作业一. 分析GENIA和AGAC语料库语言丰富度的差异

生信1802-余思克-2018317220208

实验目的

语料库指经科学取样和加工的大规模电子文本库，其中存放的是在语言的实际使用中真实出现过的语言材料。本次实验的目的是通过计算语料库GENIA和AGAC的TTR，比较TTR的差异，从而从中分析探讨两种语料库之间的语言丰富度的差异。

实验步骤

1. 数据获取

GENIA语料库是在GENIA项目范围内编译和注释的生物医学文献的主要集合。该语料库是为了支持分子生物学领域的信息提取和文本挖掘系统的开发和评估。

GENIA语料库：http://www.nactem.ac.uk/GENIA/current/GENIA-corpus/Part-of-speech/GENIAcorpus3.02p.tgz

AGAC语料库旨在捕获在致病性背景下突变基因的功能变化。 它被设计为包括与分子和细胞水平上的遗传变异和即将发生的表型改变有关的实体，重点是追踪LOF和GOF突变的生物学语义。

AGAC语料库：http://pubannotation.org/projects/AGAC_training/annotations.tgz

```
wget http://www.nactem.ac.uk/GENIA/current/GENIA-corpus/Part-of-speech/GENIAcorpus3.02p.tgz # 下载GENIA语料库

wget http://pubannotation.org/projects/AGAC_training/annotations.tgz # 下载AGAC语料库
```

2. 数据预处理

本实验需要将得到的语料库进行一些预处理，然后才能进行分析。由于语料库数据中存在一些诸如标点、下划线之类的符号，我们需要将其去除；由于单词大小写对其语义不会造成影响，所以我们需要将所有字符统一转换为小写形式；最终为了便于后续数据分析，我们需要将数据格式化为一个单词单独占据一行的格式。

```
cat GENIAcorpus3.02.pos.txt | sed -r "s/(.*\/).*/\1/" | tr -cs "[:alnum:]" "\n" | tr "[:upper:]" "[:lower:]" | tr -d "[:punct:]" > GENIA.txt # GENIA语料库数据预处理

cat AGAC_training/txt/* | tr -cs "[:alnum:]" "\n" | tr "[:upper:]" "[:lower:]" | tr -d "[:punct:]" > AGAC.txt # AGAC语料库数据预处理
```

3. TTR的计算

语义丰富度指标——TTR：通过将文本或话语中出现的类符（不同词的总数）除以其形符（词的总数）而获得的比率。高TTR意味着词汇变化程度高，低TTR则相反。为了减小误差，增加可信度，TTR的计算将采用抽样统计的方式，此步骤将使用R语言进行语料库的抽样统计并计算TTR：针对每种语料库，分别以5000为样本量大小抽样1000次，计算得到1000该语料库的个TTR值。

$$\mathrm{TTR} = \frac{\text{Amount of unique words}}{\text{Amount of total words}}$$

```
for (i in seq(1,1000)) {
    token = sample(GENIA, 5000)
    token_num = length(token)
    type_num = length(token[!duplicated(token)])
    TTR_GENIA = append(TTR_GENIA, type_num/token_num)
} # 抽样计算GENIA语料库的TTR
for (i in seq(1,1000)) {
    token = sample(AGAC, 5000)
    token_num = length(token)
    type_num = length(token[!duplicated(token)])
    TTR_AGAC = append(TTR_AGAC, type_num/token_num)
} # 抽样计算AGAC语料库的TTR
```

4. 假设检验

首先对两种语料库的TTR样本进行正态性检验，验证其是否服从正态分布。若服从正态分布，即可通过t检验两种语料库在TTR这一指标上是否具有显著差异。

```
shapiro.test(TTR_GENIA)
shapiro.test(TTR_AGAC) # 正态性检验
t.test(TTR_AGAC, TTR_GENIA) # t检验
```

