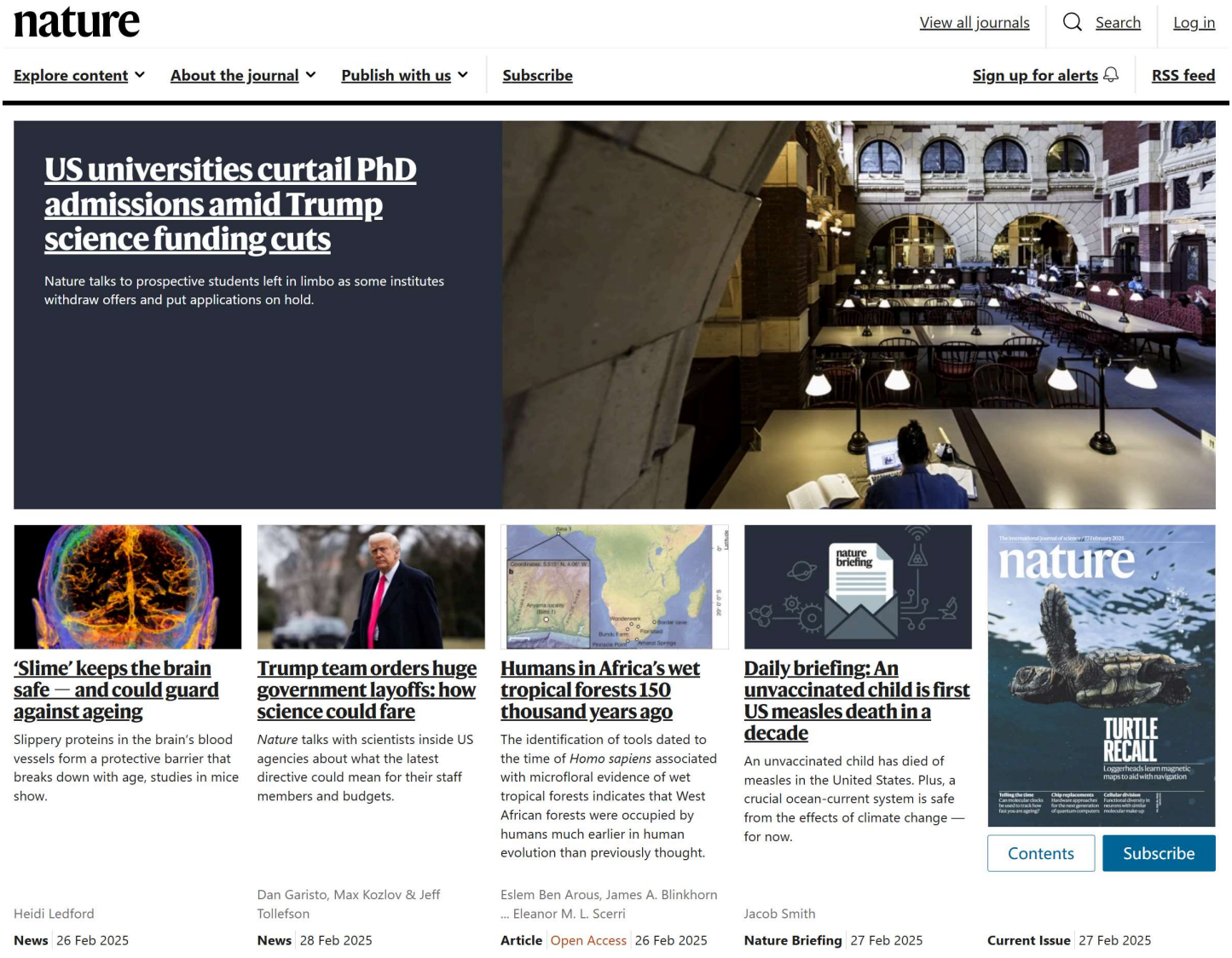
实验二

中国科学技术大学2025年春季学期“数据分析及实践”课程 - 实验二说明文档

任务概述

《自然》（英语：Nature，主页： https://www.nature.com/ ）是世界上历史悠久的、最有名望的科学杂志之一，首版于1869年11月4日。
与当今大多数科学论文杂志（例如《Cell》和《美国数学会杂志》等）专一于一个特殊的领域不同，《自然》是少数依然发表来自很多科学领域的一手研究论文的杂志（其它类似的杂志有《科学》和《美国科学院院刊》等）。在许多科学研究领域中，很多最重要、最前沿的研究结果都是以短讯的形式发表在《自然》上。
另外，Nature旗下子刊，合作期刊众多，是一个科学研究无法绕过的重要的学习平台。



现在我们希望通过Nature平台获取一些我们关心的论文信息，请你按照要求编写Python代码实现任务列表中的内容。

任务列表

1. （20%）进入Nature主页，通过高级检索功能，搜索关键词llm，限制年份为“2023-2024”，搜索得到近两年间关键词含有llm的Nature平台相关文章。

- Q. 使用python代码获取检索结果第一页的html内容并解码为文本串，将其以UTF-8编码格式写入page1.txt文件中，用于后续处理

2. （30%）本页面展示了搜索结果中前几条内容的截图

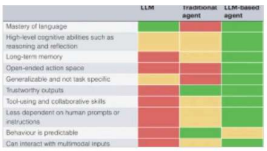
Showing 1–50 of 668 results

Comments & Opinion
05 Dec 2024
Nature Machine Intelligence
Volume: 6, P: 1418-1420

LLM-based agentic systems in medicine and healthcare

Large language model-based agentic systems can process input information, plan and decide, recall and reflect, interact and collaborate, leverage various tools and act. This opens up a wealth of opportunities within medicine and healthcare, ranging from clinical workflow automation to multi-agent-aided diagnosis.

Jianing Qiu, Kyle Lam ... Eric J. Topol

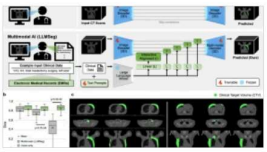


Research
Open Access
24 Oct 2024
Nature Communications
Volume: 15, P: 1-14

LLM-driven multimodal target volume contouring in radiation oncology

