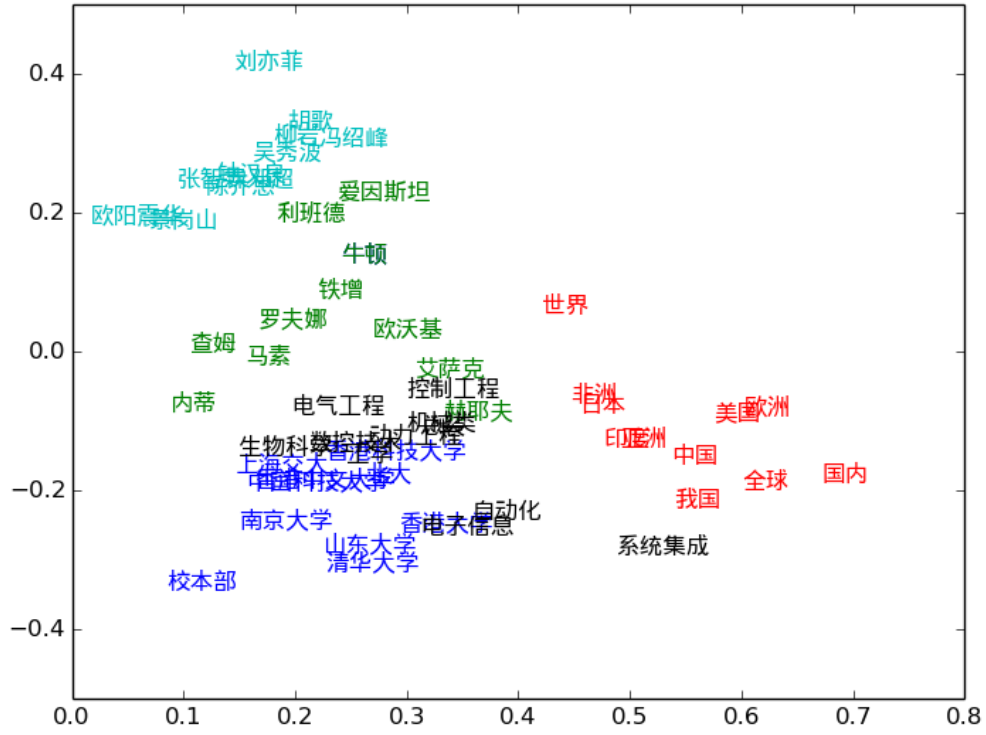
AIDemo

理解你的语言

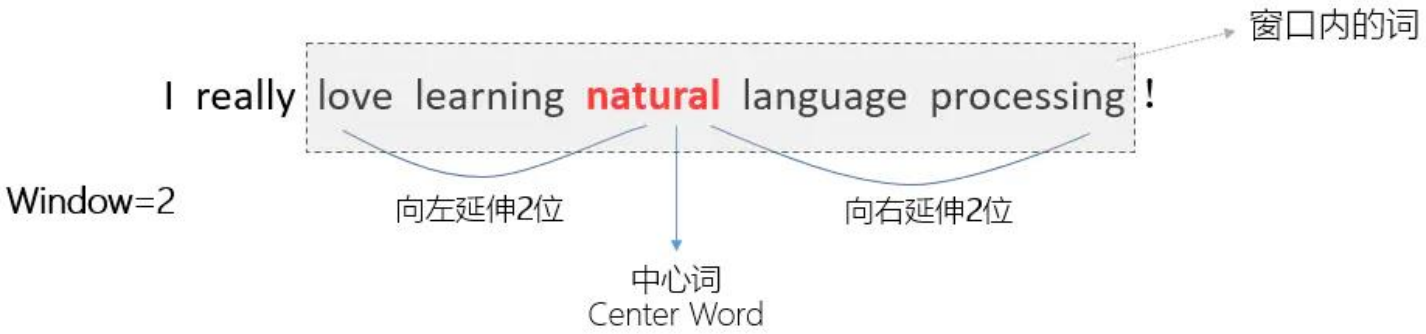
李蓝天

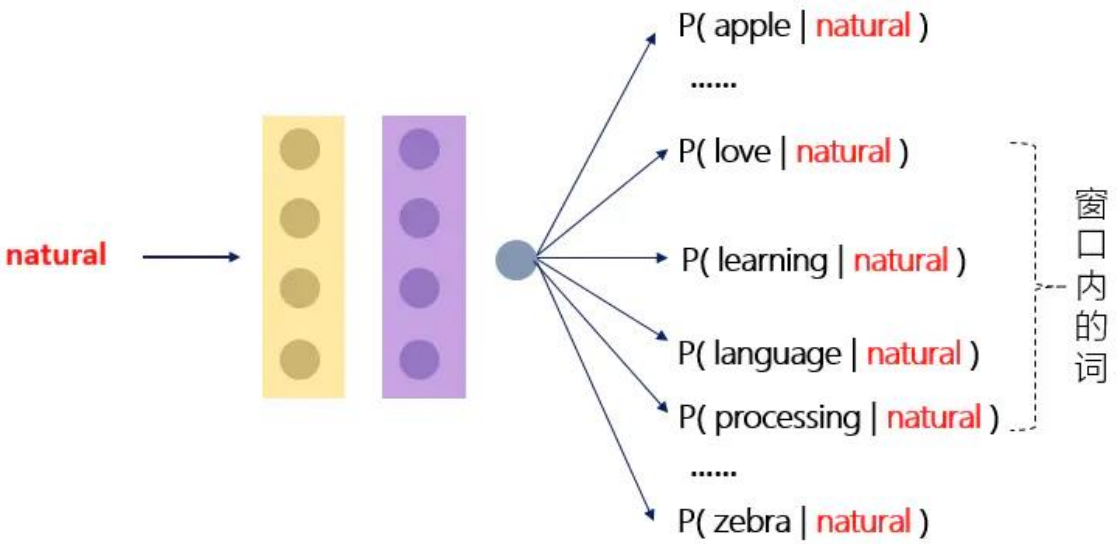
词向量

- 词向量是单词在一个连续语义空间上的映射。
- 语义接近的单词所对应的向量间距离更小。
- 语义空间低维、连续、可计算



基本思路





计算窗口内的词的概率的乘积：

P(love | natural)
× P(learning | natural)
× P(language | natural)
× P(processing | natural)

由于窗口内的这几个词是真实的语言组合，所以我们希望它们组合的概率尽可能大。

训练准则

窗口内的词

↓

语言组合

↓

组合概率高

实验一：生成《射雕英雄转》词向量

- 文本预料：金庸小说《射雕英雄传》
- 文本分词：Jieba 分词工具
- 构建词表：选出高频词汇（出现频率超过 10次 的单词）
- 训练词向量：利用 Word2Vec 工具，生成词向量
- 可视化：利用 t-SNE 工具，将词向量投影到二维空间。

实验步骤

- 认真阅读 word2vce/lang/doc/README，了解实验步骤。
  - 可在图形界面下双击打开，也可在终端下用 vim 打开
  - 关于 vim 的操作，请 aibook.cslt.org 下 Linux Shell 简易教程
- 打开终端，进入 lang 的 code 目录。
- 运行缺省程序

