学　　期 2021-2022（2）

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理第四次作业

词向量聚类问题

院（系）名称	自动化科学与电气工程学院
专业名称	电子信息
学生姓名	蔡尽云
学号	ZY2103501
指导老师	秦曾昌

2022 年 5 月

1 问题描述

利用给定语料库（或者自选语料库），利用神经语言模型（如：Word2Vec，GloVe 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。

2 实验原理

2.1 词向量

词向量（Word embedding），又叫 Word 嵌入式自然语言处理中的一组语言建模和特征学习技术的统称，其中来自词汇表的单词或短语被映射到实数的向量。 从概念上讲，它涉及从每个单词一维的空间到具有更低维度的连续向量空间的数学嵌入。

如果将 word 看作文本的最小单元，可以将 Word Embedding 理解为一种映射，其过程是：将文本空间中的某个 word，通过一定的方法，映射或者说嵌入（embedding）到另一个数值向量空间（之所以称之为 embedding，是因为这种表示方法往往伴随着一种降维的意思）。生成这种映射的方法包括神经网络，单词共生矩阵的降维，概率模型，可解释的知识库方法，和术语的显式表示单词出现的背景。

现今流行的 Word Embedding 算法携带了语义信息且维度经过压缩便于运算，因此有很多用途，例如：

- 计算相似度。
- 在一组单词中找出与众不同的一个，例如在如下词汇列表中：[dog, cat, chicken, boy]，利用词向量可以识别出 boy 和其他三个词不是一类。
- 直接进行词的运算，例如经典的：woman+king-man =queen。
- 由于携带了语义信息，还可以计算一段文字出现的可能性，也就是说，这段文字是否通顺。

本质上来说，经过 Word Embedding 之后，各个 word 就组合成了一个相对低维空间上的一组向量，这些向量之间的远近关系则由他们之间的语义关系决定。

2.2 Word2vec 模型

Word2vec 是一种可以进行高效率词嵌套学习的预测模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，

在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

Word2Vec 主要包括 CBOW 模型（连续词袋模型）和 Skip-gram 模型（跳字模型）。从算法角度看，这两种方法非常相似，其区别为 CBOW 根据源词上下文词汇（'the cat sits on the'）来预测目标词汇（例如，'mat'），而 Skip-gram 模型做法相反，它通过目标词汇来预测源词汇。

3 实验设计

本次实验以第一次作业所提供的 16 本金庸武侠小说作为数据集，利用 Word2Vec 模型训练 Word Embedding，并基于 Word Embedding 来进行聚类分析。

1、数据预处理：删去所有的隐藏符号，删除所有的非中文字符，不考虑上下文关系的前提下删去所有标点符号。以 jieba 库对中文语料进行分词。得到训练集和测试集。

2、运用 Word2Vec 进行训练：选用 gensim.models 包中的 word2Vec 模块，得到训练后的 Word Embedding。其中 Word2Vec 中的几个参数：

• sentences：可以是一个 list，对于大语料集，建议使用 BrownCorpus,Text8Corpus 或 LineSentence 构建。

• sg：用于设置训练算法，默认为 0，对应 CBOW 算法；sg=1 则采用 skip-gram 算法。

• vector_size：是指特征向量的维度，默认为 100。大的 size 需要更多的训练数据,但是效果会更好. 推荐值为几十到几百。

• window：表示当前词与预测词在一个句子中的最大距离是多少。

• alpha：是学习速率。

• seed：用于随机数发生器。与初始化词向量有关。

• min_count：可以对字典做截断，词频少于 min_count 次数的单词会被丢弃掉, 默认值为 5。

• max_vocab_size：设置词向量构建期间的 RAM 限制。如果所有独立单词个数超过这个，则就消除掉其中最不频繁的一个。每一千万个单词需要大约 1GB 的 RAM。设置成 None 则没有限制。

