**云南大学软件学院**

**实 验 报 告**

成员1： 李岳哲 学号： 20171120301 专业： 软件工程 日期： 2019/12/25 成绩：

成员2：王子陈 学号：20171050008专业：电子科学与技术 任课教师：谢仲文 成绩：

**数据挖掘技术 实验七（本实验以2人为单位完成，也可独立完成）**

1. **实验目的**
2. 掌握数据预处理的方法。
3. 掌握基本的数据统计方法。
4. **实验内容**
5. 在一个爬虫爬取的数据集（新疆的餐馆信息，数据集另给）上进行数据预处理和基本的统计分析。
6. 基本要求1：根据URL等信息，想办法定位各个餐馆所属的地区。（新疆包含以下行政区：乌鲁木齐市、克拉玛依市、吐鲁番市、哈密市、阿克苏地区、喀什地区、和田地区、塔城地区、阿勒泰地区、昌吉回族自治州、博尔塔拉蒙古自治州、巴音郭楞蒙古自治州、克孜勒苏柯尔克孜自治州、伊犁哈萨克自治州）

3. 基本要求2：根据给定数据集，做出有意义的统计分析。（必做）

4．提高要求：根据已经学习的数据挖掘知识，做出创新性结果。（选做）

**三、****实验要求**

1. 完成实验内容，源码作为实验报告附件一起打为一个压缩包提供。该压缩包要包含实验报告、代码文件。
2. 关键部分要求有注释，注释量不低于20%
3. 要求独立完成，不得抄袭代码。

**四、关键实验步骤（请粘贴关键步骤、代码、实验结果）**

1. 提取出URL中的县名，并统计各县的餐馆数量

import jieba

import re

import pandas as pd

filename='C:/Users/MY/Desktop/数据挖掘实验/实验7：数据预处理/新疆餐馆.xlsx'

dataSet=pd.read\_excel(filename)

urlSet=dataSet['url']

locatDict={}

for row in urlSet:

try:

piece=re.findall(r"food/(.+?)/",row)

location=re.sub('\d+',"",piece[0])

except IndexError:

print("问题在于:",row)

continue

if location not in locatDict.keys():

