**2022年秋数据挖掘大作业报告**

——55200129 徐明润

1. 实验内容

党代会报告分析。

1、收集二十大，十九大和十八大的报告，分别统计三份报告中的所使用词语的出现的次数除以所有词语出现的总数，得到词语的词频，再将三份报告合并为一份文档进行词频统计。

2、将每个词语当作一个项，每句话当作一条交易记录，求出每份报告中提升度最高的20条关联规则。

1. 实验过程分析

* 词频统计

1.在main.py中，导入pdfplumber包，读取pdf形式的报告，将报告中的内容保存到一个字符串中。

2.导入jieba包将得到的字符串进行分词，将分词结果保存到一个列表中

3.实现使用.txt文档保存停用词，即诸如“的”，“你”，“假如”和标点符号不需要统计的词和符号保存在stop.txt文档中，将该文档中内容读取成一个列表

4.遍历分词得到的列表，如果单词出现在停用词列表中或者长度为1则不进行统计，否则添加到词语词典中，并且将其出现次数加1

5.根据词语出现次数进行排序，将排序后的结果前10个词语打印，并将对应的出现次数除以总词语出现次数，打印。使用xlwt包将结果保存到.xls文件中

* 挖掘关联规则

1.读取pdf内容过程同词频统计第一步。

2.导入jieba包，使用句号将字符串进行拆分，并将结果保存到表示交易记录的列表中

3.加载停用词

4.遍历每一条记录，将句子拆分一个个词语的项，保存到交易记录中

5.导入efficient\_apriori模块的apriori部分，将交易记录列表作为apriori函数的第一个参数传入，指定最小支持度和最小置信度，接收返回的频繁项集和关联规则

6.根据关联规则的提升度进行排序，打印前20条关联规则，并将结果保存到.xls文件中

1. 实验结果

* 词频分析结果：
  + 二十大词频统计结果
    - 
  + 十九大词频统计结果
    - 
  + 十八大词频统计结果
    - 
* 关联规则统计结果：
  + 二十大关联规则统计结果
    - 
  + 十九大关联规则统计结果
    - 
  + 十八大关联规则统计结果

1. 结果分析
2. 总结