• sample：高频词汇的随机降采样的配置阈值，默认为 1e-3，范围是(0,1e-5)。

• workers：参数控制训练的并行数。

• cbow_mean：如果为 0，则采用上下文词向量的和，如果为 1（default）则采用均值。

只有使用 CBOW 的时候才起作用。

• hashfxn： hash 函数来初始化权重。默认使用 python 的 hash 函数。

• iter：迭代次数，默认为 5。

3、Kmeans 聚类：在得到训练后的词向量后，导入金庸小说全人物名称，用 Kmeans 算法进行聚类分析。

4 结果分析与总结

1、聚类结果：

将金庸小说全人物名称分为 n 类，以 n=10 为例，聚类结果如下：

1）类别 1:吴立身 白寒枫 归二娘 归辛树 归钟 何惕守 冯难敌 钱正伦 袁承志 何铁手 焦宛儿 穆人清 阿九 闵子华 程青竹 崔秋山 焦公礼 梅剑和 黄真 玉真子 温方达 崔希敏 洞玄 安小慧 温方义 温南扬 刘培生 孟伯飞 温正 罗立如 冯不摧 冯不破 孟铮 胡斐 程灵素 袁紫衣 马春花 商宝震 徐铮 商老太 汤沛 马行空 王剑英 陈禹 刘鹤真 石万嗔 阎基 汪铁鹗 慕容景岳 王剑杰 薛鹊 桑飞虹 秦耐之 曾铁鸥 姬晓峰 殷仲翔 姜铁山 孙刚峰 吕小妹 狄云 戚芳 丁典 万震山 万圭 花铁干 吴坎 戚长发 言达平 汪啸风 周圻 卜垣 鲁坤 沈城 孙均 冯坦

2）类别 2:杨过 小龙女 李莫愁 郭襄 陆无双 金轮法王 赵志敬 裘千尺 耶律齐 霍都 公孙止 潇湘子 武三通 武修文 尼摩星 尹克西 朱子柳 孙婆婆 达尔巴 樊一翁 完颜萍 武敦儒 马光佐 洪凌波 武三娘 耶律燕 林朝英 冯默风 刘处玄 祁志诚 郭靖 郭大侠 靖哥哥 黄蓉 洪七公 黄药师 周伯通 老顽童 欧阳克 梅超风 柯镇恶 杨康 朱聪 裘千仞 穆念慈 彭连虎 杨铁心 陆冠英 拖雷 王处一 梁子翁 韩小莹 华筝 韩宝驹 沙通天 侯通海 全金发 程瑶迦 鲁有脚 尹志平 一灯大师 南希仁 江南七怪 渔人 灵智上人 张阿生 陈玄风 陆乘风 谭处端 孙不二 郝大通 傻姑

3）类别 3:任我行 向问天 左冷禅 刘正风 劳德诺 祖千秋 桃根仙 莫大 桃枝仙 桃花仙 桃实仙 桃干仙 上官云 杨莲亭 平一指 秃笔翁 蓝凤凰 桃叶仙 天门道人 不戒和尚 郑萼 绿竹翁 游迅 王元霸 曲洋 鲍大楚 童百熊 玉灵道人 闻先生 张无忌 赵敏 谢逊 张翠山 周芷若 殷素素 张三丰 灭绝师太 殷梨亭 金花婆婆 小昭 俞岱岩 胡青牛 宋远桥 韦一笑 范遥 宋青书 何太冲 纪晓芙 周颠 殷天正 陈友谅 鹿杖客 杨不悔 张松溪 成昆 常遇春 丁敏君 说不得 莫声谷 彭莹玉 西华子 史火龙 韩林儿 殷野王 卫璧 白龟寿 简捷 掌棒龙头 宗维侠 执法长老 传功长老 掌钵龙头 薛公远 阳顶天 冷谦 唐文亮 司徒千钟 韩

千叶 卫四娘 彭长老

4）类别4:苗人凤 田归农 苗若兰 田青文 宝树 曹云奇 胡一刀 刘元鹤 陶子安 陶百岁 殷吉 赛总管 平阿四 范帮主 熊元献 于管家 周云阳 杜希孟 郑三娘

5）类别5:胖头陀 澄观 澄光 行痴 玉林 行颠 澄心 净济 澄通 心溪 净清 澄识 仪清 风清扬 易国梓 长青子 张君宝 天竺僧 李志常 觉远大师 空智 空闻 何足道 无色 空见 觉远 空性 渡厄 枯木 虚竹 鸠摩智 丁春秋 乌老大 玄难 玄慈 苏星河 李秋水 慕容博 薛慕华 玄寂 本因 玄苦 康广陵 波罗星 哲罗星 无崖子 天山童姥 玄生 止清 玄痛 范百龄 本参 缘根 神山上人 本观 本相 慧净 慧方 大智禅师 无青子 冲虚 木岛主