locatDict[location]=0

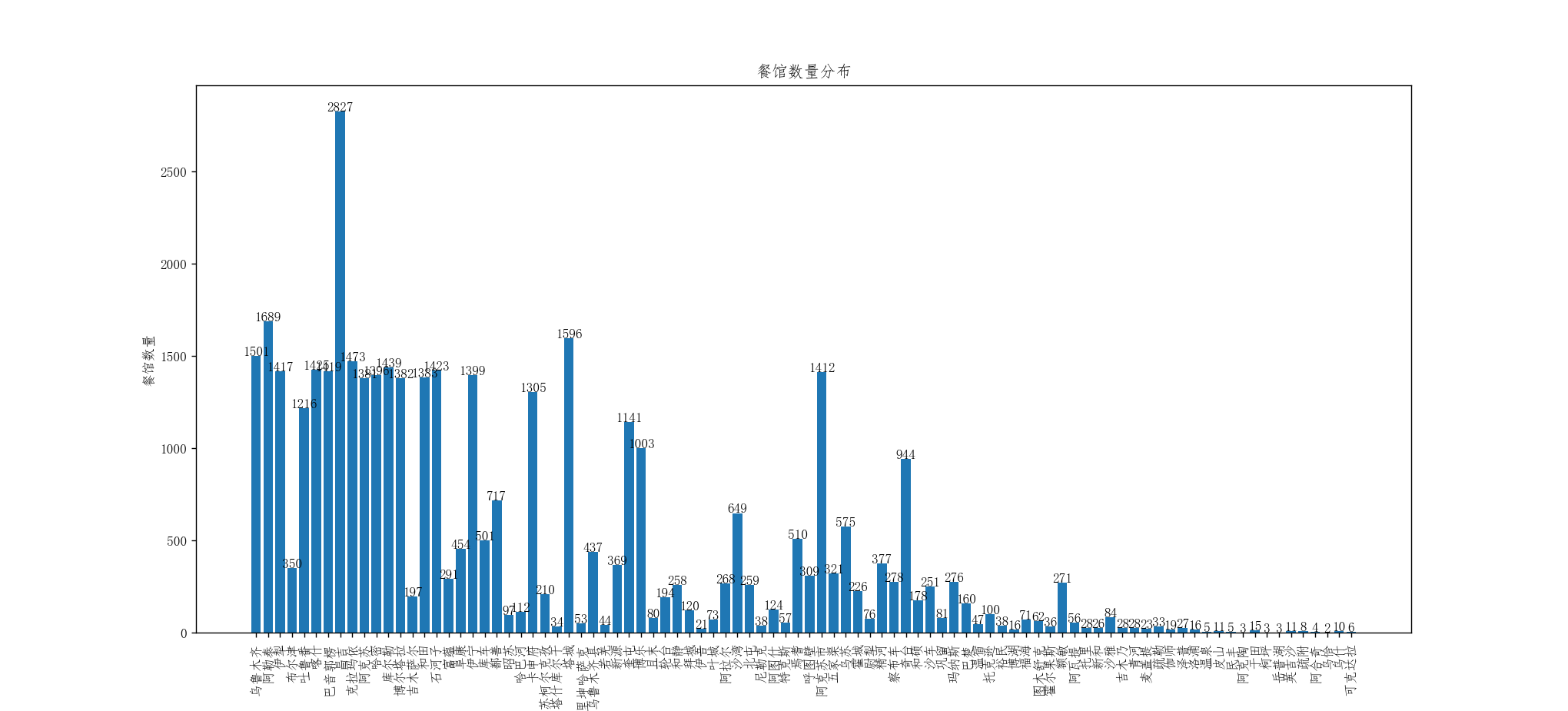
locatDict[location]+=1

1. 把英文地名翻译成中文地名

用字典存储中英文对应关系

parishDict={'urumqi': "乌鲁木齐", 'aletai': "阿勒泰", 'yili': "伊犁", 'buerjin': "布尔津", 'tulufan': "吐鲁番", 'kashi': "喀什",'bayinguoleng': "巴音郭楞", 'changji':"昌吉", "kelamayi": "克拉玛依", "akesu": "阿克苏", 'hami': "哈密", 'kuerle': "库尔勒", 'boertala': "博尔塔拉", 'jimusaer': "吉木萨尔", 'hetian': "和田", 'shihezi': "石河子", 'fuwen': "富蕴", 'fukang': "阜康", 'yining': "伊宁", 'kuche': "库车", 'shanshan': "鄯善", 'zhaosu': "昭苏", 'habahe': "哈巴河", 'kashenfu': "卡申府", 'kezilesukeerkezi': "克孜勒苏柯尔克孜", 'tashikuergan': "塔什库尔干", 'tacheng': "塔城", 'balikunhasake': "巴里坤哈萨克", 'wulumujixian': "乌鲁木齐县", 'ruoqiang': "若羌", 'xinyuan': "新源", 'kuitun': "奎屯", 'bole': "博乐", 'qiemo': "且末", 'luntai': "轮台", 'hejing': "和静", 'baicheng': "拜城", 'yiwu': "伊吾", 'yecheng': "叶城", 'alaer': "阿拉尔", 'shawan': "沙湾", 'beitun': "北屯", 'nileke': "尼勒克", 'atushi': "阿图什", 'tekesi': "特克斯", 'yanqi': "焉耆", 'hetubi': "呼图壁", 'akesushi': "阿克苏市", 'wujiaju': "五家渠", 'wusu': "乌苏", 'huocheng': "霍城", 'yuli': "尉犁", 'jinghe': "精河", 'chabuchaer': "察布车", 'qitai': "奇台", 'heshuo': "和硕", 'shache': "沙车", 'gongliu': "巩留", 'manasi': "玛纳斯", 'bachu': "巴楚", 'wensu': "温宿", 'tuokexun': "托克逊", 'yumin': "裕民", 'bohu': "博湖", 'fuhai': "福海", 'tumushuke': "图木舒克", 'huoerguosi':'霍尔果斯', 'emin': "额敏", 'awati': "阿瓦提", 'tuoli': "托里", 'xinhe': "新和", 'shaya': "沙雅", 'jimunai': "吉木乃", 'qinghe': "青河", 'maigaiti': "麦盖提", 'shule': "疏勒", 'jiashi': "伽师", 'zepu': "泽普", 'luopu': "洛浦", 'wenquan': "温泉", 'pishan': "皮山", 'minfeng': "民丰", 'aketao': "阿克陶", 'yutian': "于田", 'keping': "柯坪", 'yuepuhu': "岳普湖", 'yingjisha': '英吉沙', 'shufu': "疏附",'aheqi': "阿合奇", "wuqia":"乌恰", 'wushi': "乌什", 'kokdala': "可克达拉"}

