

****

数据分析

**疾病分析**

|  |  |  |
| --- | --- | --- |
| 组长： | 王楚 | ZF1821237 |
| 组员： | 王冰洋 | ZF1821234 |
|  | 田浩 | ZF1821228 |
|  | 薛少杰 | ZF1821316 |
|  | 邹国际 | ZF1821354 |
|  | 韩禹 | ZF1821123 |
| 2019年6月10日 | | |

目录

[1需求说明 2](#_Toc8408898)

[2分析流程及所用技术 3](#_Toc8408899)

[3 业务模块分析 4](#_Toc8408900)

[3.1数据爬取（田浩） 4](#_Toc8408901)

[3.2一级科室统计分析（邹国际） 4](#_Toc8408902)

[3.2.1需求分析 4](#_Toc8408903)

[3.2.2实现思路 4](#_Toc8408904)

[3.2.3所用技术 6](#_Toc8408905)

[3.2.4结果展示 6](#_Toc8408906)

[3.3患者疾病名称分析 9](#_Toc8408907)

[3.3.1总体病情数量统计(王楚) 9](#_Toc8408908)

[3.3.3分析不同性别易得疾病(王楚) 11](#_Toc8408909)

[3.3.4分析不同年龄易得疾病(王楚) 12](#_Toc8408910)

[3.3.5分析不同科室都有哪些疾病(王冰洋) 15](#_Toc8408911)

[3.4患者病情自述分析(王冰洋) 16](#_Toc8408912)

[3.4.1所用技术 16](#_Toc8408913)

[3.4.2代码目录结构 16](#_Toc8408914)

[3.4.3实现思路 17](#_Toc8408915)

[3.4.4分析结果 17](#_Toc8408916)

[3.4.5词云表 18](#_Toc8408917)

[3.5医生病情分析(王冰洋) 19](#_Toc8408918)

[3.5.1所用技术 19](#_Toc8408919)

[3.5.2代码目录结构 19](#_Toc8408920)

[3.5.3实现思路 20](#_Toc8408921)

[3.5.4分析结果 20](#_Toc8408922)

[3.5.5医生词云表 20](#_Toc8408923)

[3.6指导建议分析(王冰洋) 21](#_Toc8408924)

[3.6.1所用技术 21](#_Toc8408925)

[3.6.2代码目录结构 21](#_Toc8408926)

[3.6.3实现思路 22](#_Toc8408927)

[3.6.4分析结果 22](#_Toc8408928)

[3.6.5词云图 23](#_Toc8408929)

[3.7病人基本信息分析（薛少杰） 24](#_Toc8408930)

[3.7.1需求 24](#_Toc8408931)

[3.7.2实现思路 24](#_Toc8408932)

[3.7.3代码结构及部分结果 24](#_Toc8408933)

[4分析结论 31](#_Toc8408934)

# 1需求说明

爬取寻医问药网上病人与医生的问答，进行数据分析，分析角度如下

# 2分析流程及所用技术

# 3 业务模块分析

## 3.1数据爬取（田浩）

选用java技术爬取寻医问药网，并将数据存储到hdfs上，共爬取了两万条数据，下面是爬取的部分数据

|  |
| --- |
| 五官科|耳鼻喉科|2019-05-04 23:51:42|男|2岁|我儿子两岁半，鼻子一直不通气，检查得了鼻息肉，什么时候动手术比较好|专长：耳鼻喉科疾病的诊治。|问题分析：儿童鼻塞的常见原因是扁桃体腺样体肥大，其它常见问题是过敏性鼻炎引起的 |指导建议：鼻息肉常见原因是鼻腔粘膜存在过敏性炎症，长期水肿形成息肉组织，二岁儿童发生鼻息肉很少见，建议去正规医院就诊明确是否是鼻息肉| |

