作业二. 绘制比较语料库GENIA和AGAC的词云

生信1802-余思克-2018317220208

实验目的

GENIA和AGAC是两种生物医学领域内的语料库，本次实验的目的是分别统计两个语料库中的词频，并以词云的方式进行可视化，从而比较GENIA和AGAC两种语料库之间的差异。

实验步骤

1. R语言相关依赖包的加载

需要使用相关的R包进行语料库的读取，对语料库进行数据清理以及对词频信息进行可视化，因此使用了`NLP`，`tm`，`wordcloud2`三个拓展包，需要注意的是，若没有安装上述R包，则需要先使用`install.packages()`函数进行R包的安装，最好使用国内镜像进行包的下载与安装。

```
library(NLP)
library(tm)
library(wordcloud2)
```

2. 数据的读取

本步骤通过`Corpus()`函数读取数据并建立语料库。该函数对数据读取操作的格式要求与常规读取方法不同，需要以每一层路径为一个独立的单元，顺次合成一个路径变量，然后由该函数进行读取。数据读取路径仅指定到数据所在的文件夹名。

```
setwd("F:\\HZAU-course\\BioNLP\\TASK\\task2")
cname = file.path("F:", "HZAU-course", "BioNLP", "TASK", "task2", "GENIA")
docs_GENIA = Corpus(DirSource(cname)) # 读取GENIA语料库数据

setwd("F:\\HZAU-course\\BioNLP\\TASK\\task2")
cname = file.path("F:", "HZAU-course", "BioNLP", "TASK", "task2", "AGAC")
docs_AGAC = Corpus(DirSource(cname)) # 读取AGAC语料库数据
```

3. 数据清洗

由于数据中存在许多噪声，会干扰后续的分析与可视化，本步骤进行数据清洗以清除噪声。清洗过程主要包括：去除标点及特殊符号、去除数字、大写字母转换为小写字母、去除停止语、去除特殊的无意义信息、纯文本化转换、词干化处理等。

```
docs_GENIA = tm_map(docs_GENIA, removePunctuation)
for(j in seq(docs_GENIA))
{
    docs_GENIA[[j]] = gsub("/", " ", docs_GENIA[[j]])
    docs_GENIA[[j]] = gsub("@", " ", docs_GENIA[[j]])
    docs_GENIA[[j]] = gsub("\\|", " ", docs_GENIA[[j]])
}
docs_AGAC = tm_map(docs_AGAC, removePunctuation)
```

```
for(j in seq(docs_AGAC))
{
    docs_AGAC[[j]] = gsub("/", " ", docs_AGAC[[j]])
    docs_AGAC[[j]] = gsub("@", " ", docs_AGAC[[j]])
    docs_AGAC[[j]] = gsub("\\|", " ", docs_AGAC[[j]])
} # 去除标点及特殊符号

docs_GENIA = tm_map(docs_GENIA, removeNumbers)
docs_AGAC = tm_map(docs_AGAC, removeNumbers) # 去除数字

docs_GENIA = tm_map(docs_GENIA, tolower)
docs_AGAC = tm_map(docs_AGAC, tolower) # 大写字母转换为小写字母

docs_GENIA = tm_map(docs_GENIA, removeWords, stopwords("english"))
docs_AGAC = tm_map(docs_AGAC, removeWords, stopwords("english")) # 去除英语停止语

docs_GENIA = tm_map(docs_GENIA, removeWords, c("department", "email", "doi",
"center", "sciences", "pubmed", "nature", "university", "pmid", "author",
"school", "research"))
docs_AGAC = tm_map(docs_AGAC, removeWords, c("department", "email", "doi",
"center", "sciences", "pubmed", "nature", "university", "pmid", "author",
"school", "research")) # 去除特殊的无意义信息

docs_GENIA = tm_map(docs_GENIA, PlainTextDocument)
docs_AGAC = tm_map(docs_AGAC, PlainTextDocument) # 纯文本化处理

docs_GENIA = tm_map(docs_GENIA, stemDocument)
docs_AGAC = tm_map(docs_AGAC, stemDocument) # 词干化处理
```