就可得到



1. 把县定位到市级行政区

先准备一个市-县对应的字典：

zhBelongDict={"乌鲁木齐市":("乌鲁木齐","水磨沟区","天山区","沙依巴克区","新市区","头屯河区","达坂城区","米东区","乌鲁木齐县"),

"克拉玛依市":"克拉玛依、独山子区、白碱滩区、乌尔禾区",

"吐鲁番市":"高昌区、鄯善、托克逊",

"哈密市":"伊州区、伊吾、巴里坤哈萨克",

"阿克苏地区":"阿克苏、温宿、库车、沙雅、新和、拜城、乌什、阿瓦提、柯坪",

"喀什地区":"喀什、疏附、疏勒、英吉沙、泽普、莎车、叶城、麦盖提、岳普湖、伽师、巴楚、塔什库尔干塔吉",

"和田地区":"和田、墨玉、皮山、洛浦、策勒、于田、民丰",

"昌吉回族自治州":"昌吉、阜康、呼图壁、玛纳斯、奇台、吉木萨尔、木垒哈萨克" ,

"博尔塔拉蒙古自治州":"博乐、阿拉山口、精河、温泉",

"巴音郭楞蒙古自治州":"库尔勒、焉耆、轮台、尉犁、若羌、且末、和静、和硕、博湖",

"克孜勒苏柯尔克孜自治州":"阿图什、阿克陶、阿合奇、乌恰" ,

"伊犁哈萨克自治州":"伊宁、奎屯、霍尔果斯、尼勒克、霍城、巩留、新源、昭苏、特克斯、察布查尔锡伯自治县",

"塔城地区":"塔城、乌苏、额敏、沙湾、托里、裕民、和布克赛尔蒙古自治县",

"阿勒泰地区":"阿勒泰、布尔津、富蕴、福海、哈巴河、青河、吉木乃" }

#构造一个列表，存放每一条记录所属的市，再把这一列添加到dataframe中

parishList=[] #存放县名

distrList=[] #存放市名

for rowIdx in range(len(dataSet)):

try:

row=dataSet.loc[rowIdx,:]#把每一行取出来,返回值类型是Series,引用时用方括号

piece=re.findall(r"food/(.+?)/",row[0])#正则表达式，取出URL的文件名部分

location=re.sub('\d+',"",piece[0])#去掉数字，只要字符串

city=parishDict[location] #找出英文名location对应的中文名city

parishList.append(city) #顺便把县名存到列表里

for key in zhBelongDict.keys(): #在字典中找对应的市

if city in zhBelongDict[key]: #检查此地名是否在字典的value里存在

district=key

break #找到就停止循环

distrList.append(district) #把市名存放到列表里

except IndexError: #有的URL是残缺的，不能处理

print("问题在于:",row[0])

parishList.append("")

distrList.append("") #就在列表里存为空字符串

continue

#把parishList作为一列，添加到dataSet

dataSet["县"]=parishList

dataSet["市"]=distrList

dataSet.head()

其中存在残缺的URL，跳过，并未统计

得到效果如下：



1. 用jieba统计词频

canteenStr=list(dataSet["餐馆名"])

words=jieba.lcut(str(canteenStr),cut\_all=True)

counts={}

for word in canteenStr:

if(type(word)==float):

continue

elif len(word)==1: #排除单个字符串的分词结果

continue

else:

counts[word]=counts.get(word,0)+1

items=list(counts.items())

items.sort(key=lambda x:x[1],reverse=True)

for i in range(30): #输出排序前30的单词

word,count=items[i]

print("{0:<10}{1:>5}".format(word,count))

