**文本内容计算课程实验报告**

实验名称：文本表示实验

姓名：覃婧丽

学号：201911690341

院系：计算机学院

专业：数字媒体技术

二〇二二年四月

|  |
| --- |
| 1．实验目的 |
| ·深入理解文本表示各类方法的原理，并通过代码实现，提升实践能力。  ·通过实验比较不同文本表示方法的差异。在具体文本计算任务中观察结果差异。  ·通过撰写实验报告提升研究思维能力和总结归纳能力。 |
| 2．实验内容 |
| 比较三种文本表示方法  1）one-hot，可引入TF-IDF作为特征；  2）训练CBOW或Skip-gram的静态词向量；  3）预训练语言模型的词向量  •实验描述：  在Yelp数据集上进行实验，得到了词向量后，将一个句子中的词向量进行平均得到句向量，用句向量作为特征进行逻辑回归分类Yelp数据集，比较不同方法的准确率。准确率越高说明该文本表示方法更好。 |
| 3．实验步骤 |
| 获取三种文本表示方法的词向量  • One-hot词向量  1.构建one-hot词表    2.构建one-hot词向量和句向量    结果：    3.构建训练集和测试集    4.利用one-hot的句向量进行逻辑回归    5.测试结果  • 用yelp数据集训练word2ve词向量   1. 定义模型      1. 初始化参数      1. 建立损失函数      1. 存储词向量      1. 定义数据      1. 进行训练   • 开源的预训练词向量 （例如Glove预训练词向量：https://nlp.stanford.edu/projects/glove/ ）   1. 读入词向量      1. 词向量转化为句向量      1. 用句向量进行逻辑回归      1. 测试结果 |

|  |
| --- |
| 4．实验结果与分析 |
| One-hot词向量训练结果  获得更改词频后的两个结果      Skip-gram的静态词向量训练结果  词向量：      预训练语言模型的词向量训练结果    由于储存问题，one-hot训练没有办法进行40000训练集，但是在10000的次数下，其准确率已达到0.93  可以推测准确率排序为：one-hot训练>Skip-gram训练>预训练词向量训练 |
| 5．进一步工作说明 |
| 可以让三个词向量进行同频训练后多次比较，获取准确率平均数再进行三个词向量准确率的比较，这样结果会更准确，更具科学性 |