****

**实**

**验**

**报**

**告**

**课 程 大数据分析**

**题 目 项目二**

**分 数**

**学院名称 智能与计算学部**

**专 业 软件工程**

**学生姓名 黄冬梅**

**学 号 3016218170**

**年 级 2016级**

**班 级 5班**

目录

[一、实验目的 1](#_Toc527352380)

[二、实验内容 1](#_Toc527352381)

[三、实验步骤 1](#_Toc527352382)

[四、实验结果 1](#_Toc527352383)

[五、实验结论 1](#_Toc527352384)

[六、源代码 1](#_Toc527352385)

## 实验目的

针对第十九届中央委员的统计数据进行分析，展示说明这204名委员(或部分)的属性特征，或依据委员个人履历分析其调任升迁的生涯轨迹，或展示委员在职场上是否曾有共事等信息。

## 实验内容

任务1

* 1. 基本信息统计。请按照基本信息如性别、年龄、籍贯等方面对数据进行统计分析。统计角度应选取合理，具有一定的新意。
  2. 履历主题分析。使用LDA主题模型，寻找文本中所蕴含的主题。请注意选择合理的分词，停词表等预处理技术。展示每个人员在不同主题的分布，以及每个主题的在关键词上的分布。

任务2

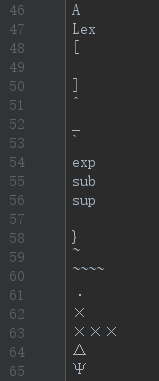
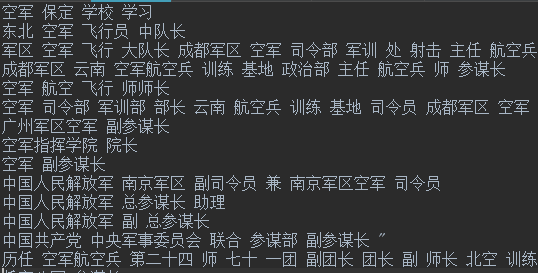
* 1. 提取履历中的地点和时间信息。抽取不同人员的履历轨迹。
  2. 分析不同人员之间的关系（可以思考的角度包括基本信息、履历经历等）。
  3. 任何其他的模式。

## 实验步骤

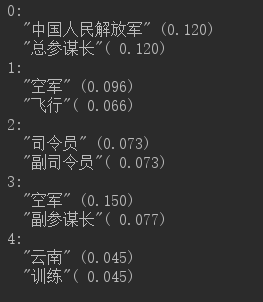
1. 使用Java处理excel表格进行基本的信息统计。分别统计性别、民族、籍贯。籍贯部分使用Java程序进行统计和比对，代码中使用依次比对籍贯中的前两个字，得到每个省份的委员数。详细代码见附件，所得到的结果如下：



1. 将履历以文本形式写入txt文件，使用停词表1893（UTF-8）.txt对文本进行处理，之后使用jieba库对文本进行分词。停词表（部分）和分词结果如下：

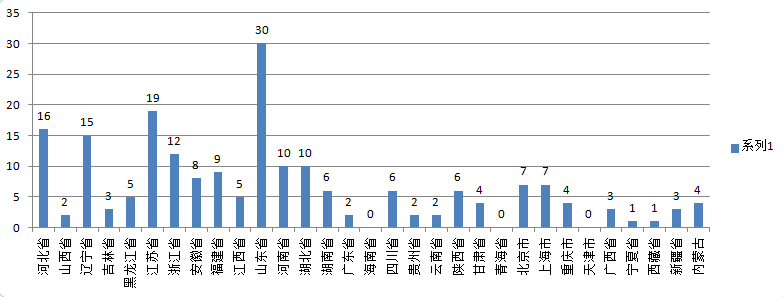
1. 将分词结果映射成字典
2. 使用LDA进行主题分析，对其中某位委员的履历主题分析结果如下：