6）类别6:令狐冲 岳不群 林平之 岳灵珊 田伯光 余沧海 林震南 木高峰 段誉 阿紫 乔峰 阿朱 慕容复 王语嫣 段正淳 木婉清 游坦之 钟灵 包不同 马夫人 阿碧 段延庆 风波恶 钟万仇 朱丹臣 云中鹤 巴天石 叶二娘 邓百川 赵钱孙 全冠清 徐长老 公冶乾 阮星竹 秦红棉 白世镜 谭婆 左子穆 崔百泉 谭公 刀白凤 萧远山 司马林 范骅 单正 摘星子 褚万里 不平道人 诸保昆 过彦之 华赫艮 郁光标 岳老三 傅思归 石破天 白万剑 闵柔 丁不四 谢烟客 贝海石 丁不三 白自在 石中玉 耿万钟 封万里 高三娘子 花万紫 范一飞 王万仞 陈冲之 成自学 柯万钧 梅芳姑 齐自勉 梁自进

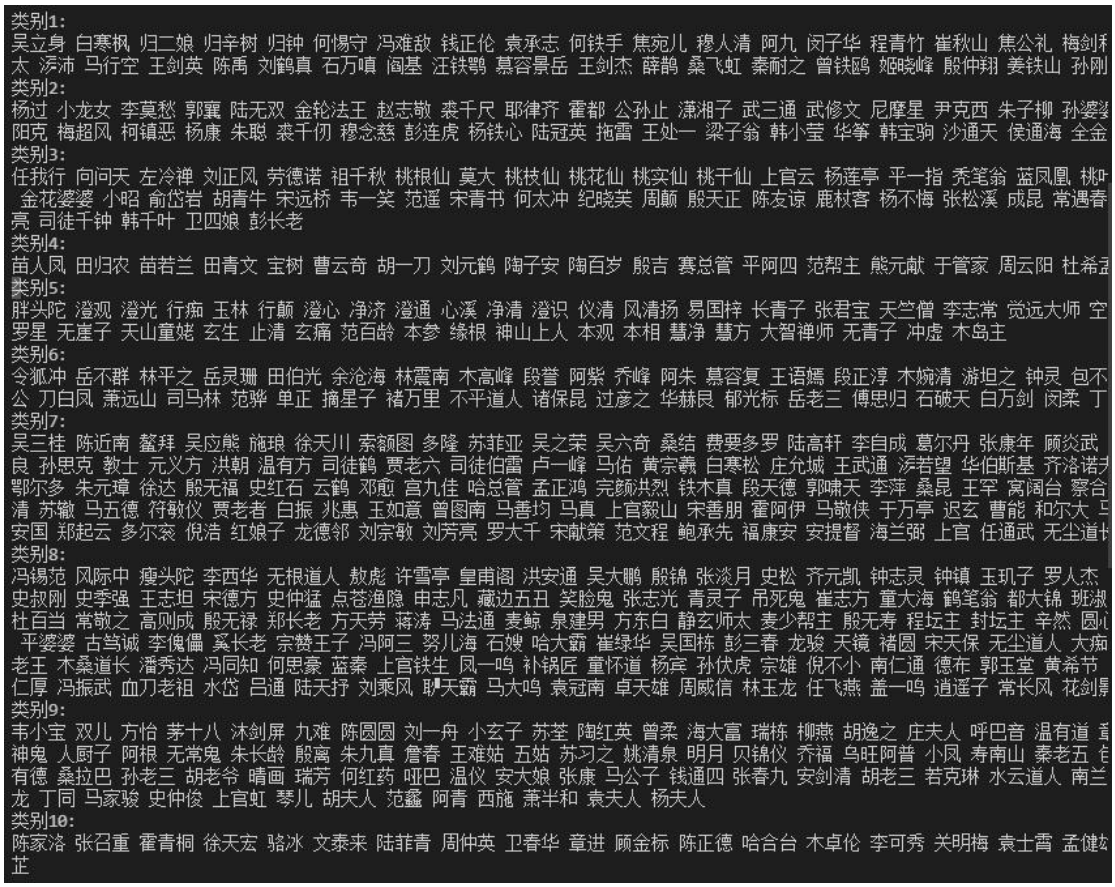
7）类别7:吴三桂 陈近南 鳌拜 吴应熊 施琅 徐天川 索额图 多隆 苏菲亚 吴之荣 吴六奇 桑结 费要多罗 陆高轩 李自成 葛尔丹 张康年 顾炎武 钱老本 杨溢之 建宁公主 李力世 柳大洪 王进宝 沐剑声 林兴珠 赵齐贤 马超兴 关安基 樊纲 夏国相 赵良栋 张勇 苏冈 图尔布青 明珠 慕天颜 佟国纲 吕留良 孙思克 教士 元义方 洪朝 温有方 司徒鹤 贾老六 司徒伯雷 卢一峰 马佑 黄宗羲 白寒松 庄允城 王武通 汤若望 华伯斯基 齐洛诺夫 南怀仁 蔡德忠 察尔珠 姚春 高里津 阿济赤 舒化龙 马博仁 祁清彪 邝天雄 于嫂 王伯奋 司马大 英额 忽必烈 耶律晋 蒙哥 耶律楚材 丁大全 萨多 陈大方 鄂尔多 朱元璋 徐达 殷无福 史红石 云鹤 邓愈 宫九佳 哈总管 孟正鸿 完颜洪烈 铁木真 段天德 郭啸天 李萍 桑昆 王罕 窝阔台 察合台 札木合 术赤 完颜洪熙 博尔忽 张十五 赤老温 博尔术 吕文德 耶律洪基 宋长老 室里 段正明 太皇太后 赫连铁树 鲍千灵 吴长风 游骥 易大彪 游驹 辛双清 苏辙 马五德 符敏仪 贾老者 白振 兆惠 玉如意 曾图南 马善均 马真 上官毅山 宋善朋 霍阿伊 马敬侠 于万亭 迟玄 曹能 和尔大 马大挺 王道 皇太后 孙克通 乾隆皇帝 呼音克 尹章垓 周阿三 袁枚 洪胜海 皇太极 张朝唐 胡桂南 孙仲寿 杨鹏举 李岩 曹化淳 祖大寿 单铁生 彼得 应松 朱安国 郑起云 多尔衮 倪浩 红娘子 龙德邻 刘宗敏 刘芳亮 罗大千 宋献策 范文程 鲍承先 福康安 安提督 海兰弼 上官 任通武 无尘道长

