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					
2	实验	材料		3		
3 实验步骤		步骤		3		
			GENIA 预料库过大内存不够的问题及解决			
	3.1	词汇频	阿率的统计以及 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	立哈	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汇频率进行一个统计,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)。

GENIA		AGAC	
word	freq	word	freq
cells	673	mutations	373
kappa	453	function	363
cell	442	mutation	260
expression	340	loss	213
binding	305	gene	209
gene	298	cells	187
alpha	298	expression	173
protein	285	mutant	168
activation	258	protein	167
transcription	244	variants	167

表 1: 两语料库出现频率较高的前十个词汇极其出现次数

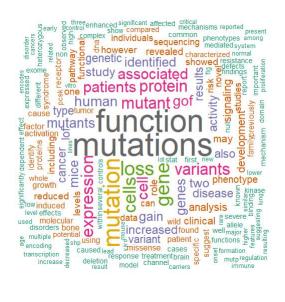


图 2: AGAC 词云

3.1.3 词云结果分析

我们可以从高频词表和词云图看出,两个语料库的高频词汇有所区别。很明显,GENIA 语料库中出现的高频词有"cell"、"gene"、"transcription"、"protein"、"human"等。与它本身的三个医学主题词: human、blood cells 以及 transcription factors 有着密切的关系。而出现频率最高的"cells"这个词也体现其主要为分子生物学水平的特点。

其中出现的高频词与"kappa"经过原始摘要检查发现实际指的是"NF-kappa-B",即一种核因子蛋白,一个转录因子蛋白家族。在几乎所有的动物细胞中都能发现 NF-kB,它们参与细胞对外界刺

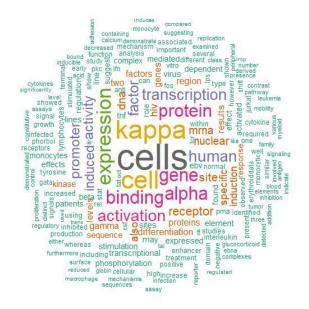


图 1: GENIA 词云

激的响应,如细胞因子、辐射、重金属、病毒等。在细胞的炎症反应、免疫应答等过程中 NF-kB 起到关键性作用。NF-kB 的错误调节会引发自身免疫病、慢性炎症以及很多癌症。NF-kB 也与突触的可塑性、记忆有关。

而 AGAC 语料库中出现的高频词是"mutations"、"mutations"、"function"、"loss"、"gene"等词语。可以明显看出其是关于突变基因和功能变化的语料库。和我们的预期一致。"loss"、"gof"等词也是后续我们实体识别中需要贴上标签的词汇。

4 实验总结