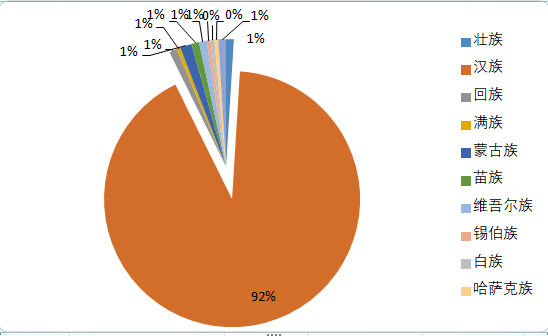


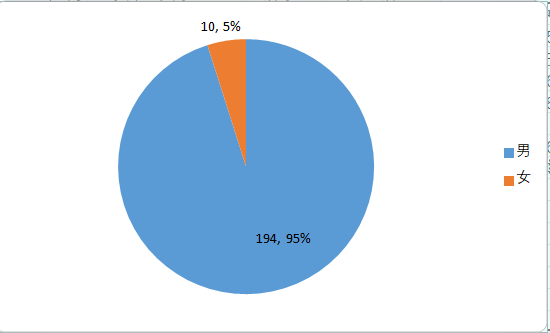
1. 履历信息中按照年份拆分，读取地名，绘制成个人的升迁路线。

## 实验结果

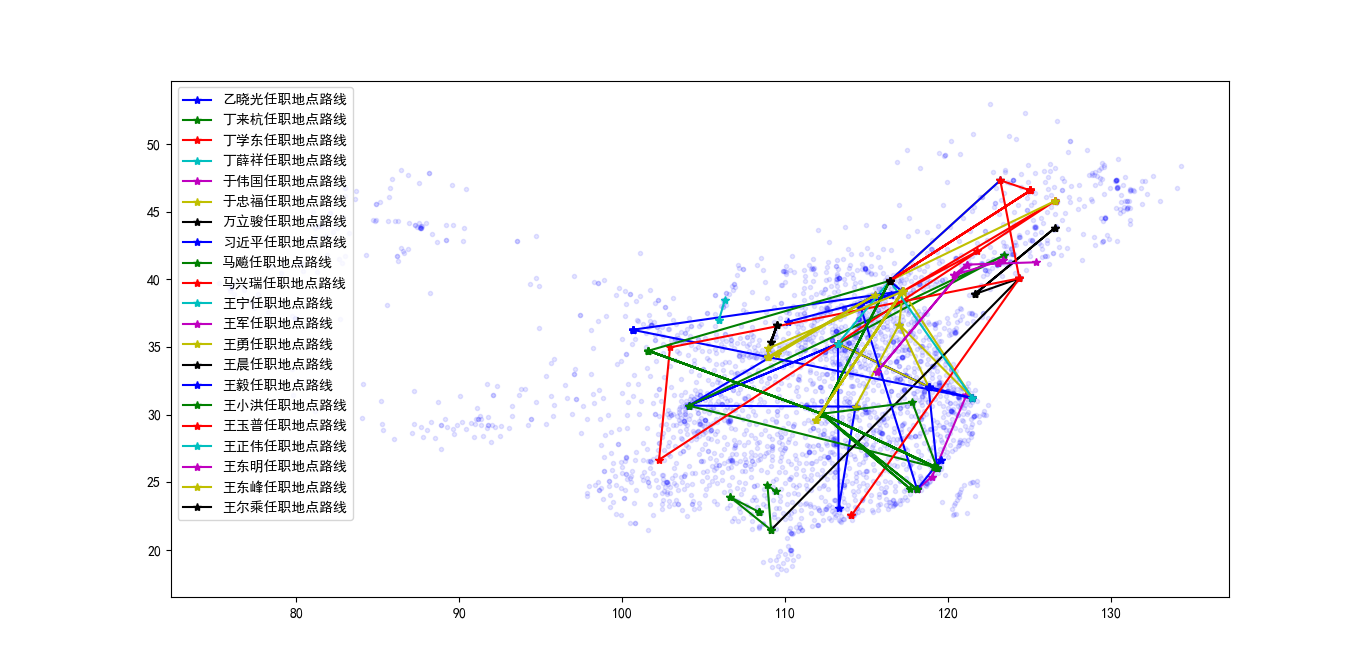
1. 基本信息统计结果如下：







1. 个人的履历轨迹如下：



## 实验结论

单就基本信息统计而言，基本以男性、汉族居多，就籍贯而言，来自山东、江苏的委员较多，可能与山东本身就是人口大省有关。就个人履历而言，基本每一位委员都经历过几个地方，这可能与避免干部培养当地人际关系网有关。

## 源代码

源代码见附件。