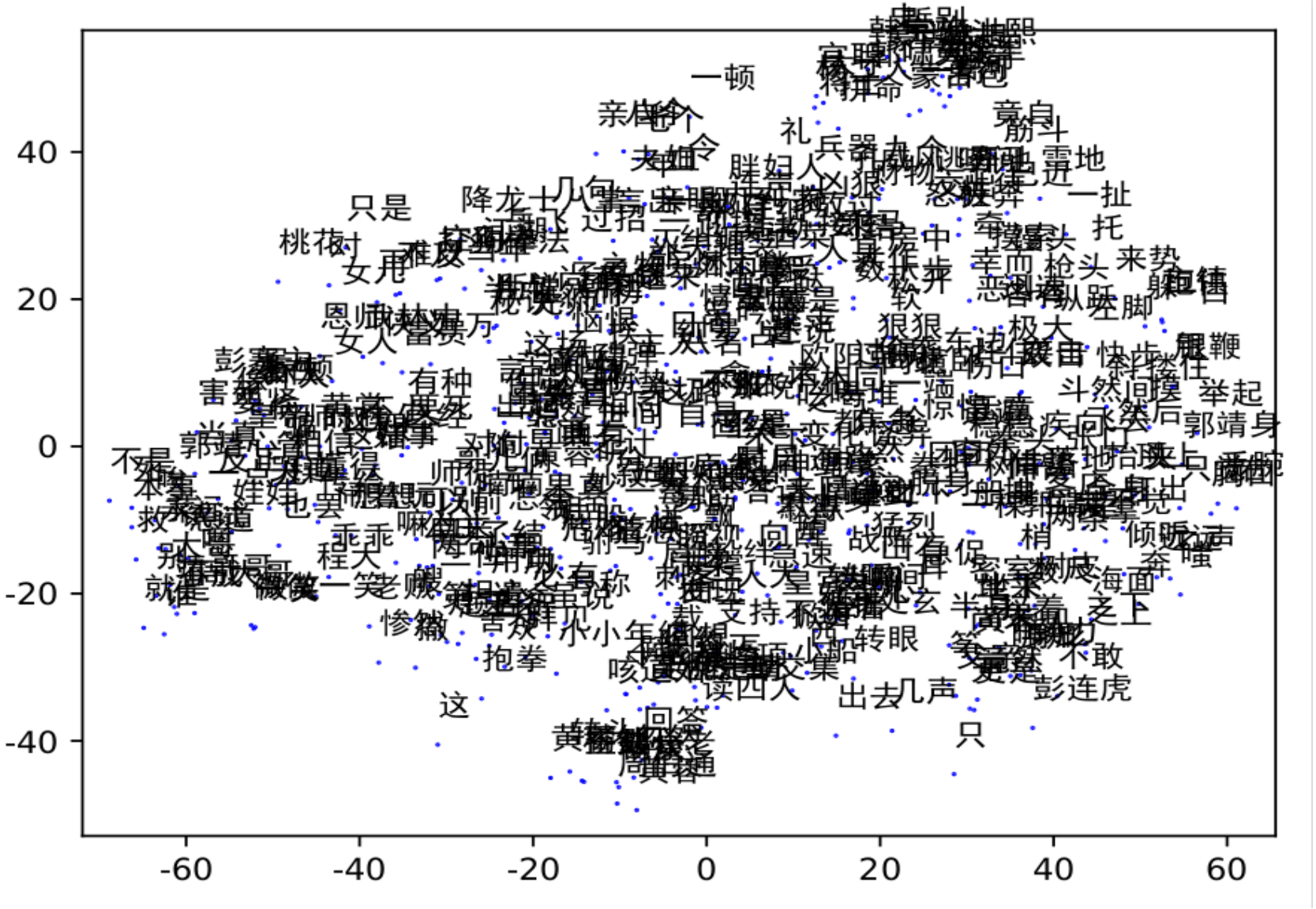
sh run.sh

- 运行上述命令得到基于《射雕英雄传》中所有文本词向量的二维映射图片，关闭图片后自动将图片转为 PDF 保存。

运行界面

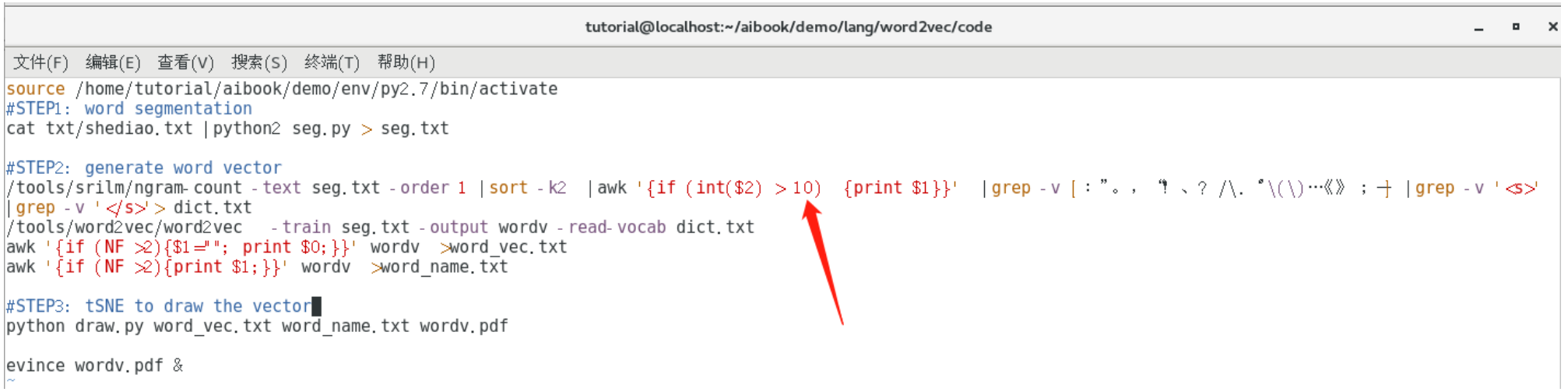
- 进入到 lang 的子文件夹 code 中运行**run.sh**

实验结果