数据内容格式为

|  |
| --- |
| 科室|病情名称|咨询时间|性别|年龄|患者自述|医生专长|病情分析|指导意见 |

## 3.2一级科室统计分析（邹国际）

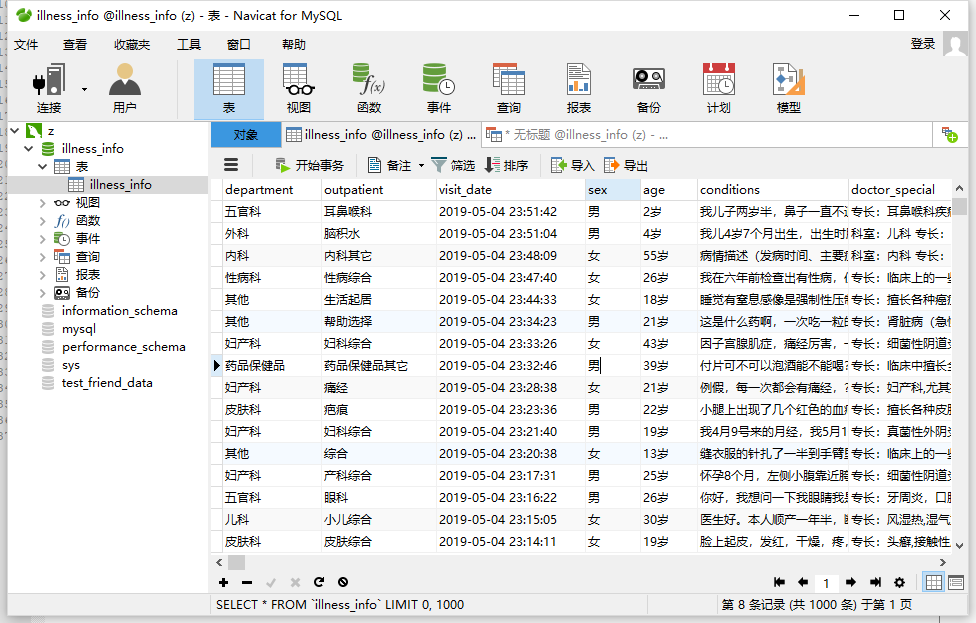
### 3.2.1需求分析

对一级科室统计，包含不区分年龄性别科室分析、不同年龄病情所属科室分析、不同性别病情所属科室分析。

### 3.2.2实现思路

由于功能相对简单，通过对爬虫的数据进行处理导入MySQl，用查询进行筛选，对筛选结果进行数据分析。

导入MySQL



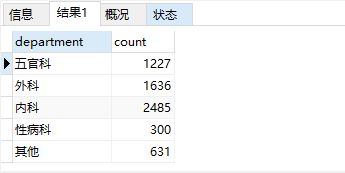
|  |
| --- |
| #各科室占百分比  SELECT  department,  COUNT(1) AS count  FROM  illness\_info  WHERE  department != ''  GROUP BY  department  #根据年龄  SELECT  department,  age,  COUNT(1) AS count  FROM  illness\_info  WHERE  age != '未知'  AND department !=''  GROUP BY  department,age  #根据性别  SELECT  department,  sex,  COUNT(1) AS count  FROM  illness\_info  WHERE  department !=''  AND sex !='未知'  GROUP BY  department,sex |

### 3.2.3所用技术

MySQL、excel

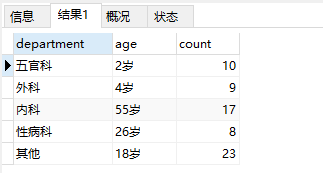
### 3.2.4结果展示

（1）查询科室，department

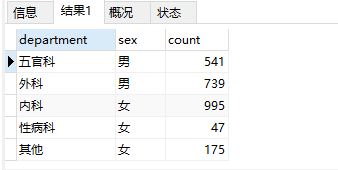


可视化结果如下

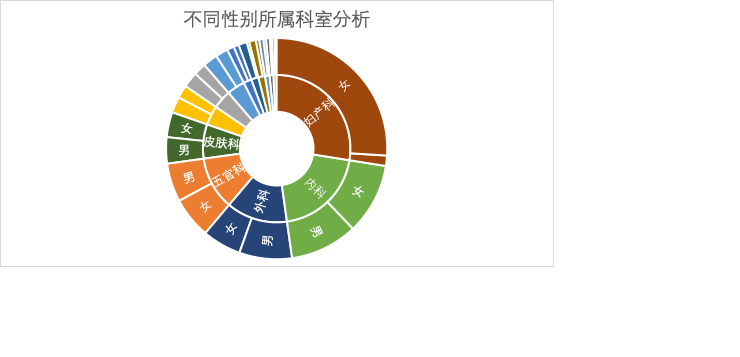
（2）按年龄，departmentBYage

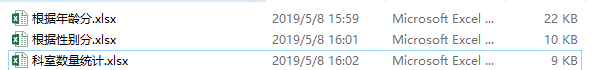


（3）按性别，departmentBYsex



可视化结果如图





## 3.3患者疾病名称分析

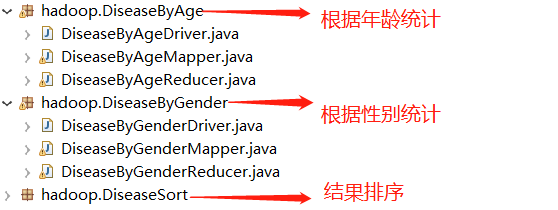
对患者疾病名称进行分析，选用技术为mapper/reducer/sparkrdd,从hdfs上面获取数据进行分析，并将结果输出在hdfs上

