**中医病案文本挖掘**

**实验二 中医病案数据清洗**

## 一．数据筛选

**步骤：**

1. 读数据
2. 筛选出失眠的病例
3. 输出，存文件

**方法：**使用pandas和numpy中的fillna、apply、contains方法

**处理过程：**

import pandas as pd

import numpy as np

data = pd.read\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\source\中医病案数据.xlsx",sheet\_name="中医病案sample")

#（1）数据筛选：根据主诉与现病史、自诉和症状两个属性，保留与由失眠相关症状的病人信息。

data["主诉"] = data["主诉"].fillna("无")

data["主诉"] = data["主诉"].apply(str)

data = data[data["主诉"].str.contains("失眠")]

data.to\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\1.数据筛选\中医病案数据(失眠).xlsx",sheet\_name="中医病案sample",index=False)

**结果：**



## 二．缺失值处理

**步骤：**

1. 读数据
2. 筛选出【现病史、自诉、症状】中含有\*病|\*痛|\*炎|失眠的所有词
3. 将上述列合并
4. 保存，输出文件一
5. 读入文件一
6. 合并相同ID，不同主诉。
7. 输出文件

**方法：**使用pandas和numpy中的fillna、apply、agg、replace、split、groupby方法

**处理过程：**

import pandas as pd

import numpy as np

import re

df = pd.read\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\source\中医病案数据.xlsx",sheet\_name="中医病案sample")

data = df['现病史、自诉、症状'].astype(str)

#筛选出【现病史、自诉、症状】中含有\*病|\*痛|\*炎|失眠的所有词

pattern1 = re.compile("[。，]([\u4e00-\u9fa5]{1,5}病){1}")

pattern2 = re.compile("[。，]([\u4e00-\u9fa5]{1,5}痛)")

pattern3 = re.compile("[。，]([\u4e00-\u9fa5]{1,5}炎)")

pattern4 = re.compile("(失眠)")

ls = pd.DataFrame()

ls[0] = data

ls[1] = df['主诉']

ls[2] = data.apply(lambda x:re.findall(pattern1,x))

ls[3] = data.apply(lambda x:re.findall(pattern2,x))

ls[4] = data.apply(lambda x:re.findall(pattern3,x))

ls[5] = data.apply(lambda x:re.findall(pattern4,x))

#合并列存放到列1中

for i in range(0,len(df['主诉'])):

#print(df['主诉'][i],i)

if(df['主诉'][i] is np.nan):

str1 = (str(ls[2][i])+str(ls[3][i])+str(ls[4][i])+str(ls[5][i]))

str2 = re.sub("['\[\]]","",str1)

ls[1][i] = str(ls[1][i]).replace('nan','')+str2

df['主诉'] = ls[1]

df.to\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\2.缺失值处理\中医病案数据(填补).xlsx",sheet\_name="中医病案sample",index=False)

import pandas as pd

import numpy as np

import re

#在现病史、自诉和症状列中，有些复诊病人的记录是病史如前又补充了一些内容，需把复诊病人的该列补充完整。

df = pd.read\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\2.缺失值处理\中医病案数据(填补).xlsx",sheet\_name="中医病案sample")

ls = pd.DataFrame()

ls[0] = df['病人ID'].astype(str)

ls[1] = df['主诉'].astype(str)

#根据id合并

ls = ls.groupby(0)[1].agg(",".join)

for i in df.index:

# print(df["病人ID"][i],ls[df["病人ID"][i]])

#set 去重

set1 = set(ls[df["病人ID"][i]].replace("nan","").split(","))

df["主诉"][i] = re.sub("['{}]","",str(set1)).replace(","," ")

#df["主诉"][i]

df.to\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\2.缺失值处理\中医病案数据(填补2).xlsx",sheet\_name="中医病案sample",index=False)

**结果：**

 

## 三．分解

**步骤：**

1. 编写中药替换字典，映射为可分解的中药
2. 编写替换函数
3. 调用自身，查找替换
4. 输出文件

**方法：**使用pandas和numpy、pandas、re中的apply等等

**处理过程：**

#（3）分解：对处方中含有各字的中药分解，如煅龙牡是煅牡蛎和煅龙骨两味药；如焦楂曲是炒山楂和炒神曲两味药，等等。

import pandas as pd

import numpy as np

import re

ls = pd.Series(np.arange(2416))

df = pd.read\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\2.缺失值处理\中医病案数据(填补2).xlsx",sheet\_name="中医病案sample")

ls = df["处方1"].astype(str)

