Covid-19科学文献知识发现

余思克[1]

[1]华中农业大学信息学院，430070，武汉，湖北，中国

摘要

新型冠状病毒（Covid-19）是一种由严重急性呼吸道综合征冠状病毒2型引发的传染病，该病毒自2019年末出现以来，由于高传染性和高隐蔽性，很快在全球范围大规模爆发并急剧扩散，成为了人类历史上致死人数最多的流行病之一[1]。疫情爆发一年多以来，有关Covid-19的研究成果在科研论文数量上已经具备了一定的规模。由于人工阅读文献获取最新研究成果的效率低下，针对探索Covid-19的入侵机理、了解基因突变与进化情况、研发治疗药物、研制疫苗等多种具有时效性的科学研究[2, 3]，使用新的技术快速获取论文中的研究成果信息这一需求极为迫切[4, 5]。因此，我们使用了生物自然语言处理领域的实体识别工具PubTator，结合自己搭建的生物信息学分析流程，对PubMed收录的122230篇Covid-19相关研究文献进行文本数据挖掘，围绕基因、突变、化合物、物种、疾病等实体进行知识发现。通过研究，我们发现了Covid-19是一种严重的呼吸道传染病，趋向于侵染哺乳动物，有着致命性，其侵染通常伴随着感冒，咳嗽，与中风，急性肾损伤，糖尿病有着密切的联系。诸如“ACE2”，“spike”，“IL-6”等来源于人类或Covid-19基因组中的基因与Covid-19的侵染可能有着密切的联系。此外，我们还发现许多与Covid19相关的基因与化合物之间有着潜在的互作关联，并构建了关联网络。通过生物医学文本挖掘与知识发现的手段，我们基于大量文献数据得出了许多与Covid19有关的信息，这些信息对于后续疫苗的研发，病毒作用机理，药物研制等研究提供了一定的参考。

关键词: Covid-19，生物医学，实体识别，文本挖掘与知识发现

1 课题概况

生物医学自然语言处理是一门利用NLP的自动化方法和手段解决生物信息领域的一些数据挖掘问题并进行知识发现的学科。由于自己在生物信息学本科阶段的学习中发现自己的兴趣更倾向于计算方面，希望以后从事人工智能在生物医学中的应用等研究，因此选修了该课程，选课的目的是为了学习NLP理论与技术，学习相关的机器学习与深度学习模型，开拓视野，为将来的科研与工作打下基础。

新型冠状病毒（Covid-19）疫情自2019年爆发以来，对各国产生了严重的影响，且仍未得到有效控制。已发表的相关研究文献有数十万篇，人工阅读文献归纳研究成果是低效且不现实的，因此，我选择了本课题，目的是使用生物医学自然语言处理的PubTator工具，结合自己搭建的分析流程，高效挖掘文献摘要中的实体信息并进行知识发现，为Covid-19的研究提供一定的参考。本项目计划挖掘出与Covid-19相关的高频基因、化合物、人类基因组和Covid-19基因组的突变情况、Covid-19侵染的物种以及Covid-19感染造成的症状；了解Covid-19相关高频基因的功能以及Covid-19相关突变对Covid-19侵染造成的影响；了解基因与化合物的关联并构建关联网络。

2 数据

本项目的实验数据来自于PubTator，该工具整合了PubMed文献数据库中收录的文献的摘要中包含的实体信息[6]。我们通过edirect工具获取了PubMed文献数据库中所有与”Covid-19“关键词有关的文献的uid，并通过调用PubTator的API接口，根据文献的uid，下载了这些文献的实体识别信息。获取的文献摘要实体

数据由三个部分组成，第一个部分是文献的标题，第二个部分是文献的摘要，第三个部分是文献中识别出来的实体。实体内容包括基因、化合物、DNA变异、蛋白质变异、SNP、物种、疾病七个种类。实体数据来源于122230篇Covid-19相关文献的摘要。本项目的研究将重点围绕实体数据中的摘要部分以及实体部分进行展开。