### 3.3.1总体病情数量统计(王楚)

#### （1）选用技术

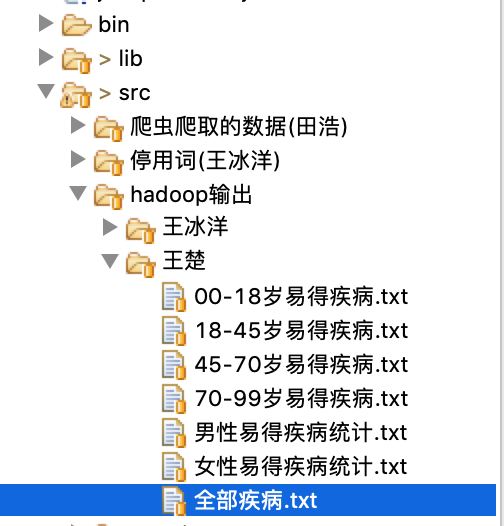
hadoop-map/reduce hdfs

#### （2）代码目录结构



#### （3）分析结果

结果文件目录



分析部分数据如下

|  |
| --- |
| 眼科 239  呼吸内科 241  男科 242  月经不调 372  皮肤综合 432  妇科综合 536  产科综合 877 |

数据可视化图形为



结果可以看出产科综合，妇科综合和皮肤综合是占比最高的

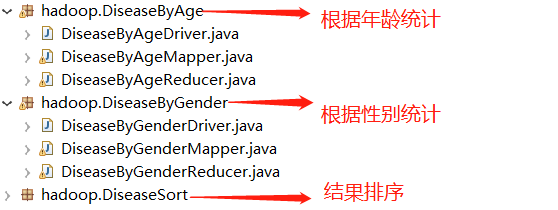
### 3.3.3分析不同性别易得疾病(王楚)

分别分析男性和女性易得的疾病，分析结果数据文件所在目录名为/Hadoop输出/王楚/男性易得疾病及女性易得疾病，

#### （1）选用技术

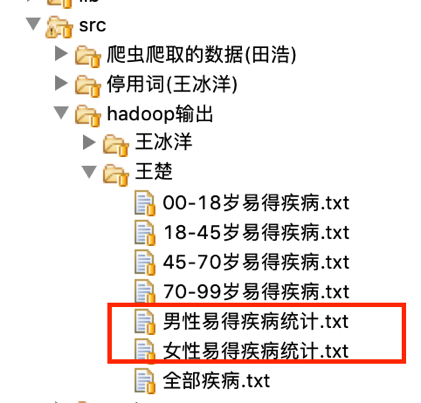
hadoop-map/reduce hdfs

#### （2）代码目录结构



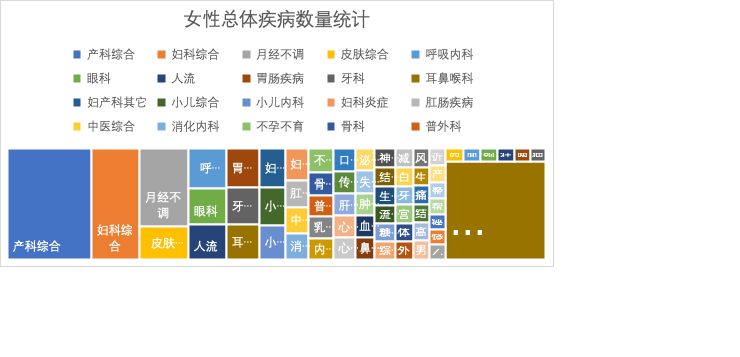
#### （3）分析结果

结果文件有两个



可视化图形为





从结果可知：

女性更多就诊产科，妇科，皮肤科，月经不调

男性更多就诊男科，眼科，耳鼻喉科和白癜风

猜想：女性月经不调概率比较高，男性的白癜风的比例要高于女性

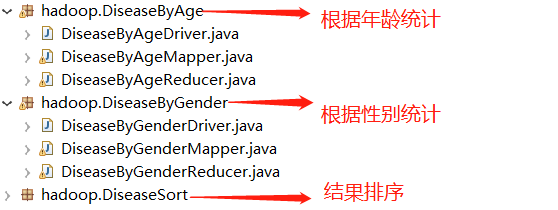
### 3.3.4分析不同年龄易得疾病(王楚)

分别分析18岁以下，18-45岁，45-70岁，70-99岁各个阶段不同年龄更容易患的疾病，分析结果数据文件所在目录名为/Hadoop输出/王楚/，分析结果图形如下