#建立替换字典

dic = {'煅龙牡':'煅龙牡(煅牡蛎 煅龙骨)',

'焦楂曲':'焦楂曲(炒山楂 炒神曲)'}

for i in dic:

print(i,dic[i])

#ls = ls.apply(lambda x:x.replace(i,dic[i]) for i in dic)

#替换函数

def substitute(string,dic):

for i in dic:

string = string.replace(i,dic[i])

return string

#string = "玄参15 焦楂曲15 龙齿15 黄芩10 黄连10 栀子10 煅龙牡30 黄芪30 桂枝10 白芍15 神曲10 夏枯草10 生晒参10 炙甘草6 14付"

#print(substitute(string,dic))

ls = ls.apply(lambda x:substitute(x,dic))

df["处方1"] = ls

df.to\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\3.分解\中医病案数据(分解).xlsx",sheet\_name="中医病案sample",index=False)

**结果：**



## 四．一致性

**步骤：**

1. 读文件
2. 查找'现病史、自诉、症状'相同而且是同一ID的所有病例
3. 存入文件，方便修改

**方法：**主要用了根据症状分类，然后过滤出相同ID相同症状的数据

**处理过程：**

import pandas as pd

import numpy as np

import re

#有复诊病人的“现病史、自诉和症状列”描述的一致性情况。各列的值是否描述了相应内容，如脉象、舌象是否混杂。

df = pd.read\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\3.分解\中医病案数据(分解).xlsx",sheet\_name="中医病案sample")

a = df.groupby(["现病史、自诉、症状"]).count()>1

price = a[a['病人ID'] == True].index

repeat\_df = df[df['现病史、自诉、症状'].isin(price)]

#输出相同'现病史、自诉、症状'，放便人工修改

df.to\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\4.一致性\中医病案数据(一致性).xlsx",sheet\_name="中医病案sample",index=False)

**结果：**



## 五．特征提取

**步骤：**

1. 读数据
2. 分别根据性别（女|男|无），年龄分类（成|小儿|无|其他）
3. 分别输出，存入不同的表中

**方法：**主要根据年龄，性别分类，方便以后分析相同中药对不同人群的反应

**处理过程：**

import pandas as pd

import numpy as np

import re

df = pd.read\_excel(r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\3.分解\中医病案数据(分解).xlsx",sheet\_name="中医病案sample")

df["性别"] = df["性别"].fillna("无")

grouped = df.groupby("性别")

#根据性别分成3个数据框

dfSex1 = grouped.get\_group('女')

dfSex2 = grouped.get\_group('男')

dfSex3 = grouped.get\_group('无')

grouped1 = df.groupby("性别")

df["年龄"] = df["年龄"].fillna("无")

dfAge1 = df[df["年龄"] == '成']

dfAge2 = df[df["年龄"] == '小儿']

dfAge3 = df[df["年龄"] == '无']

dfAge4 = df[df['年龄'].apply(lambda x:x!='成' and x!='无')]

path1 = r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\5.特征提取\中医病案数据(性别).xlsx"

path2 = r"C:\Users\meika\Desktop\Temp\蔡\实训汇总2\5.特征提取\中医病案数据(年龄).xlsx"

writer = pd.ExcelWriter(path1)

dfSex1.to\_excel(writer,sheet\_name="女",index=False)

dfSex2.to\_excel(writer,sheet\_name="男",index=False)

dfSex3.to\_excel(writer,sheet\_name="无",index=False)

writer.save()

writer = pd.ExcelWriter(path2)

dfAge1.to\_excel(writer,sheet\_name="成",index=False)

dfAge2.to\_excel(writer,sheet\_name="小儿",index=False)

dfAge3.to\_excel(writer,sheet\_name="无",index=False)

dfAge4.to\_excel(writer,sheet\_name="其他",index=False)

writer.save()

**结果：**