3 研究方法

3.1 研究方法的算法背景，与其他方法的联系与区别

本项目的实体识别过程使用的是PubTator工具。PubTator是一款基于Web，可通过使用高级文本挖掘技术来加快人工文献的管理（例如，注释生物实体及其关系）的工具[7]，作为一个多合一的系统，PubTator提供了一站式服务来注释PubMed引用[8]。与其他方法相比，PubTator提供了已经完成识别的实体数据，不需要自己完成搭建框架、训练模型、调参、然后进行实体识别等一系列过程。虽然PubTator针对实体识别的效果相较于BERT+CRF等神经网络训练出来的模型来说不占优势，但是其提供的一站式多合一服务，极大地简化了大家进行实体识别任务的操作过程，大大缩短了时间，适合快速展开研究[9]。

本项目的数据处理部分使用到了Linux Shell，R语言、Python语言以及相关的拓展包，相较于其他语言，Linux Shell可以方便地利用正则表达式快速进行数据的初步筛选，R语言以及Python语言具有强大的拓展包生态、绘图能力、数据处理能力以及简洁的语法，使得数据处理过程方便且迅速。

3.2 研究方法中的核心思路

本项目首先通过调用PubTator的API，获取了122230篇Covid-19相关文献的实体数据，并围绕基因、化合物、DNA变异、蛋白质变异、SNP、物种、疾病实体进行词频的统计，目的是分析哪些实体词汇在文献中得到了大量的报道，从而了解与Covid-19具有潜在关联的各类实体信息。然后，我们对高频基因等实体展开了进一步的探索，通过数据库了解基因功能，从而推测它们与Covid-19的入侵机制、作用机理等特征。

此外，我们还关注了基因与化合物之间的间接联系，通过对122230篇Covid-19相关文献的摘要数据，对基因与化合物实体进行共句分析，分析思路是：当某一个基因与一个化合物同时出现在了一句话中，即认为它们之间具有潜在的关联。通过编写脚本分析基因与化合物的关联并绘制关联网络，直观反映Covid-19相关的基因与化合物之间的互作关联（图 *1*）。

3.3 本文的方法部分与课堂讲授内容的联系和区别与补充

本文的方法部分与课堂相同点在于都使用了PubTator工具进行实体数据的获取。不同点在于实体数据的获取细节。课堂中通过文献搜索结果，手动复制几篇感兴趣的文献对应的uid，然后通过调用API下载实体数据，而由于本项目计划对所有的Covid-19相关文献进行实体数据的获取，一共有122230篇文献对应122230个实体数据集合需要被下载，很明显不能继续通过手动搜索并复制粘贴uid进行实体数据的下载了，于是本项目使用了edirect工具首先根据关键词自动化提取了所有Covid-19相关文献的uid，然后遍历读取uid列表文件调用PubTator API进行实体数据的下载，获得了大量丰富的实体数据。

除了课堂讲授的PubTator实体抽取基础分析，我们补充进行了各实体词频统计、基因功能分析、基因与化合物工具关联及其关联网络绘制等分析方法辅助于Covid-19文献的知识挖掘（图 *1*）。