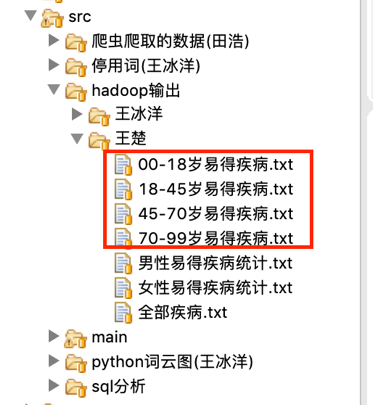
#### （1）选用技术

hadoop-map/reduce hdfs

#### （2）代码目录结构



（3）结果文件

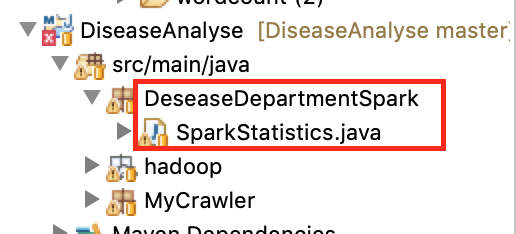


### 3.3.5分析不同科室都有哪些疾病(王冰洋)

#### （1）所用技术

Sparkrdd

#### （2）代码结构目录



#### （3）分析结果如下

|  |
| --- |
| (传染科,狂犬病/肝病/传染科综合/结核病/乙肝/流感/大小三阳/肝炎/流行性腮腺炎/寄生虫疾病/丙肝)  (整形美容,整形美容其它/美白祛斑/面部整形/生殖器整形/化妆品/减肥瘦身/丰胸美体/肢体整形/注射美容)  (心理诊所,心理咨询/心理科其它/社交恐慌/性心理障碍/强迫症)  (妇产科,妇科综合/痛经/产科综合/生殖孕育/妇科炎症/月经不调/不孕不育/人流/妇产科其它/助孕技术/产后妇科/子宫肌瘤/宫颈糜烂/女性不孕/宫外孕/卵巢囊肿/宫颈疾病/外阴白斑)  (药品保健品,药品保健品其它/西药/用药常识与不良反应/中草药问题)  (五官科,耳鼻喉科/眼科/牙科/口腔科/鼻炎/耳鸣耳聋/口腔溃疡/斜视/唇腭裂/口臭/中耳炎/咽炎/牙周炎/白内障/近视/眼科炎症/种植牙/弱视/口腔黏膜/龋齿/五官科综合)  (内科,内科其它/消化内科/神经内科/心血管内科/高血压/呼吸内科/胃炎/中风/肝胆疾病/血液内科/胆囊疾病/胃肠疾病/肿瘤科/糖尿病/肾积水/痛风/肾病综合征/心脏病/脑血管病/哮喘/风湿免疫/甲状腺疾病/面瘫/头痛/肾内科/肝血管瘤/内分泌/癫痫病/白血病/肝硬化/肝腹水/脑萎缩/截瘫/肺气肿/脊髓病变/重症肌无力/急慢性肾炎/酒精肝/三叉神经痛)  (皮肤科,疤痕/皮肤综合/斑秃/皮肤过敏/白癜风/荨麻疹/痤疮/生发/腋臭/灰指甲/植发/湿疹/鱼鳞病/胎记/牛皮癣/带状疱疹)  (外科,脑积水/神经外科/前列腺/烧烫伤/外科其它/男性不育/男科/乳腺外科/早泄/肛肠疾病/泌尿外科/普外科/强直性脊柱炎/颈椎病/骨科/心胸外科/颈腰椎病/疝气/心血管外科/结石科/阳痿/痔疮/性功能障碍/腰椎间盘突出/股骨头坏死/静脉曲张/脑外伤后遗症/肌肉萎缩)  (其他,生活起居/帮助选择/综合/两性健康/体检/饮食营养/眩晕科/更年期综合症/不良嗜好)  (中医科,中医综合/针灸按摩/中医保健/中医妇科)  (儿科,小儿综合/唐氏综合症/水痘/小儿内科/新生儿疾病/自闭症/抽动症/手足口病/血管瘤/多动症/性早熟/小儿脑瘫/小儿外科/小儿眼科)  (精神科,失眠抑郁/精神科综合/妄想症/心理障碍/精神分裂症)  (性病科,性病综合/艾滋病/尖锐湿疣/梅毒/生殖器疱疹/淋病) |

可视化结果如下

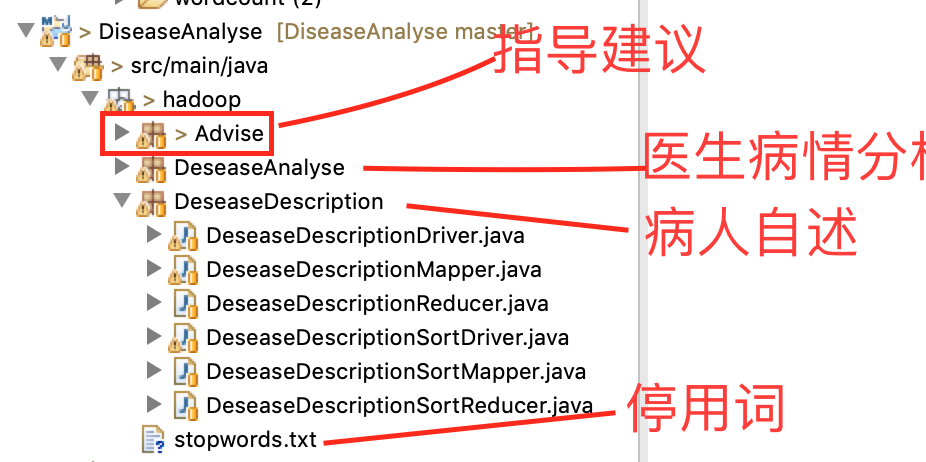
## 3.4患者病情自述分析(王冰洋)

