

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
1.1	前期工作	3
1.2	本次实验	3
2	实验材料	3
2.1	实验所需包及语料库	3
2.2	词云图介绍	3
2.3	语料库介绍	3
2.3.1	GENIA 语料库	3
2.3.2	AGAC 语料库	4
3	实验步骤	4
3.1	语料库处理	4
3.1.1	GENIA 语料库过大问题及解决	4
3.2	词频统计及绘图	4
3.3	词云结果分析	5
3.4	存在的问题	6
3.5	改进: python nltk 包词干提取	6
3.5.1	三种词干提取器对比	6
3.5.2	图形绘制	7
4	实验总结	8

1 实验目的

1.1 前期工作

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。

1.2 本次实验

本次实验我们会对语料库中出现的词语的词汇频率进行统计，并且采用词云的方式可视化呈现。我们首先简单的使用了已有的 R 代码绘制得到了一个简单的词云。然后分析其结果存在的词语重复问题，采用 python 的 jieba 包进行分词，nltk 包进行词干提取，然后用 Python 的 wordcloud 包进行更加丰富的词云图片绘制。

2 实验材料

2.1 实验所需包及语料库

上半部分：

R 语言的 tm_map 文本挖掘包、wordcloud 词云绘制包、GENIA 语料库、AGAC 语料库 (均处理成单行词语)。

下半部分：

python 的 jieba 包进行分词，nltk 包进行词干提取，然后用 Python 的 wordcloud 包进行更加丰富的词云图片绘制。

2.2 词云图介绍

wordcloud，也叫词云图，也叫文字云，是对文本中出现频率较高的“关键词”予以视觉化的展现，词云图过滤掉大量的低频低质的文本信息，使得浏览者只要一眼扫过文本就可领略文本的主旨。

2.3 语料库介绍

2.3.1 GENIA 语料库

GENIA 语料库是一个生物医学文献集合。这个语料库是为了发展和评估分子生物学信息检索及文本挖掘系统而创建的。GENIA corpus version 3.0. 包含 2000 篇来自 MEDLINE 数据库的摘要。这些摘要是由 PubMed 按照 human、blood cells 以及 transcription factors 三个医学主题词 (medical subject heading terms) 为搜索条件搜索到的。这个语料库已经被按照不同级别的语言信息、语义信息进行标注。

其包含：

1. 词性标注 2. 短语结构句法注释 4. 事件注释 5. 关系注释 6. 指代消解注释 (跨句结构里面，代词到底指的是什么) ([GENIA 下载链接](#))

2.3.2 AGAC 语料库

AGAC (Annotation of Genes with Alteration-Centric function changes) 是一个活性基因人类专家注释语料库，目的是捕捉突变基因在致病环境中的功能变化。可以用于药物再利用的案例研究，揭示了变异与广泛的人类疾病之间的潜在关联。AGAC 注释了 11 种命名实体类型。通过识别出 LOF/GOF 分类的基因-疾病关联，结合 LOF-激动剂/ GOF-拮抗剂假说可以应用于疾病候选和物质的知识片段的发现。[\(AGAC 下载链接\)](#)

3 实验步骤

3.1 语料库处理

3.1.1 GENIA 语料库过大问题及解决

在统计过程中，由于本地电脑的内存有限，而 GENIA 语料库太大，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。

3.2 词频统计及绘图

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。

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)。

前文提到, GENIA 语料库本身是在 Pubmed 上根据的三个医学主题词: human、blood cells 以及 transcription factors 得到的 2,000 篇摘要。这里的词云高频词和检索使用的主题词呈现出相似性。

而出现频率最高的“cells”这个词我们理解为这些文献的生物系统尺度主要是细胞层面的。实际上, cells 作为一个常见词, 出现的频率最高也是可以理解的。

对于其中出现的一个出现频次也很多的词“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。

AGAC 词云分析:

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。

3.4 存在的问题

首先, R 语言处理 GENIA 语料库的时候会有内存问题, 只能处理语料库的一部分。

其次, 我们看到绘制出的词云存在无法区分“mutations”, “mutant”及“mutation”的问题。除此以外的例子还有“cells”和“cell”, “genes”和“gene”等等。即单复数, 词性无法区分, 导致两个词语的语义相近或一致却分别计数。要想解决这个问题, 则需要 nltk 包的词干提取器进行词干提取。

3.5 改进: python nltk 包词干提取

我们将原语料库的文本进行以下处理: 1. 把标点符号 (除“-”’) 都替换成空格; 2. 多个空格替换成一个空格; 3. 单词转换成小写; 4. 数字替换成“NBR”。得到单个单个的小写词语语料库。分别保存为 GENIA.txt 和 AGAC.txt。这样其实已经按照句子拆分开了, 不再需要 `tokens = tk.word_tokenize(text)` 函数进行按句拆分。发现处理后 GENIA 语料库共有 407,120 个词。

3.5.1 三种词干提取器对比

我们分别使用 nltk 包的三种词干提取器, 并进行每个词语的词干抽提结果打印, 其部分结果如下 (见图2):

