验内容：  #Homework 1: VSM  #预处理文本数据集，并且得到每个文本的VSM表示。  The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.  #20news-18828.tar.gz（http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz）?- 20 Newsgroups; duplicates removed, only "From" and "Subject" headers (18828 documents) | | |
| 实验环境：  Spyder+python3.6  Win10 | | |
| 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。）  **100 5 200 10**    **300 20 500 30** | | |
| 结论分析与体会：  向量空间模型基本实现，并且能在模型基础上进行一些基本的应用测试，但由于对于文档的处理完全都是有编写的函数完成，入对单词的单复数、形式等未能处理好，还有对于过于低频和高频的用处不大的单词未处理，都有待优化。 | | |