形状, 圆圈

描述已自动生成

**卡通人物

中度可信度描述已自动生成**

**深度学习与自然语言处理第四次作业**

词向量聚类

|  |  |
| --- | --- |
| 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 | 彭程 |
| 学号 | ZY2103807 |
| 指导老师 | 秦曾昌 |

2022年 5月

# 实验背景

## 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。

## 实验原理

**1.2.1 词向量**

词向量（Word Embedding），是一种函数映射关系。在机器学习中，特征都是以数值的形式进行传递的。同样的，在NLP中，文本特征也需要被映射成数值向量。其过程是：将文本空间中的某个词（word），通过一定的方法，映射或嵌入到另一个数值向量空间。

**1.2.2 词向量的表示形式**

1. One-hot编码

One-hot编码是一种简单的词向量表示形式。具体地，用很长的一个向量来表示一个词，其中向量的长度为词典的大小；向量的所有分量只有一个为“1”，其余全为“0”，为“1”的位置为该词在词典中的字典序。

1. 分布式表示（Distributed Representation）

分布式表示引入了“距离”的概念从而可以衡量词与词之间的联系，这对建模自然语言的语义信息大有裨益；此外，有一个多维的词向量而非只包含0、1的one-hot向量对词语建模，使得向量可以包含更为丰富的语义信息。

**1.2.3 Word2Vec模型**

Word2Vec是一种可以进行高效率词嵌套学习的预测模型。其模型仅仅包括输入层、隐藏层和输出层，模型框架根据输入输出的不同，主要包括CBOW和Skip-gram模型。 CBOW的方式是在知道词的上下文的情况下预测当前词。而skip-gram实在知道了词的情况下，对词的上下文进行预测。

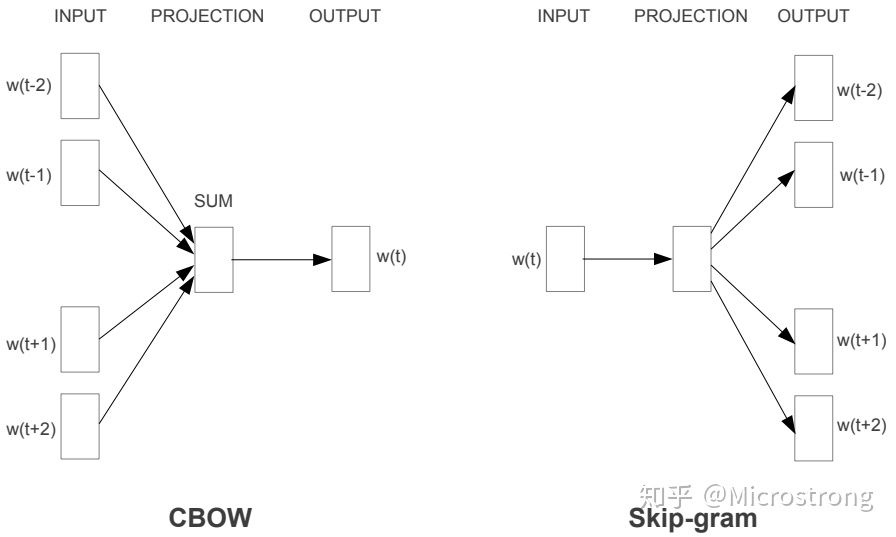


图1. CBOW和Skip-gram结构

**1.2.4 K-Means聚类**

K-Means 是一种典型的聚类算法，算法步骤如下：

1. 选取K个点作为初始聚集的簇心（也可选择非样本点）;
2. 分别计算每个样本点到 K个簇核心的距离，找到离该点最近的簇核心，将它归属到对应的簇；
3. 所有点都归属到簇之后，M个点就分为了 K个簇。之后重新计算每个簇的重心（平均距离中心），将其定为新的“簇核心”；
4. 反复迭代 2 - 3 步骤，直到达到中止条件。