吴佳米粉 35

啊臻味道米粉 32

四十九丸子汤 26

疯狂烤翅 25

黄焖鸡米饭 24

儿子娃娃椒麻鸡 22

杭州小笼包 21

海楼抓饭 21

乌孜古凉皮 20

润口肥牛 19

张大师鸭爪爪特色火锅 19

小郡肝串串香 18

名客佳大鸡排 17

贝克汉堡 16

N多寿司 16

麦多馅饼 16

柴窝堡辣子鸡 16

倒字鸡排 15

卓记米粉 15

顺水鱼馆 14

正新鸡排 13

绝味鸭脖 13

奉姐冒菜 13

香辣擀面皮 13

馋私猫卤猪蹄 12

重庆小面 12

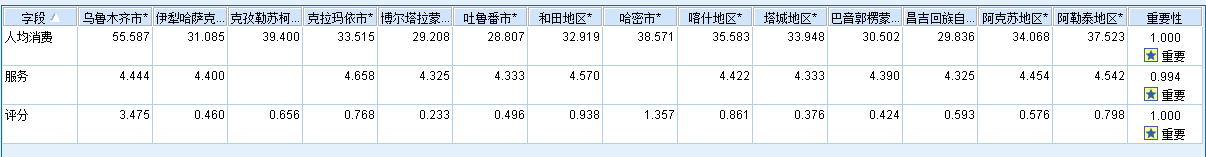
粮全其美手抓饼 12

粒粒香抓饭 12

百富烤霸 12

东北饺子馆 12

1. 各行政区的人均消费、服务的均值

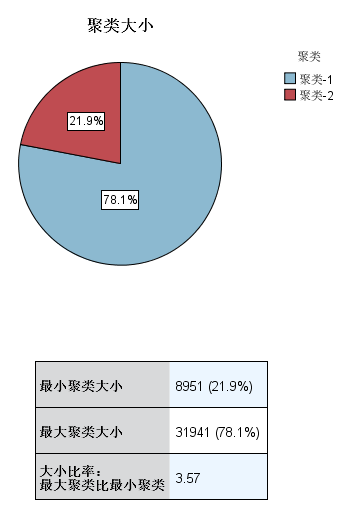
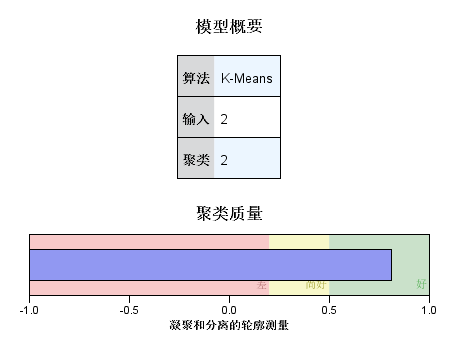


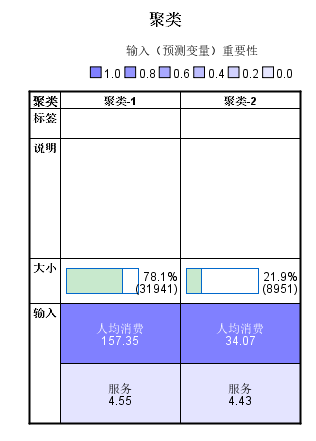
由上图的人均消费均值可看出，乌鲁木齐市作为新疆的首府，整体平均消费明显高于其它行政区，其中博尔塔拉蒙古自治州、巴音郭楞蒙古自治州和吐鲁番市三个行政区的人均消费不足三十，人均消费均值略高于乌鲁木齐市的人均消费均值的一半。由此可得出入驻美团的各种餐馆在乌鲁木齐市的发展水平是远高于其它行政区的，而吐鲁番市的发展水平最低。

由上图的服务评分的均值可看出，餐馆整体服务最好的是克拉玛依市，最差的是博尔塔拉蒙古自治州和昌吉回族自治州，而消费水平最高的乌鲁木齐市的服务均值只是处于中游，不是很高，也不是很低，可见乌鲁木齐市的餐馆还需提升服务水平才能配得上它的消费水平。虽然克拉玛依市的餐馆服务评分最高，但人均消费水平偏低，但拥有好的服务质量是餐馆发展的重中之重，可以认为克拉玛依市的各种餐馆整体的发展潜力最高。位居消费水平最低前三位的博尔塔拉蒙古自治州和昌吉回族自治州的整体服务水平也是最低，可见这两个行政区的人均消费水平低与其服务质量水平低有密切关系，二者可能互相影响，而这两个行政区的餐馆想提升发展水平，应先从提高服务质量水平开始。

1. 将人均消费和服务两个指标使用k均值算法进行聚类

最佳k值为2，结果如下





从聚类结果中可看出人均消费与服务这两项指标间存在正比关系，也就是说提升服务质量，能够为餐馆提高人均消费的水平。