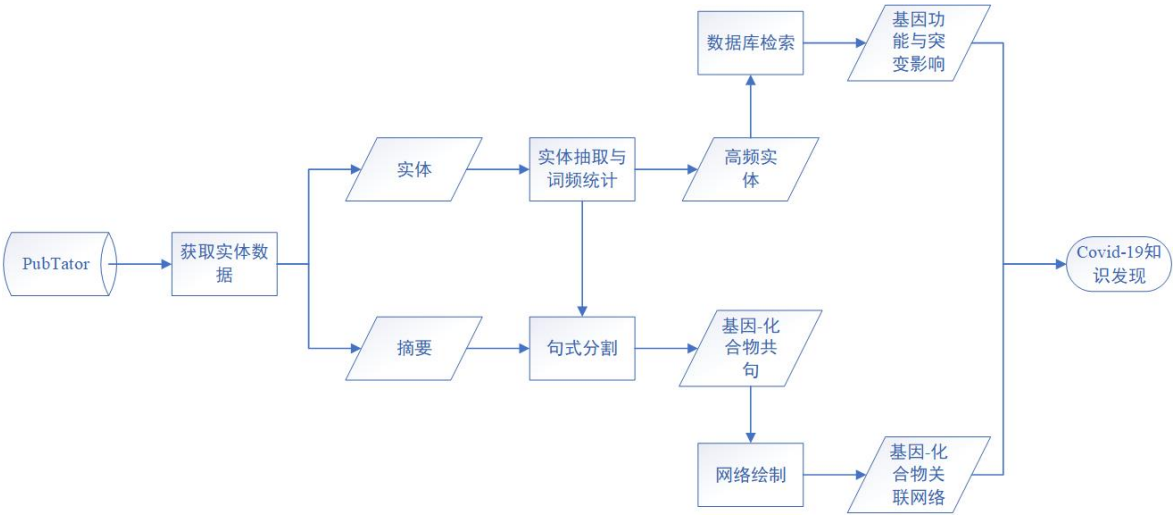

图 1: 项目流程图。图片来源：自绘，工具：Visio

4 算法实践和代码编写要求

4.1 任务描述

第一步通过编写Shell脚本进行数据的获取。此步骤将搜索与Covid-19相关的122230篇文献中uid，并通过遍历这些uid依次从PubTator中获取实体信息。由于相关的工具是linux环境下的，所以使用Shell脚本能很好地完成此任务。

第二步通过编写Shell脚本进行数据的预处理。此步骤将通过Shell中的正则表达式把实体数据分为两个子数据，一个是实体数据，另一个是摘要数据，供后续数据分析使用。由于正则表达式是Linux Shell中强大的字符处理工具，故使用Shell编写脚本可以很好地完成此任务。

第三步通过R语言编写脚本读取实体数据，按照不同的实体种类把实体数据拆分开，统计实体词汇的词频并排序，然后根据词频分别绘制不同实体的词云。由于R语言强大的绘图能力以及统计能力，其非常适用于该统计分析及绘图任务。

第四步通过Python语言编写脚本读取上一步R语言处理过后的基因实体以及化合物实体数据，同时读取摘要数据，通过判断基因与化合物是否同时出现在一句话，确定基因与化合物的潜在关联，并格式化输出关联文件，后续将依据此文件绘制关联网络。由于Python丰富的内置函数以及其拓展包Pandas，其能够很好地完成遍历、判断算法以进行共句分析及格式化输出。

4.2 实验设计

数据的获取是通过shell中的edirect工具中的esearch命令完成，它将获取所有与Covid-19相关文献的uid。然后使用shell中的curl命令，调用PubTator的API接口，依次下载这些文件。数据的预处理是通过使用正则表达式结合grep与sed命令，分别进行实体与摘要的筛选，得到待分析的数据。

实体数据的分析是通过R语言中的数据框结合table，sort等函数进行词频统计，然后通过wordcloud2函数进行词云的绘制。共句关联分析是通过python中的pandas包结合基本函数实现。

代码在运行之前需要安装相关的依赖包，R语言需要安装wordcloud2包，Python需要安装Pandas包，且需要关注文件地址信息，根据自己的数据存放地址调整代码中的读写地址。本项目使用的R为4.0.3版本，Python为3.9.5版本。

4.3 关键代码

```r
#代码来源：https://github.com/kiekie233/BioNLP-course/blob/main/script/Get_entity_freq.R
#词云绘制（以基因实体词频大于一百的词汇为例）
gene=freq[freq$bioconcep=="Gene"&freq$Freq>=100,]#筛选高频词汇
order_temp = order(gene$Freq,decreasing = T)
gene=gene[order_temp,]#排序
rownames(gene)=seq(1,nrow(gene))
gene$Freq=gene$Freq^0.4
wordcloud2(gene,size=0.6)#调整参数绘制词云
```

```python
#代码来源：https://github.com/kiekie233/BioNLP-course/blob/main/script/Get_entity_relation.py
#共句判断算法（以每一句话为单位，判断基因与化合物在摘要中的共句信息）
for gene in gene_list[0]:
    for chemical in chemical_list[0]:
        print("Gene:",gene,"chemical:",chemical)
        for sentence in sentences:
            if gene in sentence and chemical in sentence:
                relation=relation.append({"gene":gene,"relation":"relate","chemical":chemical
                    },ignore_index=True)#计算基因与化合物的共句信
                    息
```