# 实验方法

本实验采用的如下思路进行词向量训练。

1. 与前几次作业类似，采用jieba分词对数据集进行分词处理，划分训练集和测试集；
2. 运用gensim.models中的Word2Vec模型进行词向量训练；
3. 训练完成后，采用K-Means的方法进行聚类分析。

# 实验结果与分析

## 人名聚类结果

将金庸小说中所有人名分成16类，分类结果如下：

类别1:

李文秀 苏普 阿曼 苏鲁克 陈达海 车尔库 李三 霍元龙

类别2:

曲洋 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 朱长龄 成昆 常遇春 丁敏君 殷离 说不得 莫声谷 彭莹玉 韩林儿 殷野王 简捷 鲜于通 薛公远 阳顶天 王难姑

类别3:

苏菲亚 李自成 图尔布青 忽必烈 蒙哥 耶律楚材 鄂尔多 铁木真 杨铁心 拖雷 华筝 包惜弱 段天德 郭啸天 李萍 桑昆 王罕 窝阔台 察合台 札木合 术赤 博尔忽 赤老温 博尔术 吕文德 耶律洪基 兆惠 喀丝丽 阿里 霍阿伊 桑 拉巴 曹能 和尔大 孙克通 平旺先 皇太极 孙仲寿 李岩 祖大寿 彼得 应松 红娘子 钱通四 罗大千 范蠡

类别4:

无根道人 敖彪 胡逸之 许雪亭 皇甫阁 吴大鹏 殷锦 张淡月 史松 冯难敌 钟志灵 天门道人 钟镇 罗人杰 乐厚 于人豪 梁发 辛国梁 玉磬子 史登达 桑三娘 鲁连荣 白熊 申人俊 鹿清笃 冯默风 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 笑脸鬼 祁志诚 张志光 青灵子 吊死鬼 崔志方 童大海 鹤笔翁 班淑娴 西华子 史火龙 卫璧 白龟寿 掌棒龙头 常金鹏 高老者 王保保 武青婴 宗维侠 执法长老 传功长老 渡难 掌钵龙头 阿三 圆音 渡劫 矮老者 妙风使 武烈 冷谦 流云使 唐文亮 吴劲草 司徒千钟 圆业 辉月使 卫天望 夏胄 潘天耕 阿二 杜百当 常敬之 高则成 殷无禄 郑长老 卫四娘 方天劳 蒋涛 马法通 麦鲸 泉建男 方东白 静玄师太 麦少帮主 辛 然 圆心 过三拳 灵虚 简长老 黎生 哑梢公 钱青健 梁长老 汤祖德 司马林 诸保昆 姚伯当 桑土公 卓不凡 瑞婆婆 龚光杰 平婆婆 古笃诚 李傀儡 奚长老 冯阿三 哈大霸 崔绿华 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 元痛 玉真子 洞玄 温方山 温方义 温南扬 刘培生 荣彩 温方悟 温方施 沙老大 冯不摧 冯不破 石骏 潘秀达 汤沛 王剑英 陈禹 王剑杰 桑飞虹 秦耐之 曾铁鸥 殷仲翔 姜铁山 何思豪 海兰弼 蓝秦 孙刚峰 吕小妹 上官铁生 凤一鸣 童怀道 杨宾 孙伏虎 宗雄 倪不小 德布 黄希节 倪不大 尉迟连 李廷豹 欧阳公政 周隆 封万里 王万仞 天虚 吴道通 成自学 大悲老人 展飞

吕正平 周牧 柯万钧 风良 安奉日 廖自砺 齐自勉 闻万夫 照虚 梁自进 呼延万善 尤得胜 冯振武 吕通 陆天抒 刘乘风 孙均 耿天霸 马大鸣 冯坦

类别5:

韦小宝 吴三桂 鳌拜 吴应熊 施琅 索额图 多隆 吴之荣 陈圆圆 费要多罗 小玄子 葛尔丹 张康年 杨溢之 建宁公主 王进宝 林兴珠 赵齐贤 夏国相 赵良栋 张勇 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 司徒伯雷 卢一峰 马佑 黄宗羲 庄允城 汤若望 华伯斯基 齐洛诺夫 南怀仁 察尔珠 高里津 阿济赤 祁清彪 邝天雄 丁大全 陈大方 朱元璋 徐达 邓愈 段正明 太皇太后 赫连铁树 易大彪 苏辙 玉如意 于万亭 王道 皇太