8）类别 8:冯锡范 风际中 瘦头陀 李西华 无根道人 敖彪 许雪亭 皇甫阁 洪安通 吴大鹏 殷锦 张淡月 史松 齐元凯 钟志灵 钟镇 玉玑子 罗人杰 玉音子 乐厚 于人豪 施戴子 梁发 严三星 辛国梁 玉磬子 史登达 王仲强 向大年 施令威 桑三娘 鲁连荣 白熊 米为义 申人俊 祝镖头 大头鬼 鹿清笃 史伯威 史叔刚 史季强 王志坦 宋德方 史仲猛 点苍渔隐 申志凡 藏边五丑 笑脸鬼 张志光 青灵子 吊死鬼 崔志方 童大海 鹤笔翁 都大锦 班淑娴 常金鹏 鲜于通 高老者 王保保 武青婴 渡难 阿三 圆音 渡劫 矮老者 妙风使 武烈流云使 吴劲草 易三娘 圆业 辉月使 卫天望 夏胄 潘天耕 阿二 祁天彪 杜百当 常敬之高则成 殷无禄 郑长老 方天劳 蒋涛 马法通 麦鲸 泉建男 方东白 静玄师太 麦少帮主 殷无寿 程坛主 封坛主 辛然 圆心 过三拳 贺老三 静照 灵虚 简长老 农夫 黎生 樵子 童子哑梢公 钱青健 梁长老 余兆兴 汤祖德 司空玄 姚伯当 桑土公 卓不凡 出尘子 瑞婆婆 龚光杰 菊剑 平婆婆 古笃诚 李傀儡 奚长老 宗赞王子 冯阿三 努儿海 石嫂 哈大霸 崔绿华吴国栋 彭三春 龙骏 天镜 褚圆 宋天保 无尘道人 大痴 贝人龙 忽伦大虎 罗信 覃天丞范中恩 阎世魁 冯辉 韩春霖 朱祖荫 大癫 大苦 平旺先 元痛 温方山 荣彩 温方悟 齐云温方施 沙老大 董开山 万里风 石骏 老王 木桑道长 潘秀达 冯同知 何思豪 蓝秦 上官铁生 凤一鸣 补锅匠 童怀道 杨宾 孙伏虎 宗雄 倪不小 南仁通 德布 郭玉堂 黄希节 倪不大 尉迟连 蔡威 文醉翁 李廷豹 欧阳公政 周隆 天虚 吴道通 大悲老人 展飞 吕正平 周牧 风良 安奉日 廖自砺 闻万夫 照虚 呼延万善 米横野 尤得胜 温仁厚 冯振武 血刀老祖水岱 吕通 陆天抒 刘乘风 耿天霸 马大鸣 袁冠南 卓天雄 周威信 林玉龙 任飞燕 盖一鸣逍遥子 常长风 花剑影