### 3.4.1所用技术

Mapper/reducer

### 3.4.2代码目录结构

Hadoop结构代码如下



### 3.4.3实现思路

通过对病人自己描述的病情进行分词，会发现分出的词语很多包括一些形容词及跟病情无关的词语，我们把这部分的词语进行剔除，

与病情无关的词为stopwords.txt

分词后结果文件为deseaseDescription.txt

### 3.4.4分析结果

部分结果如下后面的数字代表统计出来的次数

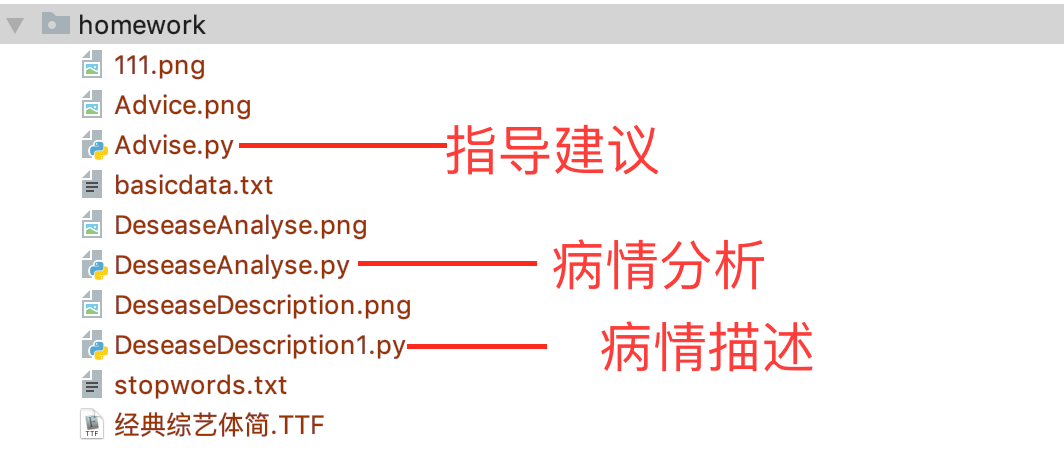
|  |
| --- |
| 晚上 328  孩子 398  宝宝 399  痛 424  手术 428  痒 519  月经 592  怀孕 812  疼 877 |

可视化图形为

更多的病人在描述时候，用到疼，痒，痛这样的词，说明很多疾病给人的直观感受是这样的，很多人女性会受到怀孕和月经的苦恼，小孩子和婴儿更容易生病，大多数疾病是在晚上发生的

### 3.4.5词云表

词云图代码为python生成，代码结构



分析结果



可以通过词频统计看出来，病人大部分描述都是痛，痒，感冒，咳嗽，白癜风为大家提出最多的，很多人都是在晚上开始身体不舒服，女性和小孩更容易生病

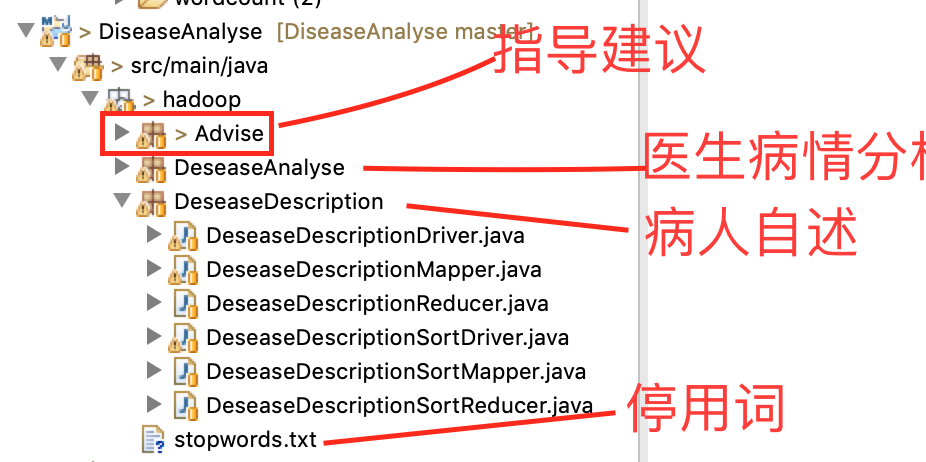
## 3.5医生病情分析(王冰洋)