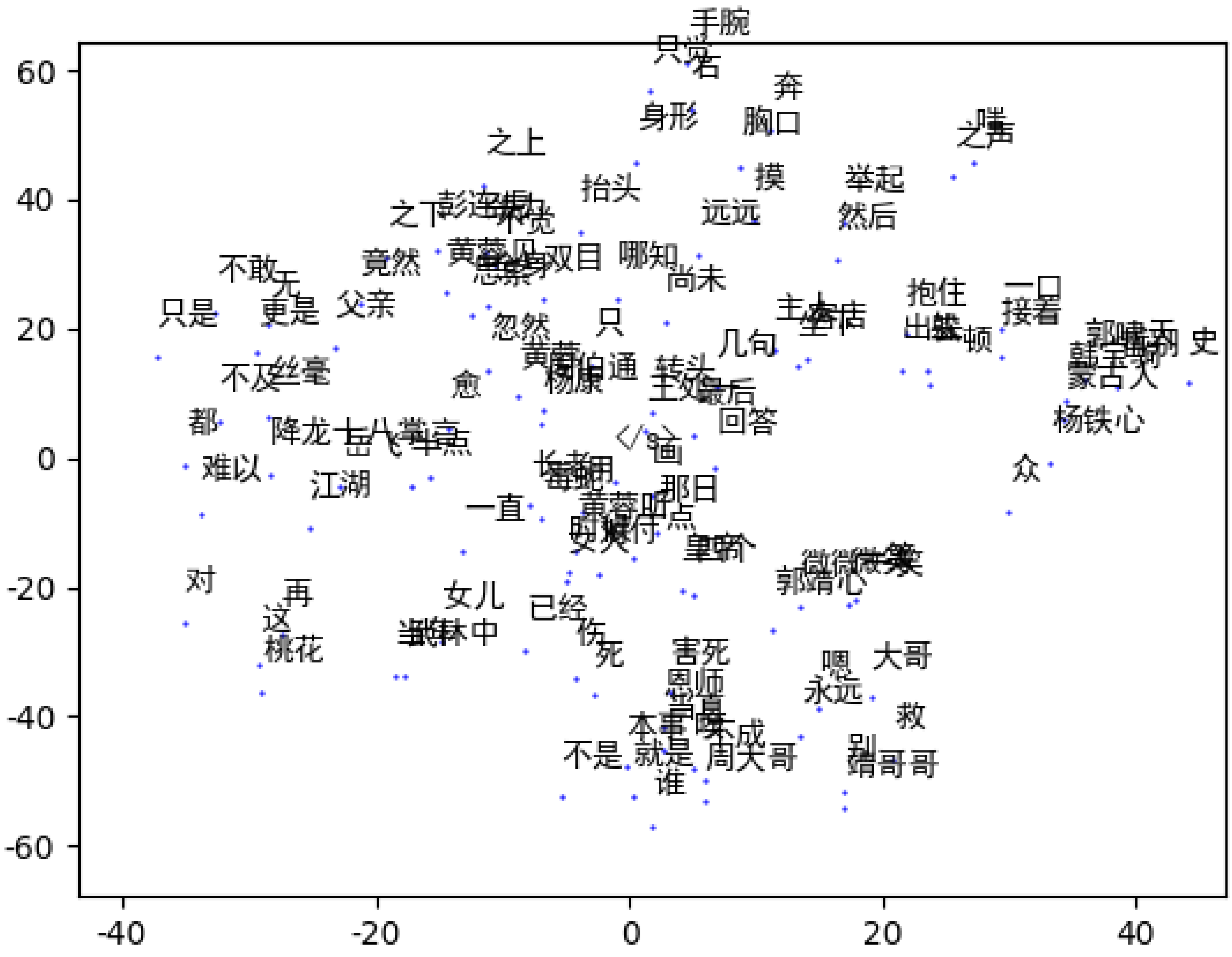


实验分析

- 从上图可以看出，由于词表数量过大，词向量在图片上显示过于稠密，并不能清晰地显示词之间的关系距离。
- 尝试通过修改 run.sh 中的参数，减少词表数量。如下图所示，将箭头所指的 10 修改为 50，代表是只保留词频大于 50 的词。



修改词频后


60
40
20
0
-20
-40
-60
-40
-20
0
20
40
手腕
奔
身形
胸口
之上
摸
举起
抬头
远远
然后
之下
竟然
双目
哪知
尚未
不敢
只是
更是
父亲
忽然
只
抱住
一口
接着
几句
史
蒙古人
杨铁心
不及
丝毫
愈
杨康
转头
最后
回答
都
降龙
难以
江湖
那日
众
一直
郭靖心
对
再
这
桃花
女儿
已经
伤
死
害死
恩师
大哥
永远
救
不是
就是
周大哥
谁