The integration of multimodal knowledge would be essential for radiation oncologist to determine the therapeutic treatment. Here, inspired by the large language models facilitating the integration of textural information and images, this group reports a 3D multimodal clinical target volume delineation model combining image and text-based clinical information for decision-making in radiation oncology.

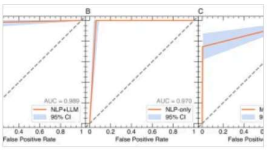
Yujin Oh, Sangjoon Park ... Jong Chul Ye



Research
Open Access
05 Nov 2024
Scientific Reports
Volume: 14, P: 1-7

Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review

Mert Marcel Dagli, Yohannes Ghenbot ... Jang W Yoon



Research
Open Access
23 Nov 2024
Nature Communications
Volume: 15, P: 1-16

An automatic end-to-end chemical synthesis development platform powered by large language models

The rise of large language model (LLM) technology offers new opportunities for advancing chemical synthesis. Here, the authors developed an LLM-based reaction development framework (LLM-RDF) to copilot the design and experimental tasks throughout the end-to-end chemical synthesis development.

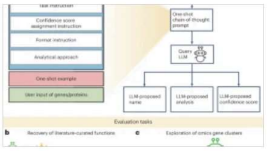
Yixiang Ruan, Chenyin Lu ... Yiming Mo

Research
28 Nov 2024
Nature Methods
Volume: 22, P: 82-91

Evaluation of large language models for discovery of gene set function

Large language models show potential in suggesting common functions for a gene set.

