南京農業大學

NANJING AGRICULTURAL UNIVERSITY

数字人文下的文本聚类

目录

CONTENT

- 文本聚类基本知识
- 非物质文化遗产的自动聚类

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

文本聚类基本知识

1、文本聚类

◆聚类又称群分析，是数据挖掘的一种重要的思想。文本分类一般和聚类一起探讨，其区数量，其根据文本内部的特征以及类别数量自动将待分类文本聚合成若干个类别。别在于有没有预先划分好的类别。文本分类任务需要预先定义好具体类别，再接着将待分类文本划分入这些类别中；而文本聚类任务则不需要预先定义具体类别，只需提前规定好划分的类别

2、特征提取

◆所提取的特征在一定程度上决定了整个分类的性能。如何从非结构化、半结构化文本中提取相应特征并转化为结构化信息是特征提取的关键。目前常用的特征提取模型如下：One-Hot、BOW词袋模型、连续词袋模型（CBOW）、Skip-Gram模型和Word2vec模型。

3、K-means

◆在整个聚类算法体系中，K-means算法是典型的基于距离的一种算法。距离是该算法相似性计算和评价的基础，如果两个文本对象在计算过程中距离较近则相似度的值较大。在具体的实现聚类的过程中，由距离靠近的文本对象所组成的簇是判定聚类结果的关键。

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

非物质文化遗产的自动聚类

语料准备

一个非遗项目字段样例如下所示，分别抽取出项目名称和详细介绍作为本章聚类的一个非遗项目。一个项目代表一条数据，如下表所示，共有2690条数据。

项目名称	苗族古歌	项目序号	1
项目编号	Ⅰ-1	公布时间	2006（第一批）
类别	民间文学	所属地区	贵州省
类型	新增项目	申报地区或单位	贵州省黄平县
详细信息	申报地区或单位：贵州省黄平县，苗族分布在我国西南数省区。按方言划分，大致可分为湘西方言区、黔东方言区、川滇黔方言区。黔东南清水江流域一带是全国苗族最大的聚居区，大致包括凯里、剑河、黄平、台江、雷山、丹寨、施秉、黄平、镇远、三穗，以及广西三江和湖南靖县等地。在此广大苗族聚居区普遍流传着一种以创世为主体内容的诗体神话，俗称“古歌”或“古歌古词”…（文本内容较长，不予全文显示）		
相关继承人信息	编号： 01-0004，姓名： 王明芝，性别： 女，出生日期： 1939.06，民族： 类别： 民间文学，项目编号： Ⅰ-1，项目名称： 苗族古歌，申报地区或单位： 贵州省黄平县，编号： 01-0003，姓名： 龙通珍，性别： 女，出生日期： 1936.04，民族： 类别： 民间文学，项目编号： Ⅰ-1，项目名称： 苗族古歌，申报地区或单位： 贵州省黄平县		
相关项目信息	Ⅰ-1，民间文学，湖南省花垣县，Ⅰ-1，民间文学，贵州省台江县		

数据预处理

◆经过数据清洗后，对文本进行分词。分词工具采用jieba分词，停用词表为哈工大停用词表。

一条数据的分词结果如下：

苗族古歌 申报 地区 单位 贵州省 黄平县　苗族 分布 我国 西南 数 省区 方言 划分 分为 湘西 方言 区 黔 东方 言区 川滇黔 方言 区 黔东南 清水江 流域 一带 全国 苗族 聚居区 包括 凯里 剑河 黄平 台江 雷山 丹寨 施秉 黄平 镇远 三穗 广西 三江 湖南靖县 苗族 聚居区 流传 一种 创世 主体 内容 诗体 神话 俗称 古歌 古歌 古词　苗族古歌 内容 包罗万象 宇宙 诞生 人类 物种 起源 开天辟地 初民 时期 滔天 洪水 苗族 迁徙 苗族 古代 社会制度 日常 生产 生活 无所不包 苗族 古代 神话 总汇　苗族古歌 古词 神话 鼓社 祭 婚丧 活动 亲友 聚会 节日 场合 演唱 演唱者 多为 中老年人 巫师 歌手 酒席 演唱 古歌 场合 苗族古歌 古词 神话 民族 心灵 记忆 苗族 古代 社会 百科全书 经典 史学 民族学 哲学 人类学 多方面 价值 古歌 古词 神话 民间 流传 唱诵　文化 市场经济 冲击 苗族古歌 濒临 失传 台江 为例 全县 13 万 苗族 同胞中 唱 完整 部 古歌 寥寥无几 二百余 人能 唱 完整 古歌 中老年人 传承 古歌 老人 年事已高 抓紧 抢救 保护 苗族古歌 这一 民族 瑰宝 最终 世间 消失

文本表示

◆将词映射到向量空间中，以下选取word2vec方法，使用的工具是Gensim，语言模型采用CBOW，训练方法采用Negative Sampling，最小词频min_count设置为0，上下文最大距离window设置为5，维度size设置了100维，可根据情况改变维度，进行不同维度的对比。其余参数为默认。最后使用均值文本表示法表示文本向量，即将该文档内所含有词的对应词向量相加求平均。

K-means聚类

◆实现K-means算法的方法有很多，本次使用的工具是sklearn中k-means++算法。关于k值范围确定，考虑到非遗分类法主要有六类法、八类法、十类法、十三类法和十六类，其中十三类法和十六类法是更为细致的划分，但是目前基于层级划分的类目都不会过多，为了更加全面对比各分类法中类别划分的科学性，同时考虑计算指标时图表的连续性和可预测性，本次K-means聚类k值范围确定为2到17，其余参数为默认。

◆轮廓系数是评价聚类效果的常用指标，将簇中所有非遗样本的轮廓系数求平均值就得到非遗平均轮廓系数，取值范围为[-1,1]。程序实现了轮廓系数的计算，并使用python画图工具实现每次聚类效果的轮廓系数可视化。如要人工评估聚类效果，可直接输出estimator.labels_观察。

K-means聚类

◆轮廓系数是评价聚类效果的常用指标，将簇中所有非遗样本的轮廓系数求平均值就得到非遗平均轮廓系数，取值范围为[-1,1]。程序实现了轮廓系数的计算，并使用python画图工具实现每次聚类效果的轮廓系数可视化。如要人工评估聚类效果，可直接输出estimator.labels_观察。