NLP 作业 3：文本词向量

顾韬
1312855584@qq.com

Abstract

利用给定语料库（金庸小说集），利用神经语言模型 Word2Vec 等模型来训练词向量，通过对词向量的聚类或者其他方法来验证词向量的有效性。

Methodology

Word2Vec 模型

Word2Vec 是一种用于生成词向量的深度学习模型，旨在将词语表示为连续向量空间中的点，从而捕捉词语之间的语义关系。它是由 Google 的 Tomas Mikolov 等人在 2013 年提出的，并已成为自然语言处理中最为流行的词向量表示方法之一。

Word2Vec 模型的核心思想是通过预测周围词语来学习每个词语的分布式表示。在训练过程中，Word2Vec 模型通过输入的文本序列来学习每个词语的向量表示，使得在向量空间中具有相似语义的词语在该空间中彼此相近。

Word2Vec 模型主要有两种架构：Continuous Bag of Words (CBOW) 和 Skip-gram。它们的训练目标是相反的，但在实践中都表现出了很好的效果。CBOW 根据上下文预测当前词语，而 Skip-gram 利用当前词语预测上下文。

Experimental Studies

模型训练

在对文本预处理后，使用使用开源的 Gensim 库提供的接口来训练 Word2vec 模型，具体代码如下：

```
word2vec_model = Word2Vec(data_txt[i], hs=1, min_count=5,
    window=5, vector_size=200, sg=0, epochs=100)
```

对每篇文章分别进行训练。

聚类分析

使用 word2vec_model.wv.similarity 分析不同词的相似度，这里以天龙八部中的主要人物为例，可以看到萧峰，虚竹，段誉三兄弟的相似明显高于其他，而萧远山和萧峰虽然关系为父子，但相似很低，可能的原因是萧远山在第五部之前出场时都没透露名字，第五部出场后不久又被扫地僧度化归隐了，有效的文本频率较低。所以和萧峰的词向量相似度很低。

Similarity between '萧峰' and '乔峰': 0.47484788
Similarity between '萧峰' and '萧远山': 0.087288216
Similarity between '萧峰' and '虚竹': 0.3493003
Similarity between '萧峰' and '段誉': 0.48614055
Similarity between '萧峰' and '慕容复': 0.2439161

Similarity between '萧峰' and '逍遥子': 0.07291481

为了可视化 Word2Vec 模型训练后的词向量，并检查相似词是否聚集在一起，使用降维算法 t-SNE 将高维词向量降到二维空间，然后进行可视化。绘制聚类图时选择前 1000 个词向量，降低运行时间。仍然以天龙八部为例。结果如下：

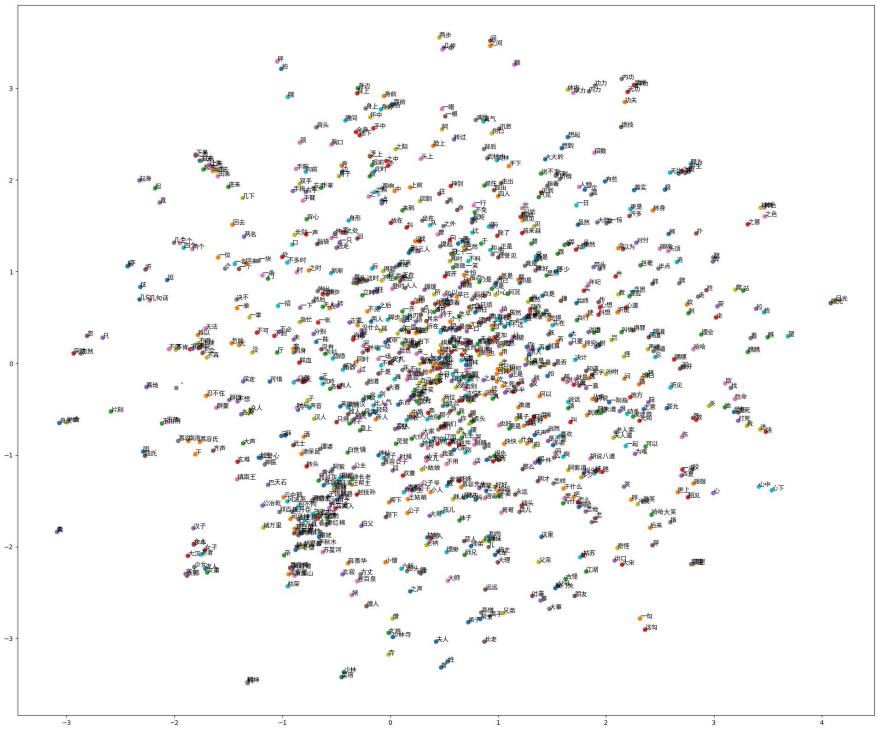

Figure1： 天龙八部词向量聚类图

Figure 2：部分放大

由图所示，主要人物基本聚类在一起，像常见的人物关系的词向量也聚类在一起（但也有部分其他类型词向量参杂）。

其他小说的词向量聚类图在附于作业文件夹下。

Conclusions

在本次实验中，使用 Word2Vec 训练了词向量，并通过计算相似度与绘制聚类图验证了词向量的有效性。