5 主要的生物信息学实验和实验结论

5.1 实体词频统计

通过对各实体词频的统计，我们发现物种实体显著富集于“人类”，“猪”，“老鼠”，“猫”，“哺乳动物”，“鸡”等词汇上，可以从中发现这些物种可能是实验验证或者是病例观察得出的Covid-19主要侵染的物种。可以推测，Covid-19偏向于侵染哺乳动物。疾病实体富集于“侵染”，“呼吸道”，“致命”，“死亡”，“糖尿病”，“感冒”，“疼痛”，“气喘”，“急性肾损伤”，“中风”，“神经”等词汇上，可以发现Covid-19是一种传染性疾病，其具有致命性，其感染伴随着感冒，疼痛，气喘等症状，可能对肾脏，神经系统等组织与器官造成伤害，且糖尿病与其关联密切。基因实体显著富集于“ACE2”，“spike”，“IL-6”等词汇上，可以推测这些基因与Covid-19有着密切的联系，可能是Covid-19病毒的组成结构基因，也可能是辅助其侵染宿主的关键功能基因。突变、化合物等实体词汇的聚集同样反映出了许多与Covid-19有关的信息。在此，我们展示出基因的词频可视化信息，完整的结果可在本项目的GitHub仓库中查看并获取。（图 2）

图 2: 基因词频词云。图片来源：自绘，工具：R

5.2 基因功能分析

通过对基因实体的词频统计，我们得到了一组频率较高的基因，针对这批基因，我们通过NCBI数据库对其进行了检索，得到了这些基因的具体信息。（表 *1*）

这些基因中一部分来源于人类基因组，一部分来源于Covid-19病毒基因组，来源于人类基因组的基因所编码的蛋白，很可能作为媒介协助Covid-19进行侵染，或者是作为Covid-19的攻击目标靶点；而来源于病毒基因组的基因，很可能是病毒的结构基因，其编码的蛋白具有一些特征可供识别，也可能是功能基因，辅助病毒入侵宿主。后续可以针对这些基因进行深入研究，为疫苗的研发，病毒作用机理，药物研制提供一定的参考。

表 1: 部分高频基因功能信息

基因	功能	物种
ACE2	血管紧张素转换酶2	人类
spike(S)	刺突糖蛋白	Covid-19
IL-6	白介素6	人类
CRP	C反应蛋白	人类
TMPRSS2	跨膜丝氨酸蛋白酶2	人类
Mpro	ORF1a多蛋白	Covid-19
CD4	CD4抗原	人类
CD8	CD8抗原	人类
N	核衣壳蛋白	Covid-19

5.3 基因与化合物互作分析

通过对基因与化合物在摘要中的共句分析，我们筛选出了共句次数超过100次的基因-化合物对，我们认为共句次数超过100次足以证明该基因与化合物之间存在着一定的潜在关联。

其中，“S-CO”，“N-CO”，“ACE2-CO”，“S-iron”，“S-oxygen”，“S-water”，“IL-6-CO”等基因-化合物共句频率非常高，可以推测它们之间有着较大的可能性存在互作关联关系，后续可以针对这些高频率的基因-化合物关联对进行进一步深入的研究以探究Covid-19的特征。（表 *2*）

表 2: 部分高频基因-化合物共句信息

基因	化合物	频率
S	CO	81014
N	CO	48717
ACE2	CO	1289
S	iron	1544
S	oxygen	1194
S	chloroquine	905
S	water	737
IL-6	CO	726
S	hydroxychloroquine	652

这些化合物可能与基因表达的产物会有互作，可能会产生激活、抑制、失活作用。这些关联具有潜在的研究意义，可以为药物靶点筛选，Covid-19病毒入侵机制的研究提供参考。对此，我们绘制了基因-化合物关联网络，直观展示它们的关联信息。（图 *3*）