9）类别 9:韦小宝 双儿 方怡 茅十八 沐剑屏 九难 陈圆圆 刘一舟 小玄子 苏荃 陶红英 曾柔 海大富 瑞栋 柳燕 胡逸之 庄夫人 呼巴音 温有道 章老三 崔瞎子 郑克爽 仪琳陆大有 曲非烟 史镖头 哑婆婆 易师爷 郑镖头 王家骏 陈七 王家驹 林远图 农妇 刘芹老不死 公孙绿萼 陆二娘 郭破虏 柔儿 煞神鬼 人厨子 阿根 无常鬼 朱长龄 殷离 朱九真詹春 王难姑 五姑 苏习之 姚清泉 明月 贝锦仪 乔福 乌旺阿普 小凤 寿南山 秦老五 包惜弱 裘千丈 曲三 简管家 瘦丐 甘宝宝 梅剑 智光大师 兰剑 严妈妈 吴光胜 竹剑 穆贵妃 来福儿 幽草 梦姑 小翠 喀丝丽 周大奶奶 曹司朋 周英杰 阿里 方有德 桑拉巴 孙老三 胡老爷 晴画 瑞芳 何红药 哑巴 温仪 安大娘 张康 马公子 钱通四 张春九 安剑清 胡老三 若克琳 水云道人 南兰 王铁匠 聂钺 平四 胖商人 张飞雄 瘦商人 侍剑 史小翠 胡大哥 梅文馨 水笙 宝象 凌退思 桃红 梅念笙 李文秀 苏普 阿曼 苏鲁克 陈达海 车尔库李三 霍元龙 丁同 马家骏 史仲俊 上官虹 琴儿 胡夫人 范蠡 阿青 西施 萧半和 袁夫人

杨夫人

10）类别 10:陈家洛 张召重 霍青桐 徐天宏 骆冰 文泰来 陆菲青 周仲英 卫春华 章进 顾金标 陈正德 哈合台 木卓伦 李可秀 关明梅 袁士霄 孟健雄 杨成协 童兆和 王维扬 言伯乾 韩文冲 蒋四根 焦文期 瑞大林 阎世章 安健刚 成璜 万庆澜 乾隆 阿凡提 赵半山 心砚 石双英 常伯志 常赫志 余鱼同 李沅芷



2、以五本小说的主人公为例，列出与其关系最相近的 10 个词汇：

1）与乔峰关系相近的 10 个词汇:[('谭公', 0.5746119618415833), ('萧峰', 0.5734392404556274), ('徐长老', 0.5710658431053162), ('马夫人', 0.5645152926445007), ('阿朱', 0.5623595714569092), ('单正', 0.5512268543243408), ('谭婆', 0.5444244742393494), ('赵钱孙', 0.5398303866386414), ('乔帮主', 0.5395501852035522), ('智光', 0.5285753011703491)]

2）与杨过关系相近的 10 个词汇:[('小龙女', 0.6759870052337646), ('金轮法王', 0.5581438541412354), ('陆无双', 0.5533270835876465), ('公孙谷主', 0.5470530986785889), ('过儿', 0.5363024473190308), ('樊一翁', 0.5357151031494141), ('二武', 0.5281738638877869), ('霍都', 0.5231220126152039), ('杨过知', 0.5211523771286011), ('洪凌波', 0.5135650634765625)]

