**武汉大学国家网络安全学院**

**实验报告**

**课程名称 自然语言处理**

**专业年级 2019级**

**姓 名 庞远心**

**学 号 2019302050244**

**协 作 者 无**

**实验学期 2021-2022　 学年 第一 学期**

**课堂时数 课外时数**

**填写时间 2021 年 10 月 13 日**

|  |
| --- |
| **实验介绍** |
| **【实验名称】：自然语言处理第一次实验** |
| **【实验目的】：**   1. **掌握jieba分词工具的基本使用** 2. **了解分词与词向量的训练。分词就是将句子、段落、文章这种长文本，分解为以字词为单位的数据结构，方便后续的处理分析工作。词向量，顾名思义就是⽤来表⽰词的向量，也可被认为是词的特征向量或表征。把词映射为实数域向量的技术也叫词嵌入。我们通常在训练语言模型的同时得到词向量。** 3. **了解Word2Vec词向量的特性。词向量本质是将一些低维、离散、不带任何意义的序号映射成带有特定任务性质的高维特征。为了研究这些特性，可以显式地通过余弦距离来比较两个词的相似性，也可以通过词间的类比关系来探查词向量的影藏特性，另外还可以通过PCA降维的方式将词向量可视化。** 4. **学习使用python**   **【实验环境】：**   1. **Pycharm** 2. **Python 3** 3. **Windows 10** 4. **genism** 5. **sklearn** 6. **matplotlib** 7. **numpy** 8. **seaborn** |
| **实验内容** |
| **【实验方案设计】：**  1.使用结巴分词，对于基本语句进行分词    2. 使用gensim中的Word2Vec模型训练词向量  model = Word2Vec(common\_texts, size=100, window=5, min\_count=1, workers=4)  其中：  min\_count指定了需要训练词语的最小出现次数，默认为5；size指定了训练时词向量维度，默认为100；worker指定了完成训练过程的线程数，默认为1不使用多线程。  3. 使用训练好的词向量对指定的词（2个例子）进行相关性比较  model.similarity('中国','毛泽东')；    4. 使用训练好的词向量选出与指定词（2个例子）最相似的10个词：model.wv.most\_similar(positive=['邓小平'], topn=10)；      5. 使用训练好的词向量选出与指定词类比最相似的5个词（2个例子）    6.对词语进行降维并输出：    **【实验结果分析】：**  **因为是复现老师的实验步骤，所以结果也没什么差别** |
| **【实验总结】：**  **1. 本次实验掌握了基本的句子转化为向量的模型训练方法，熟悉了python的使用技巧。** |
| **评语及评分（指导教师）** |
| **【评语】：**      **评分：**  **日期：** |

**附件2：**

**实验报告说明**

**1．实验名称：**要用最简练的语言反映实验的内容。

**2．实验目的**：目的要明确，要抓住重点。

**3．实验环境**：实验用的软硬件环境（配置）。

**4．实验方案设计（思路、**步骤和方法等**）**：这是实验报告极其重要的内容。包括概要设计、详细设计和核心算法说明及分析，系统开发工具等。应同时提交程序或设计电子版。

对于**设计型和综合型实验**，在上述内容基础上还应该画出流程图、设计思路和设计方法，再配以相应的文字说明。

对于**创新型实验**，还应注明其创新点、特色。

**5．实验结果分析：**即根据实验过程中所见到的现象和测得的数据，进行对比分析并做出结论（可以将部分测试结果进行截屏）。

**6．实验总结：**对本次实验的心得体会，所遇到的问题及解决方法，其他思考和建议。

**7．评语及评分：**指导教师依据学生的实际报告内容，用简练语言给出本次实验报告的评价和价值。