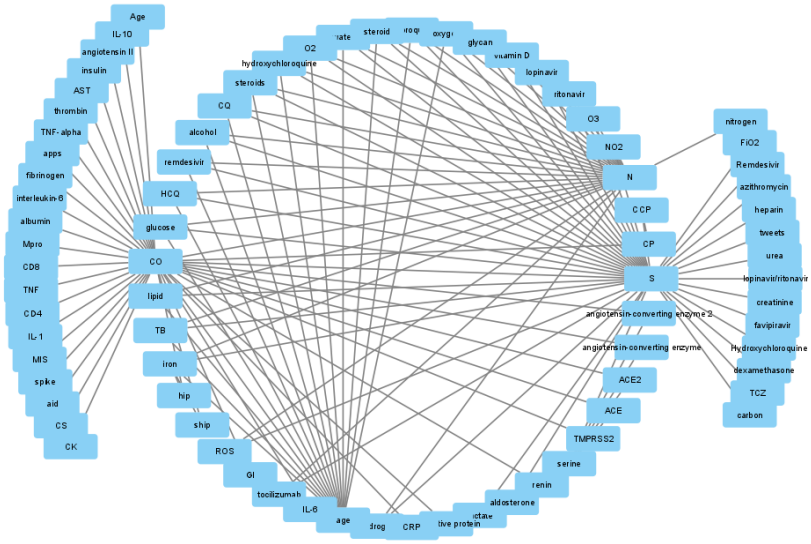

图 3: Covid-19相关基因-化合物关联网络。图片来源：自绘，工具：Cytoscape

6 后记

6.1 课程论文构思和撰写过程

在进行课程论文的构思之前，我首先深度思考了一下该项目的细节信息：这个项目是什么？目的是什么？为了达到目的，需要做些什么？如何展示自己的结果？经过了一段时间的思考，认为这个项目是针对海量的Covid-19论文进行实体抽取，发现其中潜在的，隐含的知识。此项目的目的是为了解决人工阅读文献了解信息费时费力且效率不高这一难题，并基于大量数据挖掘Covid-19的信息，为致病机理，入侵方式，药物靶点，疫苗研发等科学研究提供一定的辅助与参考。为了达到目的，需要尽可能将所有与Covid-19有关的文献全部下载下来，然后进行数据分析，最后把结果充分生动地展示出来。于是基于这些思考，我对论文中所需要呈现的内容以及呈现方法有了大致的思路：将大量的文献摘要、实体信息通过整理、可视化，直观地反映出来，并辅以精简的文字描述，做到言简意赅，图文并茂。

论文的撰写过程还是比较痛苦的，因为我需要做的不仅是把自己的结果照搬上来，我需要把项目的前因后果，实现方式，以及对于结果的讨论和自己观点的提出，通俗易懂且清晰地表达出来，这对我的归纳总结能力以及发散性思维是一个重大的考验。所以在撰写的过程中，我不断地打草稿，修改，修改，再修改，最终尽自己最大努力完成了这篇论文的撰写。

6.2 所参考的主要资源

完整项目储存在个人GitHub中：`https://github.com/kiekie233/BioNLP-course/`。项目代码均为自己编写。GitHub仓库中包含了该项目的思路、所需数据、完整代码、项目的完整结果等信息。

本项目的生物信息学流程、完整的代码及其运行方式、依赖的环境、所需要的包、运行顺序、实现的功能、可能存在的报错等帮助信息均在GitHub中有详细介绍，可以供大家更好地进行本项目分析流程的复现，欢迎大家复现本项目并提出问题、探讨见解、批评指正。

结果文件中除了上文中展示的数据之外，还包含着因篇幅限制而未能展示在文章中的数据。额外的结果包含：所有种类实体的词频信息及其可视化词云数据、基因-化合物共句信息、基因-化合物共句频率信

息等。Github中提供的完整数据可以供大家做额外的、自己感兴趣的拓展分析。

非常感谢老师和师兄师姐的课堂教学以及提供的课程资源，感谢同学们针对于项目思路以及技术实践方面的探讨与帮助，虽然本项目由我个人完成，但老师、师兄师姐以及同学们潜移默化中给了我许多帮助，这门课让我受益颇丰，再次表示感谢。