3）与张无忌关系相近的 10 个词汇:[('周芷若', 0.5977776646614075), ('赵敏', 0.5768000483512878), ('小昭', 0.5212957859039307), ('蛛儿', 0.4958591461181640 6), ('杨不悔', 0.47185859084129333), ('令狐冲', 0.45891350507736206), ('赵姑娘', 0.4583019018173218), ('宋青书', 0.45776277780532837), ('的', 0.45664215087890625), ('任我行', 0.4517929255962372)]

4）与郭靖关系相近的 10 个词汇:[('黄蓉', 0.6199606657028198), ('欧阳锋', 0.5480539202690125), ('郭靖道', 0.5408143401145935), ('拖雷', 0.5359072685241699), ('华筝', 0.5355803966522217), ('黄药师', 0.5333465933799744), ('蓉', 0.513937771320343), ('杨康', 0.5103278756141663), ('黄蓉道', 0.4947511851787567), ('洪七公', 0.4910281300544739)]

5）与韦小宝关系相近的 10 个词汇:[('索额图', 0.5446203947067261), ('韦小', 0.5411296486854553), ('张康年', 0.5389895439147949), ('茅十八', 0.5282501578330994), ('施琅', 0.5245674252510071), ('康熙', 0.508659303188324), ('海老公', 0.5030286312103271), ('双儿', 0.5020745992660522), ('洪夫人', 0.5019695162773132), ('方怡', 0.49781379103660583)]

```
与乔峰关系相近的10个词汇:
[('谭公', 0.5746119618415833), ('萧峰', 0.5734392404556274), ('徐长老', 0.5710658431053162), ('马夫人', 0.5645152926445007), ('阿朱', 0.5623595714569092), ('单正', 0.5512268543243408), ('谭婆', 0.5444244742393494), ('赵钱孙', 0.5398303866386414), ('乔帮主', 0.5395501852035522), ('智光', 0.5285753011703491)]
与杨过关系相近的10个词汇:
[('小龙女', 0.6759870052337646), ('金轮法王', 0.5581438541412354), ('陆无双', 0.5533270835876465), ('公孙谷主', 0.5470530986785889), ('过儿', 0.5363024473190308), ('樊一翁', 0.5357151031494141), ('二武', 0.5281738638877869), ('霍都', 0.5231220126152039), ('杨过知', 0.5211523771286011), ('洪凌波', 0.5135650634765625)]
与张无忌关系相近的10个词汇:
[('周芷若', 0.5977776646614075), ('赵敏', 0.5768000483512878), ('小昭', 0.5212957859039307), ('蛛儿', 0.4958591461181640 6), ('杨不悔', 0.47185859084129333), ('令狐冲', 0.45891350507736206), ('赵姑娘', 0.4583019018173218), ('宋青书', 0.45776277780532837), ('的', 0.45664215087890625), ('任我行', 0.4517929255962372)]
与郭靖关系相近的10个词汇:
[('黄蓉', 0.6199606657028198), ('欧阳锋', 0.5480539202690125), ('郭靖道', 0.5408143401145935), ('拖雷', 0.53590726852416699), ('华筝', 0.5355803966522217), ('黄药师', 0.5333465933799744), ('蓉', 0.513937771320343), ('杨康', 0.5103278756141663), ('黄蓉道', 0.4947511851787567), ('洪七公', 0.4910281300544739)]
与韦小宝关系相近的10个词汇:
[('索额图', 0.5446203947067261), ('韦小', 0.5411296486854553), ('张康年', 0.5389895439147949), ('茅十八', 0.5282501578330994), ('施琅', 0.5245674252510071), ('康熙', 0.508659303188324), ('海老公', 0.5030286312103271), ('双儿', 0.5020745992660522), ('洪夫人', 0.5019695162773132), ('方怡', 0.49781379103660583)]
```

3、以五个门派为例，列出与其关系最相近的 10 个词汇：

1）与逍遥派关系相近的 10 个词汇:[('七十余年', 0.7043185234069824), ('杂学', 0.6963227987289429), ('别派', 0.6959677934646606), ('学练', 0.6888591647148132), ('改投', 0.685207664966583 3), ('邪气', 0.6814152002334595), ('折梅手', 0.678023636341095), ('亲授', 0.6755051016807556), ('弃徒', 0.6711840033531189), ('铁剑门', 0.6678870916366577)]

