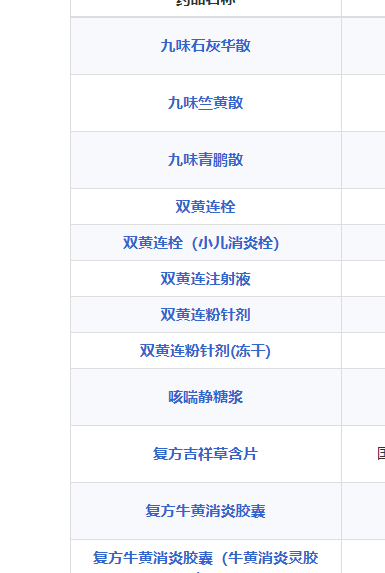
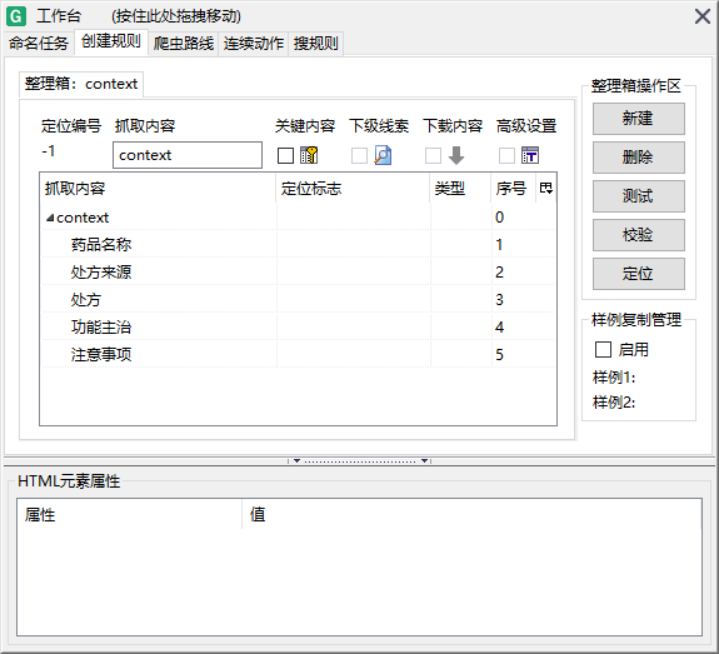
## 一．获取药方

方法：

通过使用现成的爬虫工具GooSeeker，爬取药智数据中包含治疗肺炎的所有药物（url：https://db.yaozh.com/chufang/122006.html），存放到本地文件。

步骤：

1. 制定爬取规则，爬取目录页面的所有url，方便下一步。 
2. 制定下级层次的爬取规则，爬取所需内容。
3. 执行爬虫，并将数据以xml文件的形式存放到本地（一个药方一个文件）。
4. 将所有药方合并为一个文件，方便后续操作。
5. 

## 二．构建词典

方法：

先对药方所有内容通过GetFileNewWords（）自动获取新词，存放到用户词典中，在从网上下载中药材名，治法治则名存放如用户词典。

步骤：

1. 设置软件编码格式位utf-8
2. 设置用户词典文件输出目录
3. 调用方法GetFileNewWords（）获取新词，去掉“#”存入用户词典。
4. 根据中药特征手动添加一些没有识别的词汇。
5. 

代码：

'''获取新词,构建用户词典'''

import pynlpir

pynlpir.open(encoding="utf8")

text\_path = "C:/Users/meika/Desktop/text.txt"

usrdict\_path = "C:/Users/meika/Desktop/usrdict.txt"

f = open(usrdict\_path,'a+')

#获取文件新词

result = pynlpir.nlpir.GetFileNewWords(text\_path,200,False)

for group in result.split("#"):

f.write(group+'\n')

f.close()

## 三．对药方分词处理

方法：

通过已经构建好的用户词典，将药方数据进行分词处理。

步骤：

1. 导入用户词典ImportUserDict（）
2. 对药方数据分词处理FileProcess（）
3. 输出分词好的数据存入到本地。
4. 

