**一、文献挖掘目标**

通过文本挖掘的方式寻找出不同有性反转现象的物种（表1）中被研究和报道最多的与性反转相关的基因、药物，并以词云图（图1）的形式展示。

**表1 有性反转现象的物种**

|  |  |  |
| --- | --- | --- |
| Number | Species | Common Name |
| 1 | *Betta splendens* | Siamese fighting fish |
| 2 | *Bos taurus* | Cattle |
| 3 | *Cynoglossus semilaevis* | Tongue sole |
| 4 | *Cyprinus carpio haematopterus* | Amur carp |
| 5 | *Danio rerio* | Zebrafish |
| 6 | *Epinephelus coioides* | Orange-spotted grouper |
| 7 | *Equus caballus* | Horse |
| 8 | *Gallus gallus* | Chicken |
| 9 | *Homo sapiens* | Human |
| 10 | *Lates calcarifer* | Barramundi perch |
| 11 | *Monopterus albus* | Swamp eel |
| 12 | *Mus musculus* | House mouse |
| 13 | *Oreochromis niloticus* | Nile tilapia |
| 14 | *Oryzias latipes* | Japanese medaka |
| 15 | *Paralichthys olivaceus* | Japanese flounder |
| 16 | *Thalassoma bifasciatum* | Bluehead |
| 17 | *Trachemys scripta* | Red-eared slider turtle |
| 18 | *Xenopus laevis* | African clawed frog |



**图1 词云图示例**

**二、文献挖掘步骤**

1. 通过关键字在PubMed中查找文献

针对一个物种，使用该物种的拉丁名、别名、通用名等能够表示该物种的词汇，加上“sex reversal”关键字，在PubMed中查找相关的文献（使用特定的python package提供的接口），得到PubMed ID以及文献的其它信息（标题，期刊，日期，作者等），将这些信息以适当的形式保存在本地，以备下一步使用。分析不同的关键字搜索的结果是否有重合，对查找结果和重合情况要有一个统计。

1. 获取文献摘要信息

利用上一步得到的PubMed ID，组合成一定模式的URL（例如ID为32328408的文献的URL为https://pubmed.ncbi.nlm.nih.gov/32328408/），使用爬虫技术模拟浏览器访问该URL，返回的是PubMed中关于该文献的简单描述（HTML格式），这些描述是有一定规律的。研究这些规律，获取这些文献的标题、作者、期刊、DOI号和摘要，将这些信息以适当的形式保存到本地。

1. 获取文献正文信息

利用上一步得到的DOI号，组成一定模式的URL（例如，DOI为10.1002/advs.201900997的文献的URL为https://doi.org/10.1002/advs.201900997），使用爬虫技术模拟浏览器访问该URL，将可以借助doi.org重定向到文献对应期刊的页面中，借助武大的校园网IP可以直接访问文献正文，使用爬虫技术解析HTML，从上至下遍历每个元素的内容（例如“<p>This is a case.</p>”中的“This is a case.”），有一些元素会导致换行，有一些不会，例如span元素，将这些内容文本抓取下来保存到本地。如果可以，掌握一些规律，尽可能避免抓取一些非正文的信息，例如期刊版面固有的一些信息。

1. 解析文献

针对一篇文献的全文，过滤掉一些特殊的符号（小括号、小括号、逗号、问号等）后，将全文拆分成一个一个单词，统计一篇文章中的所有单词；构建一个黑名单，用于过滤掉一些词汇，尽可能让剩下的词汇是基因、药物、激素，这个构建过程是随着不断地分析文献全文不断地扩增，不断完善黑名单；这个黑名单列表可以是单词的一部分，例如，gene这个字符串，可以过滤掉单词中含有这四个字符的都过滤掉（GENE，genes，Genes），无视大小写；统计所有文章中出现的非黑名单词汇（基因、激素、药物），分析他们出现的频率，绘画成词云。

1. 数据回收

对于一些由2个或者以上单词组成的药物、激素（基因一般都是只有一个词汇），在设置黑名单的时候保留它们（额外设置一个白名单？），然后再分析这些词组出现的频率。

1. 不同物种有一个词云图，所有物种合并起来有一个词云图。

**三、可能遇到的问题**

1. 抓取速度过慢

使用python的线程池技术（threading模块）同时执行多个任务，对于访问超时的任务，设置重试的间隔和次数，如果最终都失败了，要记录下来；开发一套可以继续执行的程序，例如，有很多个任务，每执行完一个任务，都对该任务打上标记，如果出现意外或者任务执行到一半没成功，在下一次程序启动时，只执行那些没有被打上标记的任务（类似snakemake的机制）。

1. 单词太多

常出现的单词大概几千个，很多都是同一个单词的变形，掌握响应的规律，添加合适的黑名单项目，一般每执行完一篇文章都会添加几个黑名单项目，在开始的时候需要人为监督，执行了一部分解析任务后，逐渐达到能够只保留基因、激素、药物的词汇。

1. 不同期刊的正文形式不同

无论正文形式如何，都是分散在不同的元素当中，遍历所有元素的内容，拼接在一起即可，最终只保留文本，不要图片。

提示：在执行的过程中多思考，多总结，针对没有预料到的情况，要么调整策略，要么略过，不要为了解决一个小问题浪费大量时间。所有中间结果都要保留下来，要整理好整个项目的文件结构，设计好中间结果的保存形式和保存位置，特别是保存形式，要通俗易懂，否则过几天自己看起来都费劲。