6.3 代码撰写的构思和体会

在生物信息本科阶段的学习中，我发觉自己对于代码编写有着强烈的热爱。每当需要实现一个功能，在编写代码之前，需要先把这个功能拆分为几个板块，然后分板块实现这些小目标，然后把它们拼凑在一起。例如共句分析的代码构思，我将其分为了四步，第一步是对基因、化合物进行遍历，第二步是对摘要中拆分好的每一句话进行遍历，第三步是将基因、化合与与摘要中的每一句话进行匹配，第四步是结果的输出。

虽然在代码的构思过程中，头脑中思维逻辑的飞快转动使我头痛不已，但是当自己完成了debug，实现了自己预想的功能，顺利跑通代码之后，内心有着无法用言语形容的一种畅快感。尤其在自己的代码绘制出了精美的图片，巧妙完成了各类计算，解决了生物学问题时，我领悟到了代码的魅力，交叉学科的魅力。代码虐我千百遍，我待代码如初恋，不论在编写代码时，遇到了多大的困难，只要逻辑清晰，稳扎稳打，冷静内心，都能写出漂亮的代码（当然也得保持不错的代码风格，方便维护以及共享）。我想，今后我依旧会保持对代码的热爱，因为它真的太神奇了。

6.4 生物信息学实验设计的构思和体会

因为自己在去年疫情期间尝试过针对Covid-19的序列进行一定的分析，以及自己有过一定的生信项目经历，所以自己对Covid-19及其实验设计有着一定的了解。本项目的关键目的是替代人工阅读，从文献摘要中挖掘出与Covid-19有关的知识。所以，在得到实体数据之后，我们首先关注的是实体的频率并根据频率发现一些规律与特征，进而尝试探索不同实体之间存在的潜在关联，深入挖掘实体相关的信息。同时，还要充分地利用得到的觉果，将数据以图片的形式生动地展现出来。

设计生物信息学实验最重要的不是写代码，而是好好地把握住生物问题，首先需要弄懂我们为什么要研究、开展这个项目，这个项目的目的是弄清楚什么科学问题。只有把生物学问题把握住了，才能更好地设计生物信息学实验，将生物学问题转化为概率问题，计算问题，算法问题等，从而对其进行实践，这便是我理解的交叉学科的一个特点：跨学科的科学问题转化以及多种类的技术工具应用。

7 附录

S1. 本项目的GitHub页面. `https://github.com/kiekie233/BioNLP-course`

S2. 2021年BioNLP课程网页. `https://hzaubionlp.com/course-bionlp-and-kd/`

S3. PubTator网页. `https://www.ncbi.nlm.nih.gov/research/pubtator/`

参考文献

[1] Thirumalaisamy P Velavan and Christian G Meyer. The covid-19 epidemic. *Tropical medicine & international health*, 25(3):278, 2020.

[2] T Thanh Le, Zacharias Andreadakis, Arun Kumar, R Gómez Román, Stig Tollefsen, Melanie Saville, Stephen Mayhew, et al. The covid-19 vaccine development landscape. *Nat Rev Drug Discov*, 19(5):305–306, 2020.

[3] Xuetao Cao. Covid-19: immunopathology and its implications for therapy. *Nature reviews immunology*, 20(5):269–270, 2020.

[4] Gemelli Against COVID, Post-Acute Care Study Group, et al. Post-covid-19 global health strategies: the need for an interdisciplinary approach. *Aging Clinical and Experimental Research*, page 1, 2020.

[5] Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. Covid-19 named entity recognition for vietnamese. *arXiv preprint arXiv:2104.03879*, 2021.

[6] Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics*, 32(12):1907–1910, 2016.

[7] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522, 2013.

[8] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593, 2019.

[9] Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, and Zhiyong Lu. Pmc text mining subset in bioc: about three million full-text articles and growing. *Bioinformatics*, 35(18):3533–3535, 2019.