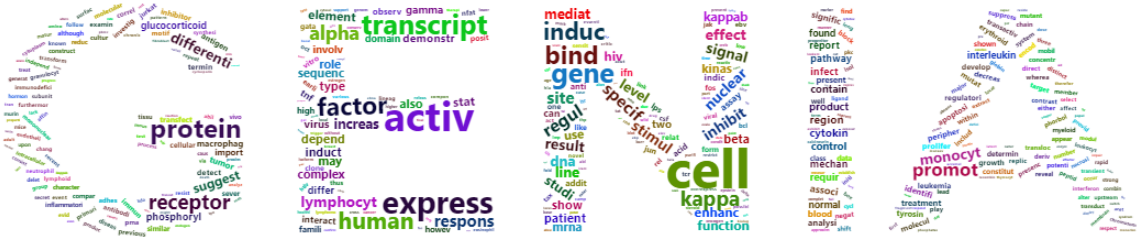
4. 词频统计与可视化词云绘制

本步骤将针对前一步清洗后得到的数据，进行词频统计，将每个单词与其出现的次数相对应，并对词频超过5的单词使用 `wordcloud2` 包中的 `letterCloud()` 函数绘制字符形状的词云，直观反映出GENIA语料库与AGAC语料库的内容构成特点。由于不同单词的词频跨度较大，为了使绘制的词云更加便于观察，将词频进行了取根号的处理从而使得不同单词的词频差距变得柔和。

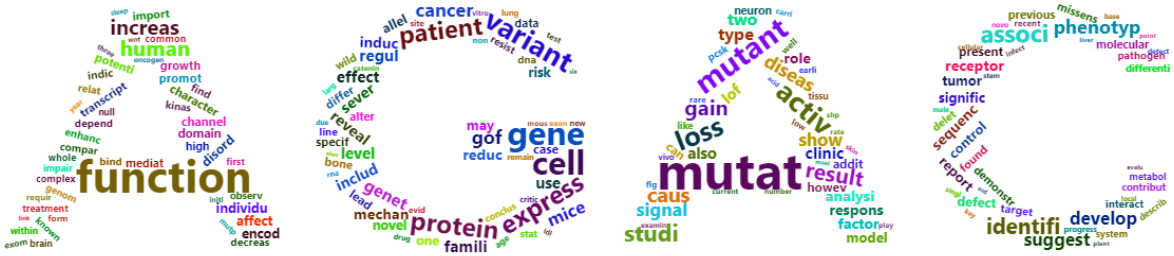
```
dtm_GENIA = DocumentTermMatrix(docs_GENIA)
freq_GENIA <- colSums(as.matrix(dtm_GENIA))
freq_GENIA = data.frame(word = names(freq_GENIA), freq =
data.frame(freq_GENIA)$freq_GENIA)
dtm_AGAC = DocumentTermMatrix(docs_AGAC)
freq_AGAC <- colSums(as.matrix(dtm_AGAC))
freq_AGAC = data.frame(word = names(freq_AGAC), freq =
data.frame(freq_AGAC)$freq_AGAC) # 词频统计

order_temp = order(freq_GENIA$freq, decreasing = T)
freq_GENIA = freq_GENIA[order_temp,]
freq_GENIA$freq = freq_GENIA$freq ^ 0.6
letterCloud(freq_GENIA[freq_GENIA$freq >= 5,], "GENIA", size = 0.27)
order_temp = order(freq_AGAC$freq, decreasing = T)
a = freq_AGAC[order_temp,]
a$freq = a$freq ^ 0.6
letterCloud(a[a$freq >= 5,], "AGAC", size = 0.27) # 绘制词云
```

实验结果

1. GENIA语料库词云



2. AGAC语料库词云



讨论

为了突出语料库的值观特征，本实验以语料库名称为形状绘制了词云，然而在绘制时，出现了错误，绘制出的词云没有单词的显示，只有黑色的形状背景。

通过查询资料发现，其原因是因为在R语言中直接使用`install.packages()`函数安装的`wordcloud2`包不是最新的，存在一些bug，于是通过`install_github('lchffon/wordcloud2')`函数直接从GitHub地址中重新安装`wordcloud2`包之后，再次运行`letterCloud()`函数进行词云绘制，bug成功消除。

通过对GENIA和AGAC语料库的可视化词云图片进行观察可以发现，GENIA语料库中出现频率较多的单词有"cell", "activ", "express", "　transcript", "gene", "protein", "bind", "factor", "induc"等，这些单词广泛涵盖了分子生物学层面的各个研究方向的信息。而AGAC语料库中出现频率较多的单词有"mutat", "function", "gene", "cell", "variant"等，可以看出来这些单词涵盖面并不广泛，而是集中在了基因突变、功能变化这一方面。结合上一项课程作业对TTR分析的结果可以认为，GENIA语料库是生物医学方面内容覆盖面广，宽泛的语料库，AGAC语料库是生物医学方面专注于疾病与基因突变、功能之间联系的小而精的语料库，若专注于开展基因突变与功能相关研究，AGAC语料库可能更加合适更加出色。