实验二：生成人名的词向量

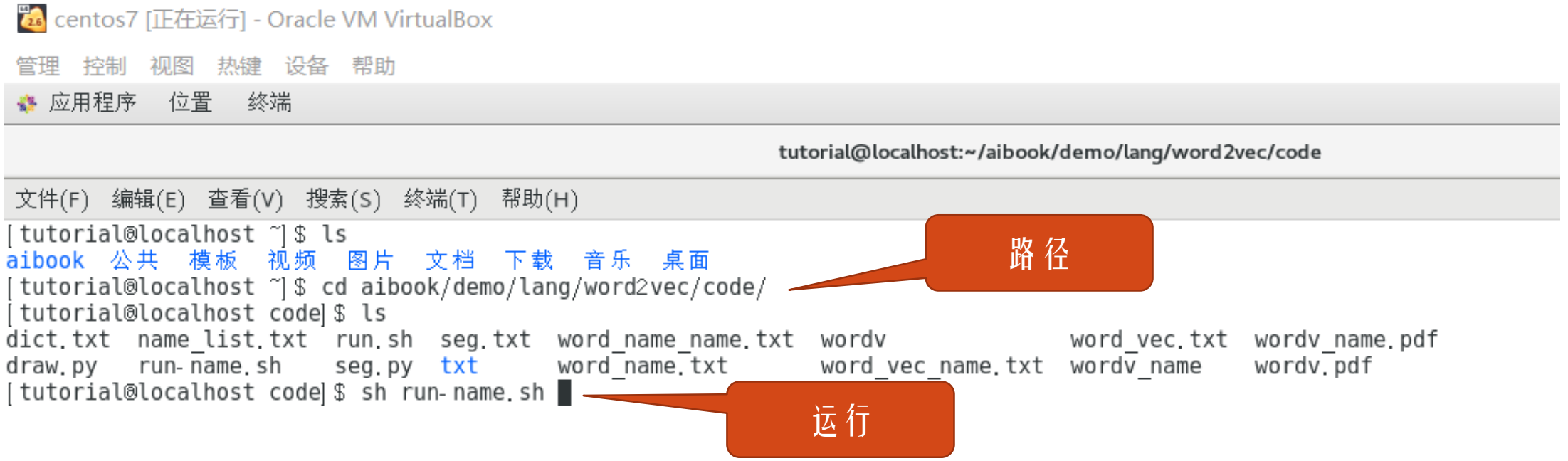
- 绘制《射雕英雄传》中人名的词向量空间，观察人物之间的关系

- 对词表 dict.txt 进行筛选，保留人名存至 name_list.txt 中
- 在词向量文件 wordv 中筛选出对应人名的词向量
- 利用 t-SNE 工具，将人名词向量投影到二维空间

实验步骤

- 本次实验我们单独选出《射雕英雄传》中部分人物，保存到 name_list.txt
- 运行 run-name.sh，查看他们之间的词向量的关系距离。

sh run-name.sh



实验结果

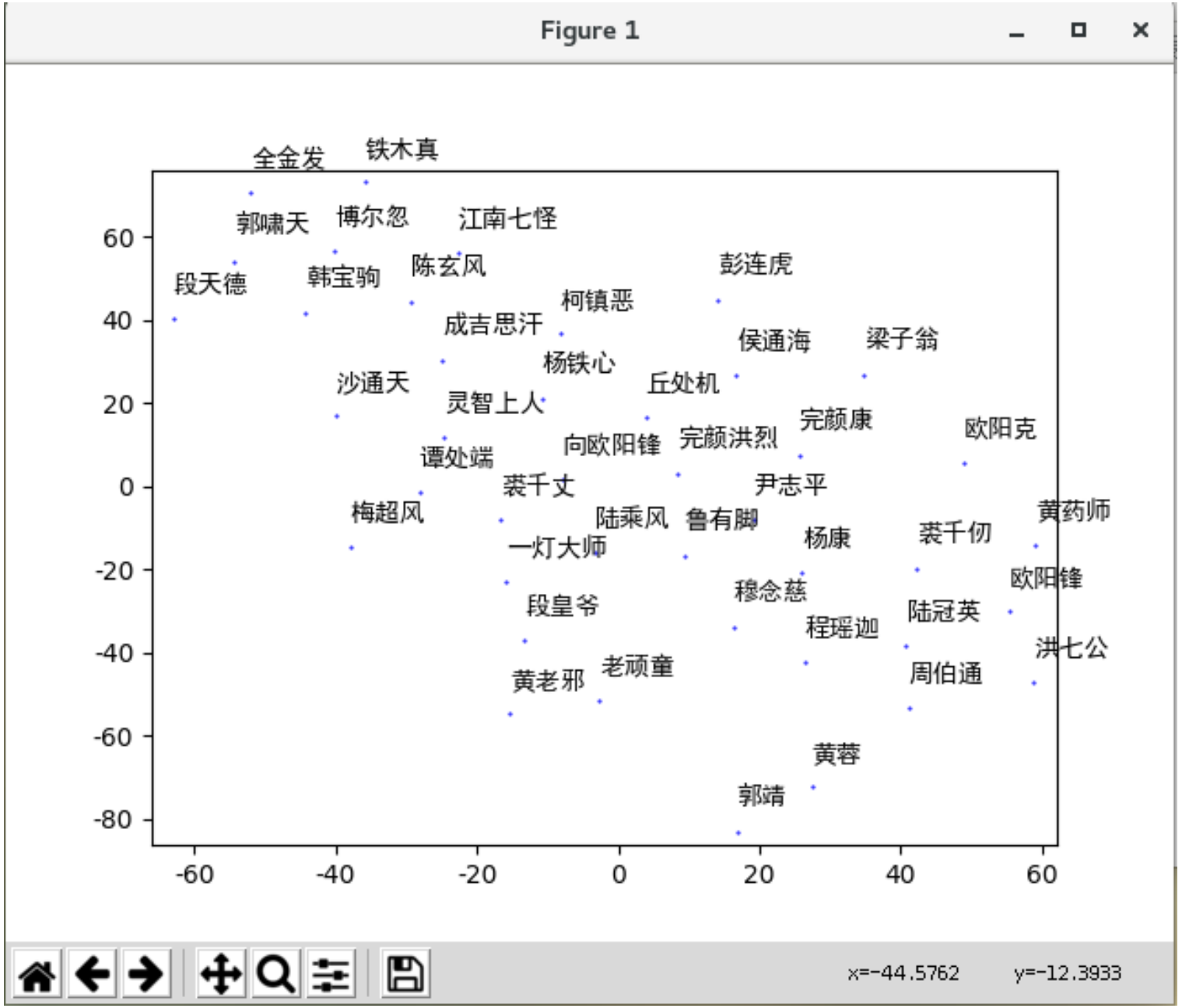
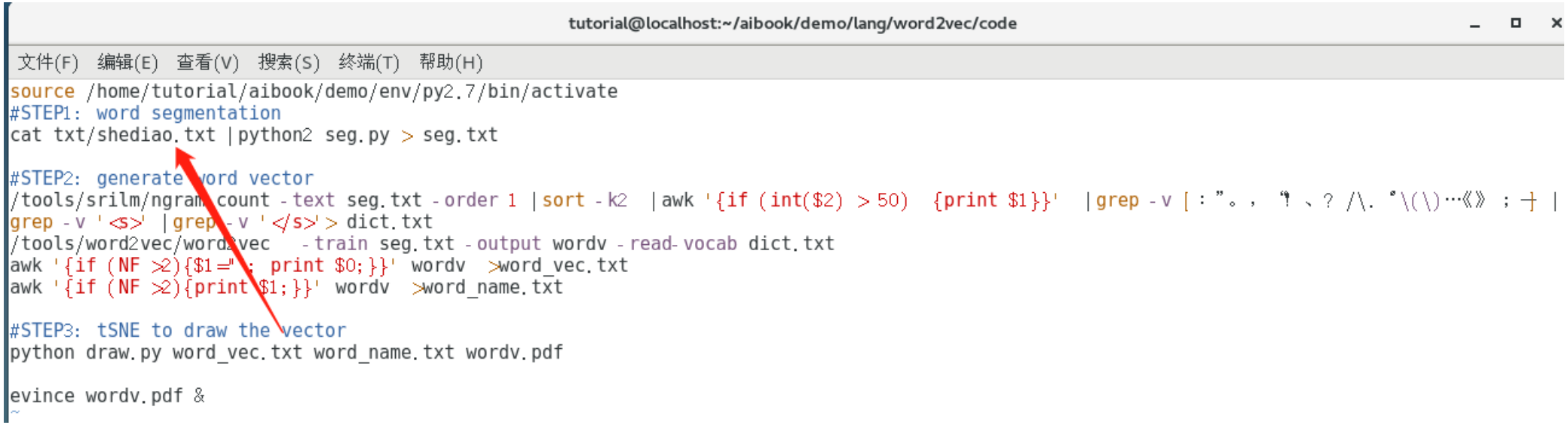


Figure 1
全金发
铁木真
郭啸天
博尔忽
江南七怪
段天德
韩宝驹
陈玄风
彭连虎
柯镇恶
成吉思汗
侯通海
梁子翁
杨铁心
沙通天
丘处机
灵智上人
完颜康
欧阳克
向欧阳锋
完颜洪烈
谭处端
裘千丈
尹志平
梅超风
陆乘风
鲁有脚
黄药师
杨康
裘千仞
一灯大师
欧阳锋
穆念慈
段皇爷
陆冠英
程瑶迦
洪七公
老顽童
周伯通
黄老邪
黄蓉
郭靖
60
40
20
0
-20
-40
-60
-80
-60
-40
-20
0
20
40
60
x=-44.5762
y=-12.3933

实验三：生成自己的词向量

- 找到一部小说的电子文本，命名为 xiaoshuo.txt
- 修改 code 下 run.sh（图形界面下双击或 vim）的文件名（shediao.txt）为 xiaoshuo.txt（下图所示）
- 运行 run.sh，查看小说中词向量的关系



The end