### 3.5.1所用技术

Mapper/reducer

### 3.5.2代码目录结构

Hadoop结构代码如下



### 3.5.3实现思路

同病情自述分析

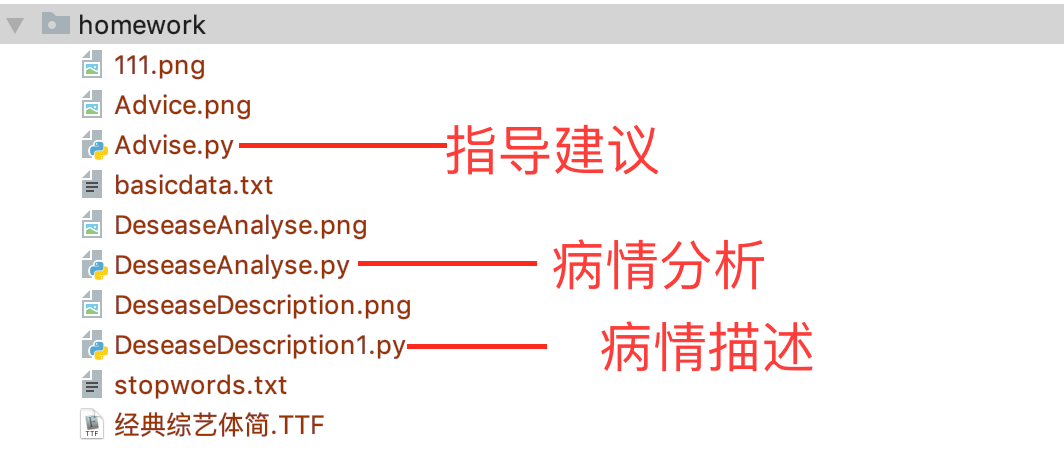
### 3.5.4分析结果

结果文件deseaseanalyse.txt

|  |
| --- |
| 慢性 197  出血 209  炎 214  子宫 214  饮食 233  皮肤 272  疼痛 276  孩子 278  手术 336  医院 362  炎症 363  月经 421  药物 446  怀孕 573  感染 589 |

### 3.5.5医生词云表

词云图代码为python生成，代码结构



分析结果



图形里面显示和患者自述很像

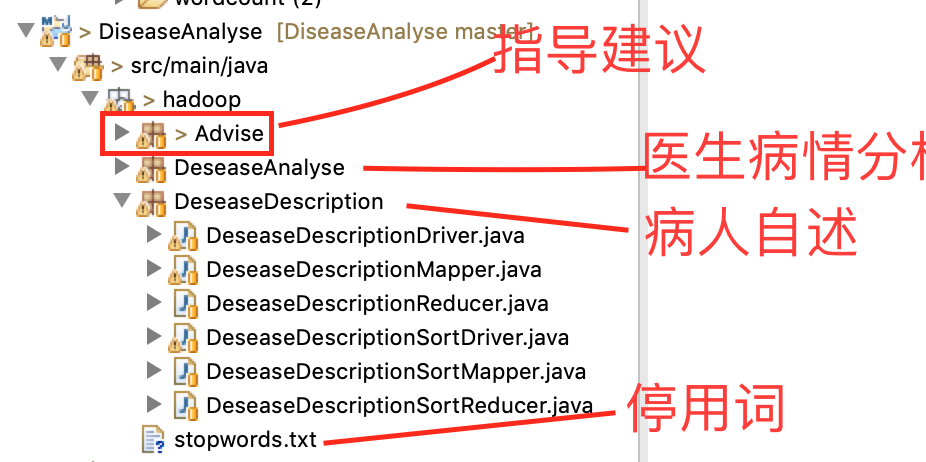
## 3.6指导建议分析(王冰洋)

