**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. **实验过程及方法**

* *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里。

3.实现使用.txt文档保存停用词，即诸如“的”，“你”，“假如”和标点符号等不需要统计的词和符号保存在stop.txt文档中，并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，否则添加到词语字典中，并且将其出现次数加1。

5.将4中的词语字典再次转化为列表，然后根据词语的出现次数进行排序，将排序后的列表的前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，将句子拆分包含一个个词语的项集，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，接收作为返回值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*
    - 
  + *三份报告合并词频统计结果：*
  + 
* 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(模式评估)**

通过对三份报告以及合并后的报告进行词频分析，可以发现三份报告出现频率最高的词都是“发展”，三份报告中词频前三的词都包含建设，结合当下中国的发展战略以及总书记多次讲话的主要思想，可以了解到时下中国强调的主题即为发展，发展生产力，发展社会主义制度，发展人类命运共同体，发展综合国力等，为达到两个百年目标中国的主要任务就是不断地进行发展和建设。从十八大到二十大的报告以及词频统计结果来看，除了主题是发展和建设之外，关键词汇还包含了“中国”，“特色”两词，说明当下在发展的同时，仍需要坚持中国特色社会主义道路，坚持社会主义核心价值观的建设，坚持建设我国的社会主义制度，时刻提防西方国家进行的和平演变，保持我国社会主义制度的优越性，首要就是建设社会主义文化，强调社会主义的进步性和发展的特点，同时，对于需要创新的部分，需要敢于创新，积极体现出我国社会主义中的“特色”性。

通过对三份报告关联规则的挖掘发现，人大报告的重点由十八大强调的发展，特色，科学依存于发展观到十九大二十大共同强调的复兴和中国民族的关系，说明我国的社会主义建设取得了长足的进步，为中华民族的复兴准备了充足的条件，也说明了中华民族的复兴是一个更为深刻，意义更加远大的主题，需要付出更多的努力，发展的目的就是为了实现中国民族的复兴，明确指明了第二个百年我党和全国人民要实现的共同目标。通过复兴和中华民族关联规则提升度可以发现，提升度的值都远大于1，且置信度大于60%，说明十九大二十大多次强调了复兴和中国民族之间的不可分割，二者具有高度的相关性，二者是紧紧相连的关系，中华民族必然实现复兴，复兴需要中国民族全体人民共同奋斗。

此外，三份报告中提升度最高的二十条规则中名列前茅的规则还有('中国', '时代') --> ('特色')，说明十余年来我党和我国的发展离不开中国特色社会主义的建设，该关联规则的提升度大于1，与此同时置信度也大于50%，二者的关联效果较好，说明当前时代中国发展依赖于中国特色社会主义。

除上述分析，三份报告的关联规则挖掘结果中都包含了制度和完善、现代化和建设、发展特色和科学中国等项集之间的关联规则，且这些关联规则的提升度都大于1，说明三次人大会议在指出我国发展目标和主题的同时也指明了实现这些目标的方法论，制度需要不断完善，以保证在优越的社会制度下飞速发展，建设现代化国家，以实现国民生产力质的飞跃。

结合词频分析和关联规则挖掘来看，三次人大共同强调了当下中国仍需要坚持发展，以现代化的手段发展，以中国特色社会主义制度为保障发展，为实现中国民族复兴而发展，通过阅读三份报告，可以发现数据挖掘的结果和报告主旨大意基本相符，挖掘结果较为准确。

1. **总结**

通过此次大作业，我充分运用了在课堂中学习的知识，掌握了在大数据之中提取到有效模式的方法，熟悉了Python语言的使用和各种强大的库函数的调用，自己的编程能力有所提升。此外。作业使用的材料更是令人兴奋不已，通过此次作业我得以仔细揣摩三次人大会议的报告，对人大报告的数据分析，我充分感受到我国历史的车轮在轰隆前滚，国家正在势不可挡地走在复兴的道路上，这些都更加深沉了我的爱国情怀，同个人能力的提升相比，我认为这才是这次作业的真正意义所在。