山东大学 计算机科学与技术 学院

信息检索 课程实验报告

|  |  |  |
| --- | --- | --- |
| 学号：201600150109 | 姓名：沈棋韬 | 班级： 16人工智能班 |
| 实验题目：: VSM | | |
| 实验内容：  预处理文本数据集，并且得到每个文本的VSM表示 | | |
| 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。）  遍历读取文件问题，通过创建一个保留每个文件路径的列表完成。  文件格式不能解码问题，通过errors=’ignore’避免。  文件中去除标点，使用replace完成。  文件分句，使用nltk.sent\_tokenize()完成。  文件分词，使用nltk.word\_wokenize()完成。  文件停用词，使用from nltk.corpus import stopwords导入停用词表。 | | |
| 结论分析与体会：  在预处理时要尽可能完善，否则词袋中会出现很多没有意义的词，如邮箱地址。 | | |