### 3.6.1所用技术

Mapper/reducer

### 3.6.2代码目录结构

Hadoop结构代码如下



### 3.6.3实现思路

同病情自述分析

### 3.6.4分析结果

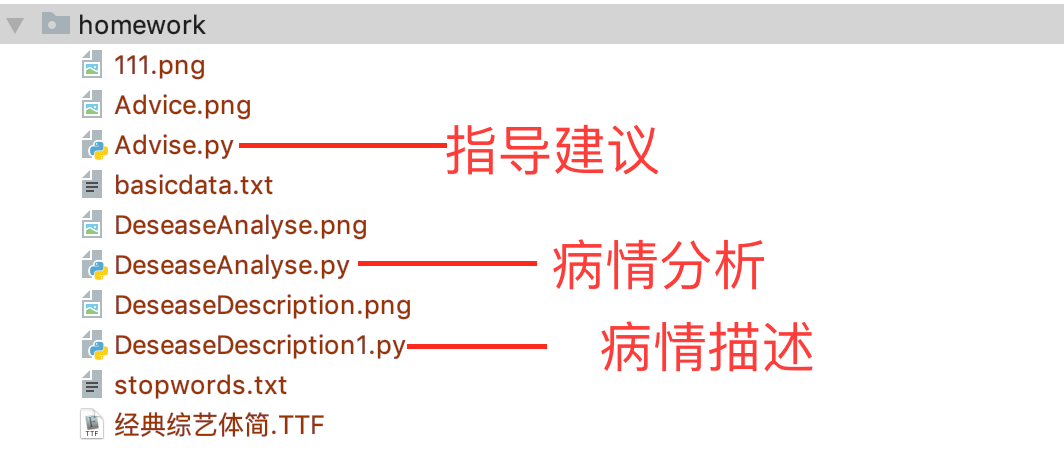
结果文件advise.txt

|  |
| --- |
| 清淡 506  多吃 548  保持 598  月经 653  感染 692  休息 694  观察 706  口服 789  怀孕 791  手术 848  服用 912  辛辣 926  饮食 1288  食物 1395  药物 1582  医院 2034 |

从图中可以看出，医生更多的建议是去医院，以及用药，通过食物来调整。避免辛辣饮食，多休息，多吃清淡食物，少吃刺激性的食物。

### 3.6.5词云图

词云图代码为python生成，代码结构



分析结果



可以看出医生给出的建议更多的是去医院就诊和观察，注意休息，避免刺激性的食物，多休息

## 3.7病人基本信息分析（薛少杰）

### 3.7.1需求

根据小组成员采用爬虫获得的数据，获得了广大网民在网络求医过程中，人员年龄，性别，咨询科室，访问时间等一系列信息。但是我们更关注，在这个过程中，网络求医这种新型医疗保健方式所隐藏的背后的规律。究竟是什么样的一群人带动了这个现象，他们以怎样的方式参与其中？这些都是我们所关心并且希望了解的。

### 3.7.2实现思路

通过SQL和Excel对原始数据进行初步清理，并导入数据库中，采用SQL语言进行相关信息的提取

### 3.7.3代码结构及部分结果

#### （1）病人咨询时间分析

代码结构：

#用HOUR()函数将取时间的小时，按小时分组判断出不同时间段网络求医人数

select

HOUR(d\_time) as 看病时间 ,count(d\_name) as 相同时间看病人数量

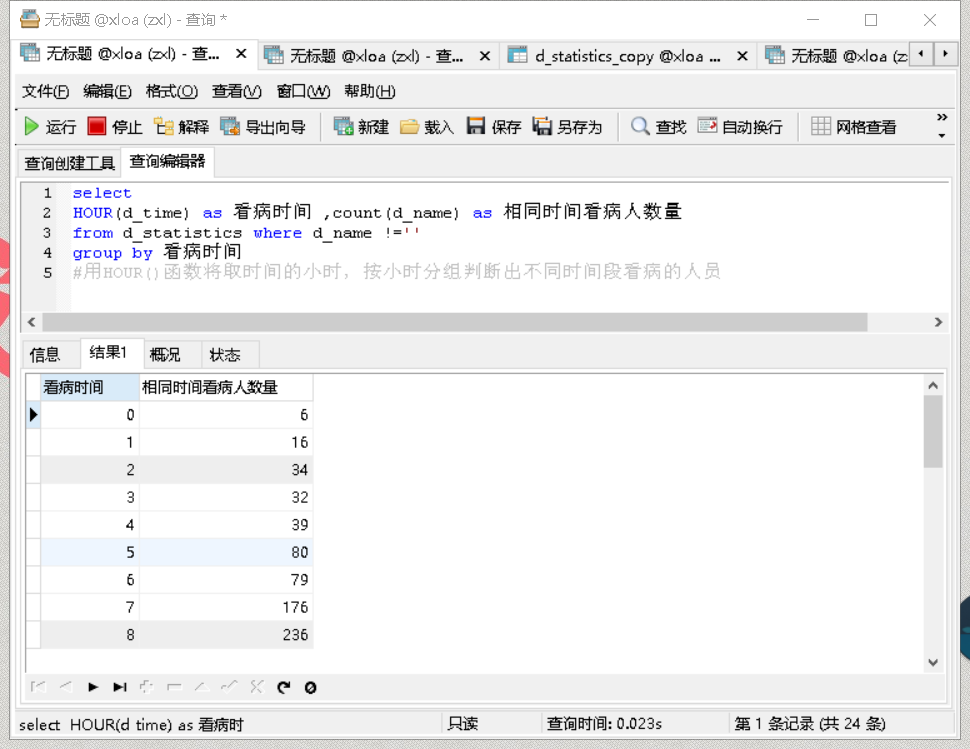
from d\_statistics where d\_name !=''

group by 看病时间

查询结果：网络求医时间分布.txt

|  |
| --- |
| 0 6  1 16  2 34  3 32  4 39  5 80  6 79  7 176  8 236  9 153  10 16  11 7  12 18  13 21  14 32  15 21  16 29  17 65  18 105  19 183  20 600  21 463  22 276  23 18501 |