代码：

import pynlpir

pynlpir.open(encoding="utf8")

#导入用户词典

usrdict\_path = "C:/Users/meika/Desktop/usrdict.txt"

out\_path = "C:/Users/meika/Desktop/text\_out.txt"

text\_path = "C:/Users/meika/Desktop/text.txt"

pynlpir.nlpir.ImportUserDict(usrdict\_path)

pynlpir.nlpir.FileProcess(text\_path,out\_path,True)

## 四．文档特征提取

方法：

提取关键词，看到词性和频率，对于一些不希望看到的词语可添加到停用词库中，过滤出的内容生成新的文件。

步骤：

1. GetFileKeyWords（）获取关键词，输出词语，词频，和频数
2. 手动创建停用词库，并将其导入
3. 调用方法，返回不在停用词库中的词，生成过滤后的新文件。
4. 

代码：

'''关键词提取'''

import pynlpir

text\_path = "C:/Users/meika/Desktop/text.txt"

result = pynlpir.nlpir.GetFileKeyWords(text\_path,200,True)

for group in result.split("#"):

try:

w,f,wt,fq = group.split("/")

print w,f,wt,fq

except:

pass

'''停用词过滤'''

textout\_path = "C:/Users/meika/Desktop/text\_out.txt"

stopwords\_path = "C:/Users/meika/Desktop/stopwords.txt"

stopwords = []

for line in open(stopwords\_path).readlines():

stopwords.append(line.strip())

out = open("C:/Users/meika/Desktop/text\_filtered.txt",'a+')

for line in open(textout\_path,"r").readlines():

remain=[]

word\_list = line.strip().split(" ")

if len(word\_list)>1:

for word in word\_list:

g = word.split("/")

if len(g)>1:

[w,f] = g

if 'g' not in w: #去除中药的重量

if w not in stopwords:

remain.append(word)

out.write(" ".join(remain)+"\n")

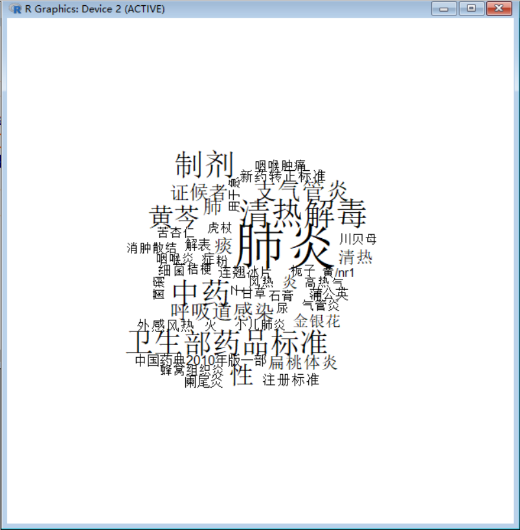
out.close();

## 五．词频统计

方法：

这里我们主要统计名词。先统计所有词性，在调用方法抽取名词，最后做出统计。

步骤：

1. 读入之前过滤完后的文件。
2. 获取所有名词存放到列表中。
3. 通过nltk.FreqDist（）统计所有名词词频，输出文件。
4. 
5. 使用R获得词云。

代码：

'''统计词频'''

import nltk

words = []

#统计所有词性

for line in open("C:/Users/meika/Desktop/text\_filtered.txt","r").readlines():

remain = []

word\_list = line.strip().split(" ")

if len(word\_list)>1:

for word in word\_list:

[w,f] = word.split("/")

words.append(w)

#抽取名词

words\_n =[]

for line in open("C:/Users/meika/Desktop/text\_filtered.txt","r").readlines():

word\_list = line.strip().split(" ")

if len(word\_list)>1:

for word in word\_list:

g = word.split("/")

if len(g)>1:

[w,f]=g

if f[0] in ["n"]: #v,a

print w,f

words\_n.append(word)

