学　　期 2023-2024（2）

北京航空航天大學
BEIHANG UNIVERSITY

NLP 第三次大作业

院（系）名称	自动化科学与电气工程学院
专业名称	控制工程
学生姓名	胡正皓
学　　号	ZY2303205

2024 年 5 月

一、研究背景

利用给定语料库（金庸语小说料如下链接），利用 1～2 种神经语言模型（如：基于 Word2Vec ， LSTM， GloVe 等模型）来训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联、或者其他方法来验证词向量的有效性。

二、验证方式

1、计算词向量之间的语义距离。
2、利用 k-means 聚类，并展示聚类结果。
3、分析不同段落之间词向量的平均语义，判断关联性。
三种验证方式均使用 Word2Vec 对文本进行训练，因此前面的步骤相同：
1、 对文本进行预处理，去除停词等多余符号；
2、 设置参数，用 Word2Vec 对模型进行建模。

三、结果分析

（一）语义距离的计算

		对比	
（郭靖，黄蓉）	0.757	（郭靖，韦小宝）	0.418
（丐帮，少林）	0.630	（丐帮，华山派）	0.007
（反清复明，天地会）	0.683	（反清复明，抗元）	0.105

可以看到，在同一本小说中的词，语义的距离较高，不同小说的词语语义距离较小。验证了模型训练的有效性。

（二）k-means 聚类

利用 k-means 聚类方法对模型进行聚类，聚类结果如图所示：

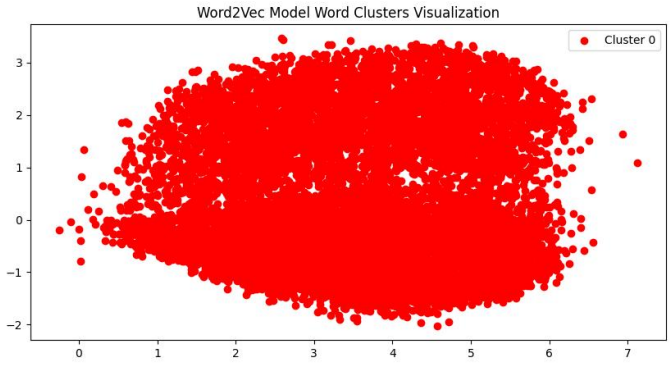

图 1 聚类结果

将所有词语进行聚类分析，共计 10 类聚类。选取其中的一类进行结果展示，可以看到，聚类的效果很好。

（三）不同段落的关联性

选取如下段落：

1、 盈盈见他包裹严密，足见对自己所赠之物极是重视，心下甚喜，道：“你一天要说几句谎话，心里才舒服？”接过琴来，轻轻拨弄，随即奏起那曲《清心普善咒》来，问道：“你都学会了没有？”令狐冲道：“差得远呢。”静听她指下优雅的琴音，甚是愉悦。

2、 任我行冷笑道：“是吗？因此你将我关在西湖湖底，教我不见天日。：东方不败道：”我没杀你，是不是？只须我叫江南四友不送水给你喝，你能捱得十天半月吗？“任我行道：”这样说来，你待我还算不错了？“东方不败道：”正是。我让你在杭州西湖颐养天年。常言道，上有天堂，下有苏杭。西湖风景，那是天下有名的了，孤山梅庄，更是西湖景色绝佳之处。”

通过计算可得，关联度为 0.73

作为对比，选择不同小说的段落来计算关联度：

通过计算，关联度为 0.203.

综上所述，三种方法均有校的验证了模型的有效性。