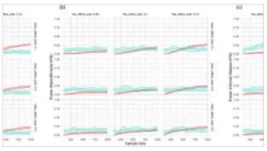
Mengzhou Hu, Sahar Alkhairy ... Dexter Pratt



Research
Open Access
18 Nov 2024
npj Digital Medicine
Volume: 7, P: 1-8

Simulating A/B testing versus SMART designs for LLM-driven patient engagement to close preventive care gaps

Sanjay Basu, Dean Schillinger ... Joseph Rigdon



- Q. 打开page1.txt，观察相关数据的组织格式规律。从这些文本串中提取一些论文相关的重要信息（包括文章标题，文章地址，文章简介，作者列表，文章类型，期刊名称，卷宗/页面信息），并按照发表的期刊名进行分类，以下面的格式存储到一个字典列表中：

```
[
    {
        "journal": "Nature",
        "papers": [
            {
                "title": "‘Fighting fire with fire’ — using LLMs to combat LLM hallucinations",
                "authors": [
                    Karin Verspoor
                ],
                "url": "/articles/d41586-024-01641-0",
                "description": The number of errors produced by an LLM can be reduced by groupi
                "type": "News & Views",
                "volume_page_info": "Volume: 630, P: 569-570"
            },
            ...
        ]
    },
    {
        "journal": "Nature Machine Intelligence",
        "papers": [
            ...
        ]
    }
]
```


Hint: 有些文章欠缺部分内容，如description模块，若无description则对应部分填写"no description"

- Q. 基于上述结果，输出每个期刊在page1中包含的论文数量

3. （30%）下面展示的是某篇文章打开后的界面

- Q. 观察可以发现该文章的作者信息实际上多于搜索结果中展示的内容，请你仔细观察此界面的html数据组织格式，依据此编写python程序，将上一步骤中的字典提取内容中的作者列表中的内容进行替换，替换为文章主页面显示的全部作者

4. （20%）JSON是一种轻量级的数据交互格式，常用于存储和表示数据。

- Q. 将上一步获得的字典列表转化为 json 对象（可使用json包的相应方法），并以2字符缩进的方式写入nature_llm.json文件中。

5. （选做）Nature平台的Advanced Reasearch页面中，还可以依据期刊名称和Volume进行检索

- Q. 对于上述提取到的page1中的文章，请你尝试提取所有发表在以Nature为开头的期刊（不包括Nature Reviews开头的期刊）的文章，请你检索得到这些文章所在的相应Volume中的所有文章信息，提取检索结果的page1中的所有文章，并以如下格式存储在volume.json文件中，重复的Volume只需要提取一次。查找过程需要使用python代码自动实现，不可以手动获取网址。（此处的author信息只使用检索界面的部分信息即可）

```
[
    {
        "journal": "Nature",
        "volume": "630"
        "papers": [
            {
                "title": ,
                "authors": [

                ],
                "url": ,
                "description": ,
                "type": ,
            },
            ...
        ]
    },
    {
        "journal": "Nature Machine Intelligence",
        "volume": "7",
        "papers": [
            ...
        ]
    }
]
```

Hint: 尝试使用高级检索功能后，观察搜索结果对应的网址发现，每个期刊在nature内部的编码中都有一个简写的名称，如期刊"Nature Machine Intelligence"对应的名称为"natmachintell"，这个期刊的简写信息可以从文章的主页中获取（第3步中打开的网页）

格式要求

1. 请按具体任务分步编写代码，存储于.ipynb格式文件中用于复现，并增加注释。
2. 本实验可使用 Python 标准库以及 Requests 库中的所有方法实现，无需局限于任务要求中指定的方法。
3. 实验报告必须涵盖任务列表中的所有内容和相应结果，并请存储于.pdf格式文件中。

提交要求

1. 将实验原代码与实验报告保存至一个压缩包中，命名为学号-姓名-实验二.zip
2. 邮件命名为学号-姓名-实验二，发送至USTC_AD2025@163.com（设有自动回复）
3. 提交截止日期为2025.3.27晚11：59，请注意不要迟交

参考资料

以下资料可能会对你顺利完成实验有所帮助。

1. 使用 Conda 配置虚拟环境与管理安装包：点击这里
2. Conda 轻量级版本 Miniconda 的安装地址：点击这里
3. 在 VSCode 中使用 Jupyter Notebook 进行代码实现：点击这里
4. Python 官方教程（版本：3.10.13）：点击这里