

实验二：绘制比较语料库 GENIA 和 AGAC 的词云

2021 年 5 月 8 日

学 校:	华中农业大学
学院班级:	信息学院生信 1801 班
姓 名:	邓启东
学 号:	2018317220103
指导教师:	夏静波

目录

1	实验目的	3
2	实验材料	3
3	实验步骤	3
3.0.1	GENIA 预料库过大内存不够的问题及解决	3
3.1	词汇频率的统计以及 wordcloud 词云绘制	3
3.1.1	GENIA 预料库过大内存不够的问题及解决	3
3.1.2	词频统计和词云绘制	3
3.1.3	词云结果分析	4
4	实验总结	5

1 实验目的

在之前的作业一中我们已经衡量了两个语料库的 TTR (类符/形符比), 并且分析了两个语料库的 TTR 衡量可能存在的问题。并且在考虑后引入了 STTR (标准化类符/形符比), 排除了语料库长度造成的影响, 两个语料库得到了近乎一致的结果。本次实验我们会对语料库中出现的词语的词频进行一个统计, 并且采用词云的方式可视化呈现。

2 实验材料

R 语言的 `tm_map` 文本挖掘包

3 实验步骤

之前的 TTR 和 STTR 的计算都是为了观察语料库的复杂程度, 也就是词语不重复的程度。为了直观地观察两个语料库重复的词语的程度, 我们使用 R 语言的 `tm_map` 这个文本挖掘包对语料库中出现的词语进行统计。

3.0.1 GENIA 预料库过大内存不够的问题及解决

在统计过程中, 由于本地电脑的内存有限, 而 GENIA 太大, 电脑再将变量 `dtm` 转换成矩阵的时候无法分配大小为 36.7 Gb 的矢量。因此我们书写代码 (见代码 `get_GENIA.py`) 将 GENIA 截取前 55,414 行, 和 AGAC 的行数保持一致 (保存为 `get_part_of_GENIA.txt` 文件)。这种方式其实有一个问题就是不是随机抽取行, 可能会因为文献顺序问题影响结果, 不能反映 GENIA 全文的词频。不过随机抽行也很简单, 这里知识以截取前 55,414 行作为示例进行分析。

3.1 词汇频率的统计以及 wordcloud 词云绘制

之前的 TTR 和 STTR 的计算都是为了观察语料库的复杂程度, 也就是词语不重复的程度。为了直观地观察两个语料库重复的词语的程度, 我们使用 R 语言的 `tm_map` 这个文本挖掘包对语料库中出现的词语进行统计。

3.1.1 GENIA 预料库过大内存不够的问题及解决

在统计过程中, 由于本地电脑的内存有限, 而 GENIA 太大, 电脑再将变量 `dtm` 转换成矩阵的时候无法分配大小为 36.7 Gb 的矢量。因此我们书写代码 (见代码 `get_GENIA.py`) 将 GENIA 截取前 55,414 行, 和 AGAC 的行数保持一致 (保存为 `get_part_of_GENIA.txt` 文件)。这种方式其实有一个问题就是不是随机抽取行, 可能会因为文献顺序问题影响结果, 不能反映 GENIA 全文的词频。不过随机抽行也很简单, 这里知识以截取前 55,414 行作为示例进行分析。

3.1.2 词频统计和词云绘制

之后过小写转换、去掉数字、去除停用词 (the/of/a 之类的无实义的词语)、去除标点、空格之后进行词语频率的统计。分别得到词频如表1所示以及词云 (见图 1、2)。