2）与全真教关系相近的 10 个词汇:[('重阳', 0.6720374226570129), ('创教', 0.6526528000831604), ('全真', 0.6425071954727173), ('反出', 0.6379581689834595), ('掌教', 0.6345388889312744), ('丘道长', 0.6181472539901733), ('子马钰', 0.6121949553489685), ('我门', 0.6108074188232422), ('首徒', 0.6050494909286499), ('郝大通', 0.6022130846977234)]

3）与明教关系相近的 10 个词汇:[('明教中', 0.6419917345046997), ('四分五裂', 0.6335322856903076), ('总教', 0.6259443759918213), ('光明顶', 0.6175218224525452), ('右使', 0.6127390265464783), ('左使', 0.6122970581054688), ('正教', 0.6043582558631897), ('护教', 0.6040052771568298), ('教规', 0.6018332242965698), ('六大', 0.5983508229255676)]

4）与少林关系相近的 10 个词汇:[('高僧', 0.6759040355682373), ('僧众', 0.6277364492416382), ('武当', 0.6066418886184692), ('群僧', 0.60446697473526), ('本寺', 0.6042309403419495), ('少林寺', 0.6024782061576843), ('下院', 0.6001448035240173), ('于少林', 0.599332332611084), ('神僧空', 0.5992250442504883), ('十八罗汉', 0.5976754426956177)]

5）与华山派关系相近的 10 个词汇:[('弃徒', 0.6405172348022461), ('岳先生', 0.6217259764671326), ('本门', 0.595853865146637), ('华山', 0.588897168636322), ('投师', 0.578190267086029), ('带艺', 0.5764662027359009), ('门下', 0.5759971141815186), ('气宗', 0.5682054162025452), ('首徒', 0.5666701793670654), ('剑宗', 0.5652083158493042)]

```
与逍遥派关系相近的10个词汇:
[('七十余年', 0.7043185234069824), ('杂学', 0.6963227987289429), ('另派', 0.6959677934646606), ('学练', 0.6888591647148184003353 1189), ('铁剑门', 0.6678870916366577)]
与全真教关系相近的10个词汇:
[('重阳', 0.6720374226570129), ('创教', 0.6526528000831604), ('全真', 0.6425071954727173), ('反出', 0.6379581689834595), 94909286499), ('郝大通', 0.6022130846977234)]
与明教关系相近的10个词汇:
[('明教中', 0.6419917345046997), ('四分五裂', 0.6335322856903076), ('总教', 0.6259443759918213), ('光明顶', 0.6175218224 525452), ('右使', 0.6127390265464783), ('左使', 0.6122970581054688), ('正教', 0.6043582558631897), ('护教', 0.6040052771 568298), ('教规', 0.6018332242965698), ('六大', 0.5983508229255676)]
与少林关系相近的10个词汇:
[('高僧', 0.6759040355682373), ('僧众', 0.6277364492416382), ('武当', 0.6066418886184692), ('群僧', 0.60446697473526), ( '本寺', 0.6042309403419495), ('少林寺', 0.6024782061576843), ('下院', 0.6001448035240173), ('于少林', 0.599332332611084) , ('神僧空', 0.5992250442504883), ('十八罗汉', 0.5976754426956177)]
与华山派关系相近的10个词汇:
[('弃徒', 0.6405172348022461), ('岳先生', 0.6217259764671326), ('本门', 0.595853865146637), ('华山', 0.588897168636322),  ('投师', 0.578190267086029), ('带艺', 0.5764662027359009), ('门下', 0.5759971141815186), ('气宗', 0.5682054162025452), ('首徒', 0.5666701793670654), ('剑宗', 0.5652083158493042)]
```

3、结果分析：词向量训练结果较好，最终聚类得到的结果基本与小说内容相符，人物与门派之间的关系关联度均较高，符合小说的实际情况。

5 代码链接

见 https://github.com/ErrricCai/DL-NLP/tree/main/HW4

参考：

https://blog.csdn.net/weixin_42663984/article/details/116739799?spm=1001.2014.3001.5502

https://blog.csdn.net/weixin_44966965/article/details/124732760?spm=1001.2014.3001.5502

https://blog.csdn.net/weixin_50891266/article/details/116750204?spm=1001.2014.3001.5502