#统计词频

out = open("C:/Users/meika/Desktop/sum\_freq\_n.csv",'a+')

freq = nltk.FreqDist(words\_n)

for(w,f) in freq.most\_common(50):

print w.rstrip("//ng)"),f

out.write(w.rstrip("//ng)")+","+str(f)+"\n")

out.close()

library(wordcloud)

Data = read.csv(file = "C:/Users/meika/Desktop/sum\_freq\_n.csv")

attach(Data)

attach(Data)

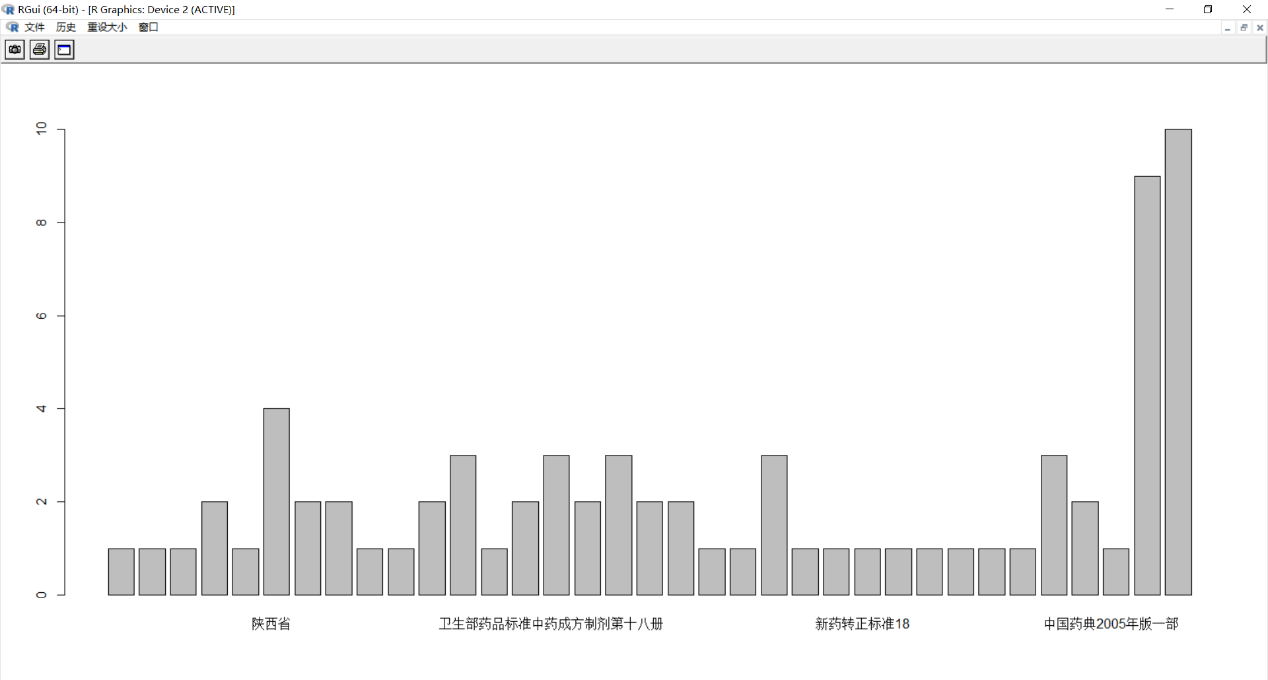
wordcloud(words=word,freq=freq,min.freq=5,max.words=50,random.order=TRUE)

## 六．中药来源分析

方法：

这里用之前爬取的药方汇总文件，统计药方来源分布。

步骤：

1. 读入xlsx格式的数据。
2. 用Table（）转化数据，方便统计
3. 使用barplot画出药方分布条形图
4. 

代码：

library(xlsx)

read.xlsx("C:/Users/meika/Desktop/text/context\_1583853237845.xlsx", 1,encoding="UTF-8")

attach(data)

attach(data)

data1 = 处方来源

barplot(table(data1))