后 乾隆皇帝 呼音克 尹章垓 袁枚 张朝唐 曹化淳 多尔衮 刘宗敏 宋献策 范文程 鲍承先 福康安

类别6:

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

类别7:

石万嗔 慕容景岳 薛鹊 狄云 水笙 戚芳 丁典 万震山 万圭 花铁干 吴坎 血刀老祖 戚长发 宝象 言达平 水岱 汪啸风 周圻 卜垣 鲁坤 沈城 凌退思 桃红

类别8:

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 尼摩星 尹克西 朱子柳 孙婆婆 达尔巴 樊一翁 洪凌波 耶律燕 刘处玄 郭靖 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 完颜洪烈 裘千仞 穆念慈 彭连虎 陆冠英 王处一 梁子翁 韩小莹 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 灵智上人 张阿生 陈玄风 陆乘风 谭处端 孙不二 郝大通 傻姑

类别9:

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 慕容博 谭婆 左子穆 崔百泉 谭公 刀白凤 萧远山 范骅 单正 褚万里 不平道人 过彦之 华赫艮 甘宝宝 宋长老 郁光标 傅思归 耿万钟 高三娘子 范一飞

类别10:

归二娘 归辛树 何惕守 袁承志 何铁手 焦宛儿 洪胜海 穆人清 阿九 闵子华 程青竹 何红药 崔秋山 焦公礼 梅剑和 黄真 哑巴 胡桂南 温方达 崔希敏 温仪 安大娘 安小慧 单铁生 罗立如 朱安国

类别11:

令狐冲 岳不群 林平之 岳灵珊 仪琳 田伯光 任我行 向问天 左冷禅 余沧海 刘正风 劳德诺 祖千秋 陆大有 桃根仙 木高峰 莫大 桃枝仙 桃花仙 桃实仙 桃干仙 仪清 曲非烟 平一指 秃笔翁 蓝凤凰 桃叶仙 不戒和尚 郑萼 风清扬 玉玑子 玉音子 施戴子 丁春秋 乌老大 苏星河 李秋水 薛慕华 摘星子 康广陵 无崖子 出尘子 范百龄 冲虚

类别12:

武修文 完颜萍 武敦儒 马光佐 公孙绿萼 武三娘 胡斐 程灵素 袁紫衣 苗人凤 马春花 田归农 商宝震 徐铮 商老太 马行空 刘鹤真 阎基 补锅匠 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 琴儿 袁冠南 卓天雄 周威信 林玉龙 任飞燕 盖一鸣

类别13:

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 白振 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 曾图南 吴国栋 马善均 钱正伦 上官毅山 马大挺 乾隆 阿凡提 杨鹏举 赵半山 心砚 石双英 常伯志 常赫志 余鱼同 李沅芷

类别14:

陈近南 茅十八 徐天川 冯锡范 吴立身 吴六奇 风际中 白寒枫 苏荃 顾炎武 钱老本 李力世 柳大洪 沐剑声 马超兴 关安基 樊纲 李西华 苏冈 贾老六 白寒松 王武通 马博仁 郑克爽 孟伯飞 汪铁鹗

类别15:

澄观 桑结 澄光 行痴 玉林 行颠 澄心 净济 澄通 心溪 净清 澄识 易国梓 张君宝 林朝英 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 玄难 玄慈 玄寂 本因 玄苦 波罗星 哲罗星 玄生 止清 玄痛 本参 缘根 神山上人 本观 本相 慧方 大智禅师 无青子

类别16:

双儿 方怡 沐剑屏 胖头陀 九难 陆高轩 刘一舟 瘦头陀 陶红英 曾柔 海大富 瑞栋 归钟 柳燕 洪安通 庄夫人 呼巴音 蔡德忠 温有道 姚春 齐元凯 章老三 舒化龙 崔瞎子 林震南 上官云 杨莲亭 绿竹翁 游迅 王元霸 鲍大楚 童百熊 于嫂 玉灵道人 史镖头 哑婆婆 易师爷 郑镖头 王家骏 严三星 陈七 王家驹 林远图 王仲强 向大年 施令威 王伯奋 司马大 长青子 米为义 闻先生 英颚 农妇 刘芹 老不死 祝镖头 大头鬼 耶律晋 陆二娘 藏边五丑 郭破虏 柔儿 煞神鬼 萨多 人厨子 阿根 无常鬼 朱九真 都大锦 詹春 五姑 易三娘 苏习之 姚清泉 明月 祁天彪 殷无福 贝锦仪 乔福 史红石 韩千叶 云鹤 乌旺阿普 小凤 殷无寿 程坛主 封坛主 寿南山 宫九佳 贺老三 哈总管 秦老五 静照 孟正鸿 郭大侠 渔人 农夫 彭长老 完颜洪熙 樵子 裘千丈 曲三 简管家 张十五 童子 余兆兴 瘦丐 司空玄 天山童姥 室里 梅剑 岳老三 智光大师 兰剑 菊剑 鲍千灵 严妈妈 吴长风 吴光胜 游骥 慧净 宗赞王子 竹剑 游 驹 穆贵妃 辛双清 努儿海 石嫂 来福儿 马五德 幽草 梦姑 符敏仪 贾老者 小翠 周大奶奶 马真 曹司朋 周英杰 宋善朋 方有德 马敬侠 迟玄 孙老三 胡老爷 晴画 瑞芳 周阿三 温正 张康 齐云 郑起云 倪浩 龙德邻 马公子 张 春九 董开山 安剑清 胡老三 刘芳亮 孟铮 若克琳 万里风 水云道人 老王 木桑道长 冯同知 安提督 姬晓峰 上官 南兰 任通武 王铁匠 南仁通 郭玉堂 聂钺 蔡威 文醉翁 平四 胖商人 张飞雄 无尘道长 瘦商人 侍剑 花万紫 陈 冲之 木岛主 梅芳姑 米横野 史小翠 胡大哥 梅文馨 温仁厚 梅念笙 丁同 马家骏 史仲俊 上官虹 胡夫人 阿青 西施 萧半和 逍遥子 常长风 花剑影 袁夫人 杨夫人

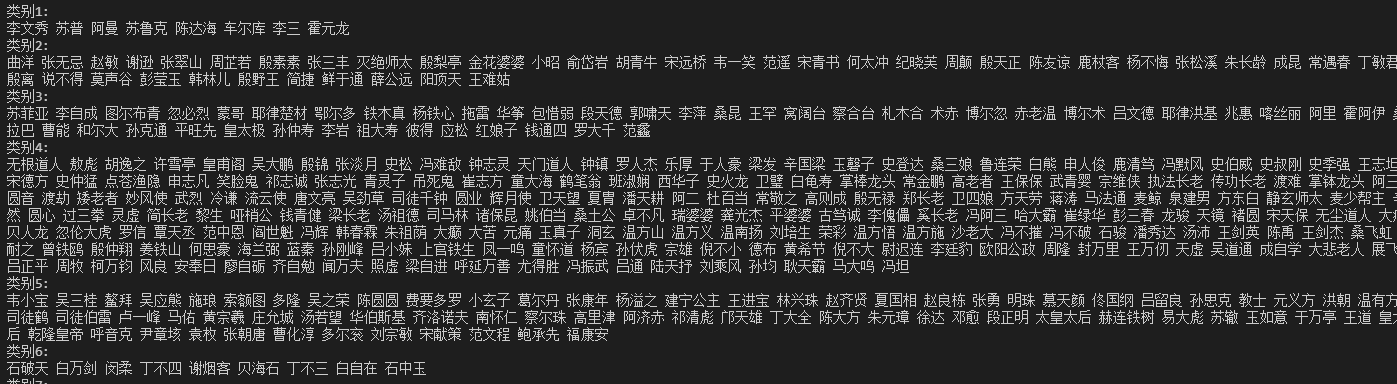


图1. 人名聚类结果

## 主人公关系检测

以小说中的男女主人公各5位，分别列出与其关系最接近的10个词语，检测结果如下：

与狄云关系相近的10个词汇:

[('丁典', 0.700218915939331), ('水笙', 0.6870795488357544), ('花铁干', 0.6746068596839905), ('戚芳', 0.6455783843994141), ('万圭', 0.630989670753479), ('戚长发', 0.6083515286445618), ('万震山', 0.5971252918243408), ('血刀', 0.5920051336288452), ('水岱', 0.5842193365097046), ('宝象', 0.5779240131378174)]

与杨过关系相近的10个词汇:

[('小龙女', 0.7263299226760864), ('金轮法王', 0.5793139338493347), ('陆无双', 0.5677275061607361), ('过儿', 0.5586572885513306), ('神雕侠侣', 0.5476442575454712), ('法王', 0.5469890236854553), ('杨过正

', 0.5467565059661865), ('李莫愁', 0.5367921590805054), ('公孙谷主', 0.5333265066146851), ('返回', 0.5246555209159851)]

与张无忌关系相近的10个词汇:

[('赵敏', 0.5924034118652344), ('周芷若', 0.5618893504142761), ('小昭', 0.5018176436424255), ('蛛儿', 0.4994775056838989), ('韦一笑', 0.49287089705467224), ('杨不悔', 0.492387056350708), ('赵姑娘', 0.4772055149078369), ('明教', 0.46578261256217957), ('赵敏道', 0.45743316411972046), ('灭绝师太', 0.445392370223999)]

与郭靖关系相近的10个词汇:

[('黄蓉', 0.6384910345077515), ('柯镇恶', 0.5225673913955688), ('拖雷', 0.5145242214202881), ('蓉', 0.5028324723243713), ('华筝', 0.49208447337150574), ('郭靖自', 0.4894099235534668), ('欧阳锋', 0.48843854665756226), ('黄药师', 0.48703575134277344), ('郭靖正', 0.4825728237628937), ('金轮法王', 0.475029855966568)]

与萧峰关系相近的10个词汇:

[('耶律洪基', 0.5813088417053223), ('阿紫', 0.5456578731536865), ('游坦之', 0.5182104706764221), ('阿骨', 0.5102747082710266), ('乔峰', 0.5085758566856384), ('阿朱', 0.5079723000526428), ('契丹', 0.5012590289115906), ('紫', 0.4992639422416687), ('辽国', 0.49379417300224304), ('萧峰心', 0.48918840289115906)]

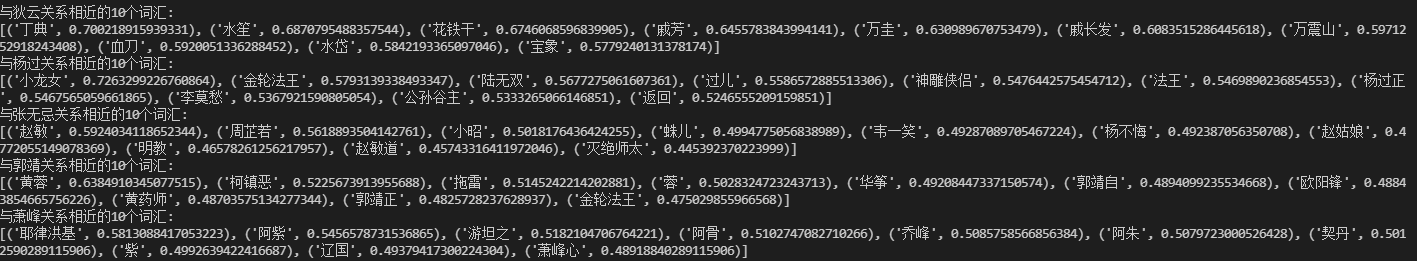


图2 男主人公关系检测结果

与小昭关系相近的10个词汇:

[('黛绮丝', 0.6254581212997437), ('不悔', 0.5591026544570923), ('周芷若', 0.5517778992652893), ('杨不悔', 0.5515418648719788), ('张公子', 0.5507571697235107), ('赵姑娘', 0.5469911694526672), ('小昭道', 0.5459510087966919), ('赵敏', 0.536781907081604), ('蛛儿', 0.5362133979797363), ('赵敏道', 0.5359558463096619)]

与周芷若关系相近的10个词汇:

[('赵敏', 0.6280105113983154), ('张无忌', 0.6034364700317383), ('灭绝师太', 0.5799347162246704), ('殷离', 0.5641778111457825), ('道芷若', 0.558218240737915), ('金花婆婆', 0.5526871085166931), ('小昭',

0.5517779588699341), ('宋青书', 0.5336641669273376), ('蛛儿', 0.5321539044380188), ('赵姑娘', 0.5128013491630554)]

与小龙女关系相近的10个词汇:

[('杨过', 0.7150152921676636), ('过儿', 0.6214496493339539), ('姑姑', 0.5915101766586304), ('古墓', 0.591119110584259), ('孙婆婆', 0.5771390795707703), ('李莫愁', 0.5766944289207458), ('神雕侠侣', 0.5655080676078796), ('陆无双', 0.5640255808830261), ('金轮法王', 0.5619088411331177), ('洪凌波', 0.5495826005935669)]

与王语嫣关系相近的10个词汇:

[('王姑娘', 0.6779236793518066), ('表哥', 0.6634266376495361), ('段誉', 0.6531925201416016), ('段公子', 0.6412140130996704), ('慕容复', 0.6211093664169312), ('誉', 0.5825321078300476), ('段', 0.5382914543151855), ('对段', 0.5359795689582825), ('阿碧', 0.5163213014602661), ('鸠摩智', 0.5116162300109863)]

与黄蓉关系相近的10个词汇:

[('郭靖', 0.6287184357643127), ('黄药师', 0.6053279042243958), ('黄蓉道', 0.5906912684440613), ('靖哥哥', 0.5800280570983887), ('欧阳克', 0.5744040012359619), ('黄蓉见', 0.5694435238838196), ('杨康', 0.5527618527412415), ('鲁有脚', 0.5413798689842224), ('道蓉儿', 0.5246714949607849), ('黄蓉知', 0.5200697183609009)]

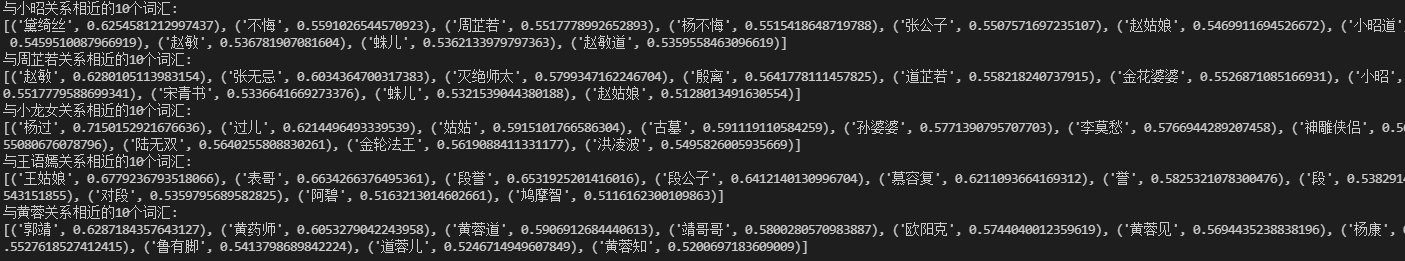


图3 女主人公关系检测结果

## 结果分析

通过以上实验可以看出，处于同一个类别中的名字基本来源于同一本小说；而在主人公关系检测中，所得到的结果于人名关系较为密切，可以看出词向量的训练结果较好。

**附录：**

**实验代码见工程文件夹。**

GetData对文本数据进行处理；

Word2Vec 为主函数，包括训练过程和测试结果。