注：其中第一列为时间，第二列为对应人数



可视化结果如下

大多数医生回复的时间在晚上11点以后，可以说明很多医生在那个时间段是一天中比较空闲的时间，由此可见，医务人员的工作非常累

#### （2）病人年龄分析

代码结构：

#用LENGTH()筛选出，正确的年龄看病记录，分组查看不同病情，不同性别的人员

select d\_age as 年龄,count(d\_age) as 数量

from d\_statistics\_copy where

d\_name !='' and LENGTH(d\_age) <10 and LENGTH(d\_age)>1 and d\_age!=' 0个月'

GROUP BY d\_age

查询结果：网络求医年龄分布.txt

|  |
| --- |
| 1 209  2 76  3 63  4 59  5 64  6 58  7 50  8 33  9 28  10 29  11 36  12 38  13 45  14 61  15 84  16 139  17 178  18 255  19 219  20 491  21 360  22 414  23 464  24 418  25 501  26 437  27 313  28 327  29 248  30 297  31 136  32 216  33 127  34 118  35 145  36 130  37 74  38 78  39 65  40 113  41 52  42 70  43 71  44 51  45 98  46 68  47 47  48 71  49 45  50 70  51 23  52 58  53 47  54 25  55 37  56 52  57 27  58 31  59 20  60 46  61 22  62 19  63 13  64 18  65 33  66 17  67 14  68 13  69 9  70 31  71 6  72 16  73 13  74 6  75 11  76 9  77 4  78 10  79 3  80 6  81 3  82 2  83 2  84 3  85 3  86 1  87 1  89 1  90 2  91 2  92 1  95 1  106 1 |

注：第一列为年龄，第二列为人数

可视化结果为

#### （3）病人性别分析

代码结构：

select d\_six as 性别,count(d\_six) as 数量

from d\_statistics\_copy

GROUP BY d\_six

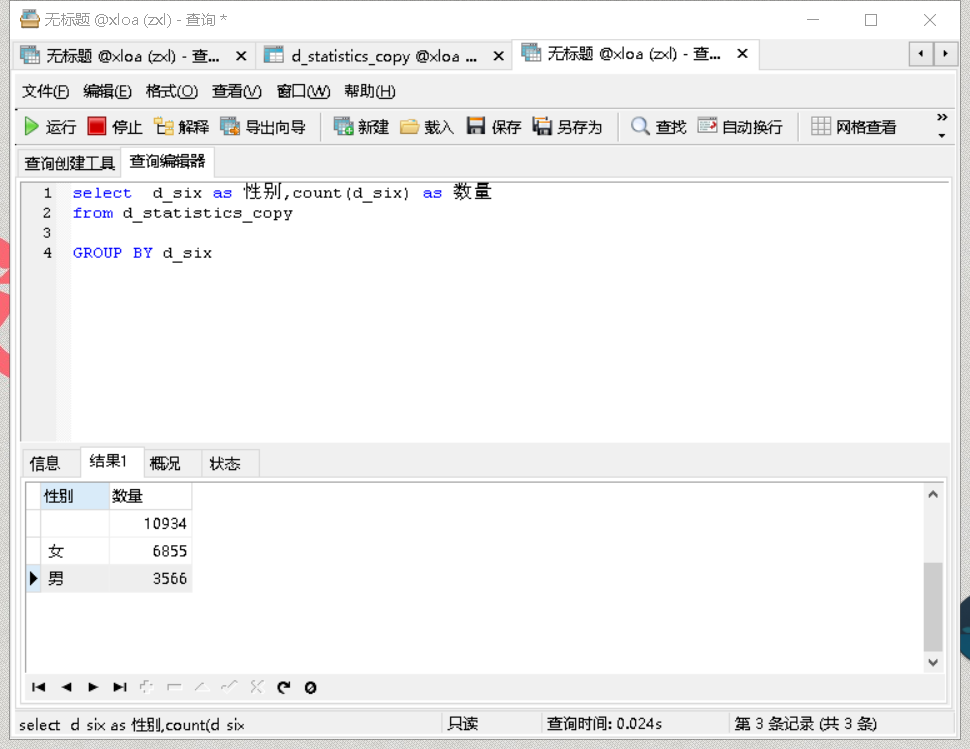
查询结果：网络求医性别分布.txt

10934

女 6855

男 3566

注：空白项为求医者未提供性别信息



可视化结果如下

没有提供性别信息的人员提供半数，不利于病情分析，剔除了未提供性别信息的人群后的性别分析

求医者体现出明显的性别差异，女士在关心生理健康方面远远超过男士

# 4分析结论