作业报告三

Weijia Long
wjial@buaa.edu.cn

Abstract

利用给定语料库，利用基于 Word2Vec 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。

Part 1

利用给定语料库，基于 Word2Vec 模型来训练词向量，通过计算词向量之间的语意距离。

Experimental 1

在训练完词向量后，分别计算“刀光，剑影“，”猫，狗“，”牛肉，剑影“这三对词向量之间的语义距离，实验结果如下：

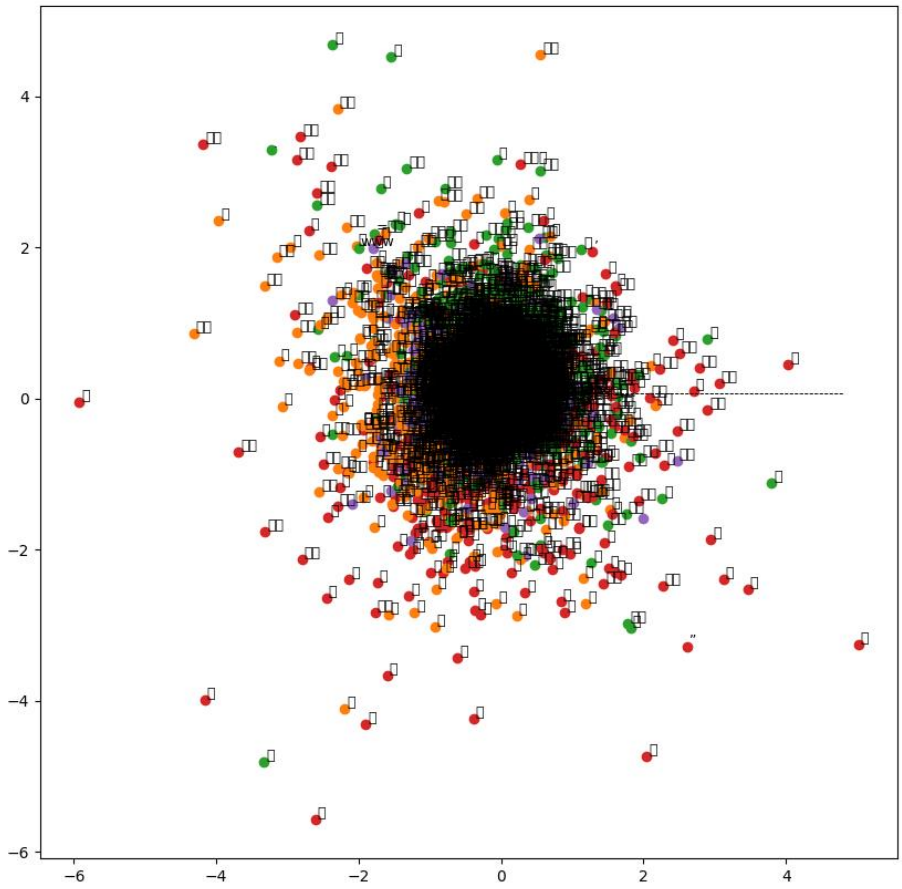
词向量对	语义距离
刀光，剑影	0.7926
猫，狗	0.5065
牛肉，剑影	0.5524

可以看到，在武侠领域的词向量语义距离计算较为准确，但在日常生活语义距离计算不太准确。

Part 2

利用给定语料库，基于 Word2Vec 模型来训练词向量，并展示词语的聚类效果。

Experimental 2



Part3

利用给定语料库，基于 Word2Vec 模型来训练词向量，计算段落的语意关联

Experimental 3

分别利用语料库中段落以及其他语料段落计算语义关联，语料库段落为:

“梅剑四姊妹开动机关 ， 移开大门上的巨岩 ， 放了朱天 、 昊天 、 玄天九部诸女进入大厅。“

“韦小宝想起身上怀有皇帝亲笔御札， 可以调遣文武官员， 说 ： “ 眼下事情紧急 ， 我们少林僧武功虽高，可是寡不敌众， 三十七个和尚 ， 怎敌得过他三千名喇嘛 ？ 我须得立刻下山求救 。”澄通道 ： “ 只怕远水救不着近火 。 ”韦小宝道 ： “ 那么咱们 护送

行痴大师 ， 冲了出去 。”澄通点头道 ：“ 看来只有这个法子 。咱们三十七名少林僧 ，再加上师叔的僮儿 ， 要抵挡三千多名喇嘛 ， 那是万万不能 ， 但要从空隙中冲 ， 却也不是什么难事。”韦小宝道：“就只怕行痴大师和他师父玉林大师不肯，他们说 生死都是一般，逃不逃也没什么分别。”澄通皱眉道：“这就须请师叔劝上一劝。”“

其他语料段落为：

“文本标注子系统的功能目标为提供一个能够完成文本标注与管理的图形化界面平台，以直接提供可供科研使用的有效标注数据或为智能挖掘模型训练提供训练数据“

“标注管理员创建标注项目，填写项目相关信息，加入标注文本，指定标注人员，添加标签后为标注人员分配任务“

计算关联度结果为：

段落对	语义距离
语料库随机抽取	0.9790
其他语料库	0.5872

可以看到，只有对语料库内抽取的段落关联性较强。