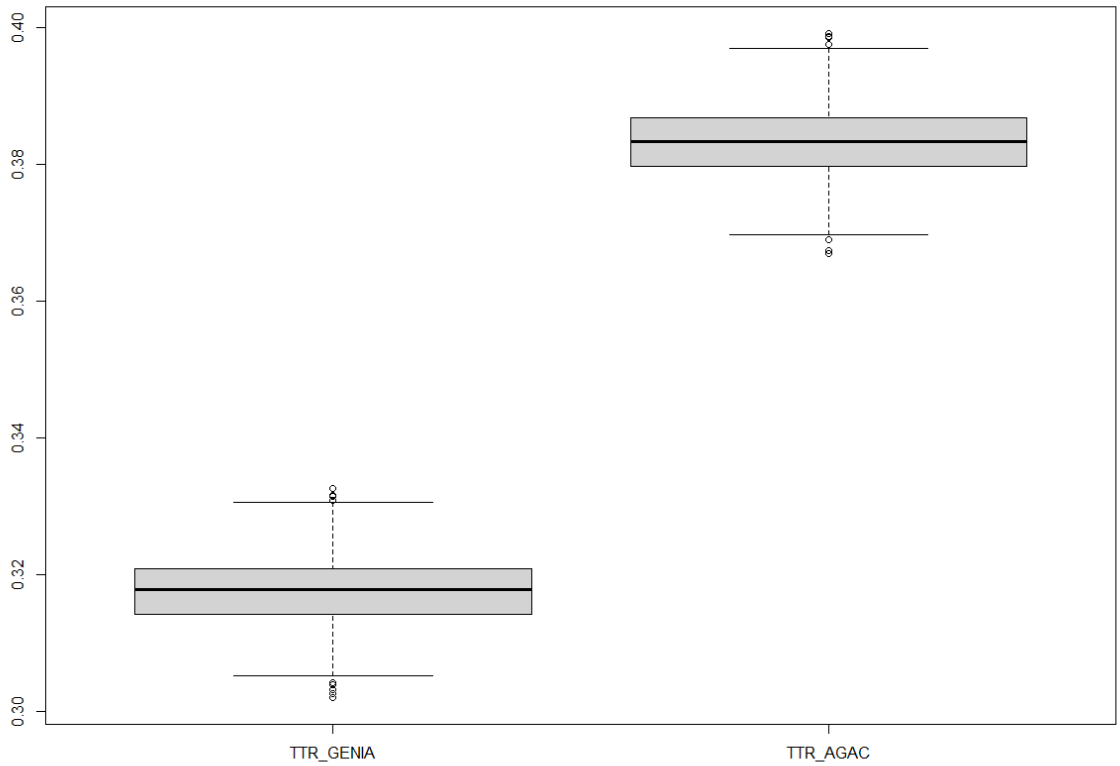
实验结果

1. 语料库的规模与基本TTR信息

GENIA语料库的TTR均值为0.3177，AGAC语料库的TTR均值为0.3834。GENIA语料库的规模为AGAC语料库的近十倍。

2. TTR样本分布情况

通过对TTR样本绘制箱线图来大致了解TTR的大小在两种语料库中的分布情况。



3. TTR样本正态性

通过正态性检验，两种语料库抽样得到的TTR样本均服从正态分布（GENIA: p = 0.7232, AGAC: p = 0.5912）,可以进行t检验。

```
        Shapiro-Wilk normality test

data:  TTR_GENIA
W = 0.99875, p-value = 0.7232


        Shapiro-Wilk normality test

data:  TTR_AGAC
W = 0.99856, p-value = 0.5912
```

2. 假设检验

通过t检验证明，AGAC语料库的TTR值显著大于GENIA语料库（p < 0.05）。

```
> t.test(TTR_AGAC, TTR_GENIA)

        Welch Two Sample t-test

data:  TTR_AGAC and TTR_GENIA
t = 286.27, df = 1997.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06525566 0.06615594
sample estimates:
mean of x mean of y
0.3834032 0.3176974
```

讨论

由于GENIA语料库与AGAC语料库的规模不同，所以进行TTR比较时采取了抽样统计的方法。经过t检验，AGAC语料库的TTR值显著高于GENIA语料库，这表明AGAC语料库在语义丰富度上显著高于GENIA语料库。然而，由于GENIA语料库的规模庞大，是AGAC语料库的近十倍，GENIA语料库在信息量上还是多于AGAC语料库。

GENIA语料库是涵盖了广泛的生物医学领域中的分子生物学信息，而AGAC语料库侧重于捕获在致病性背景下突变基因的功能变化，其更加专注于基因、突变相关的信息。通过TTR分析以及语料库规模分析，可以认为AGAC作为一个规模小于GENIA的语料库，其所专注的基因、突变等方面的相关信息的丰富度是高于GENIA的，更加适合对关于致病性背景下突变基因的功能变化相关研究提供有力的参考于帮助。