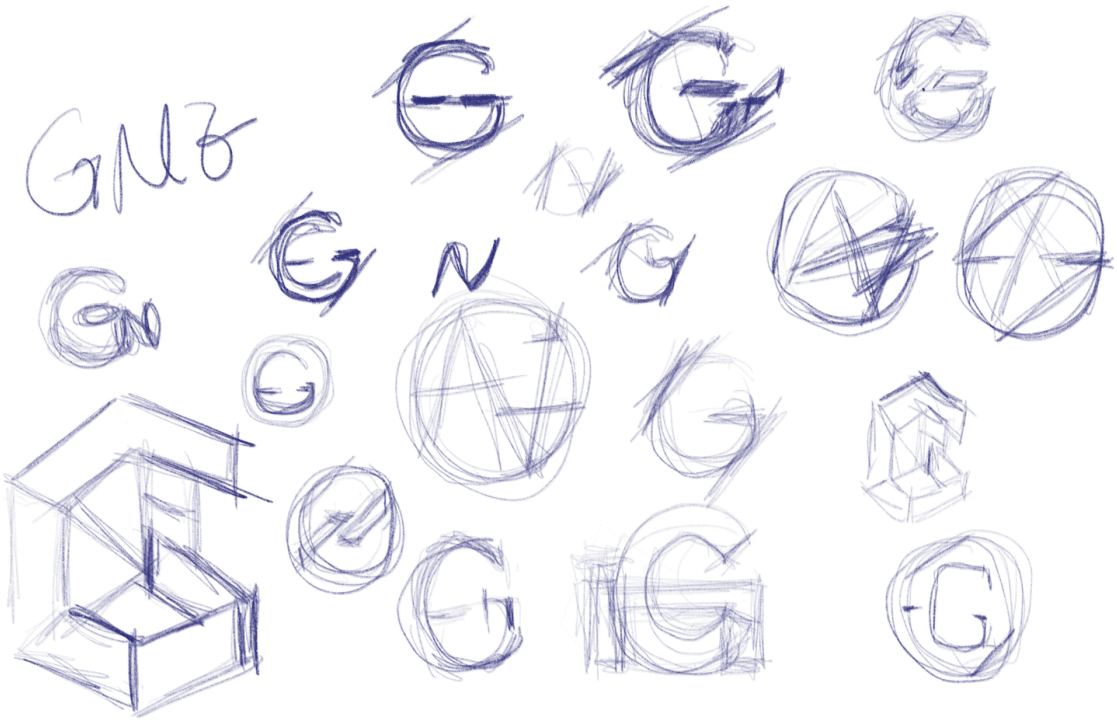
GeneralNewsExtractor

发布 *0.0.6*

2022 年 09 月 25 日

Contents:



GeneralNewsExtractor（GNE）是一个通用新闻网站正文抽取模块，输入一篇新闻网页的 HTML，输出正文内容、标题、作者、发布时间、正文中的图片地址和正文所在的标签源代码。GNE 在提取今日头条、网易新闻、游民星空、观察者网、凤凰网、腾讯新闻、ReadHub、新浪新闻等数百个中文新闻网站上效果非常出色，几乎能够达到 100% 的准确率。

使用方式也非常简单：

```
from gne import GeneralNewsExtractor

extractor = GeneralNewsExtractor()
html = '网站源代码'
result = extractor.extract(html)
print(result)
```

本项目取名为 抽取器，而不是 爬虫，是为了规避不必要的风险，因此，本项目的输入是 HTML 源代码，输出是一个字典。请自行使用恰当的方法获取目标网站的 HTML。

GNE 现在不会，将来也不会提供主动请求网站 HTML 的功能。

CHAPTER 1

如何使用

如果你想体验 GNE 的功能，请按照如下步骤进行：

0. 在线体验

如果你想先体验 GNE 的提取效果，那么你可以访问 http://gne.kingname.info。一般情况下，你只需要把网页粘贴到最上面的多行文本框中，然后点 提取按钮即可。通过附加更多的参数，可以让提取更精确。具体参数的写法与作用，请参阅 API

1. 安装 GNE

```
# 以下两种方案任选一种即可

# 使用 pip 安装
pip install --upgrade gne

# 使用 pipenv 安装
pipenv install gne
```

2. 使用 GNE

```
>>> from gne import GeneralNewsExtractor
>>> html = '''经过渲染的网页 HTML 代码'''
>>> extractor = GeneralNewsExtractor()
>>> result = extractor.extract(html, noise_node_list=['//div[@class="comment-list"]'])
```

(下页继续)

(续上页)

```
>>> print(result)
{"title": "xxxx", "publish_time": "2019-09-10 11:12:13", "author": "yyy", "content":
↪"zzzz", "images": ["/xxx.jpg", "/yyy.png"]}
```

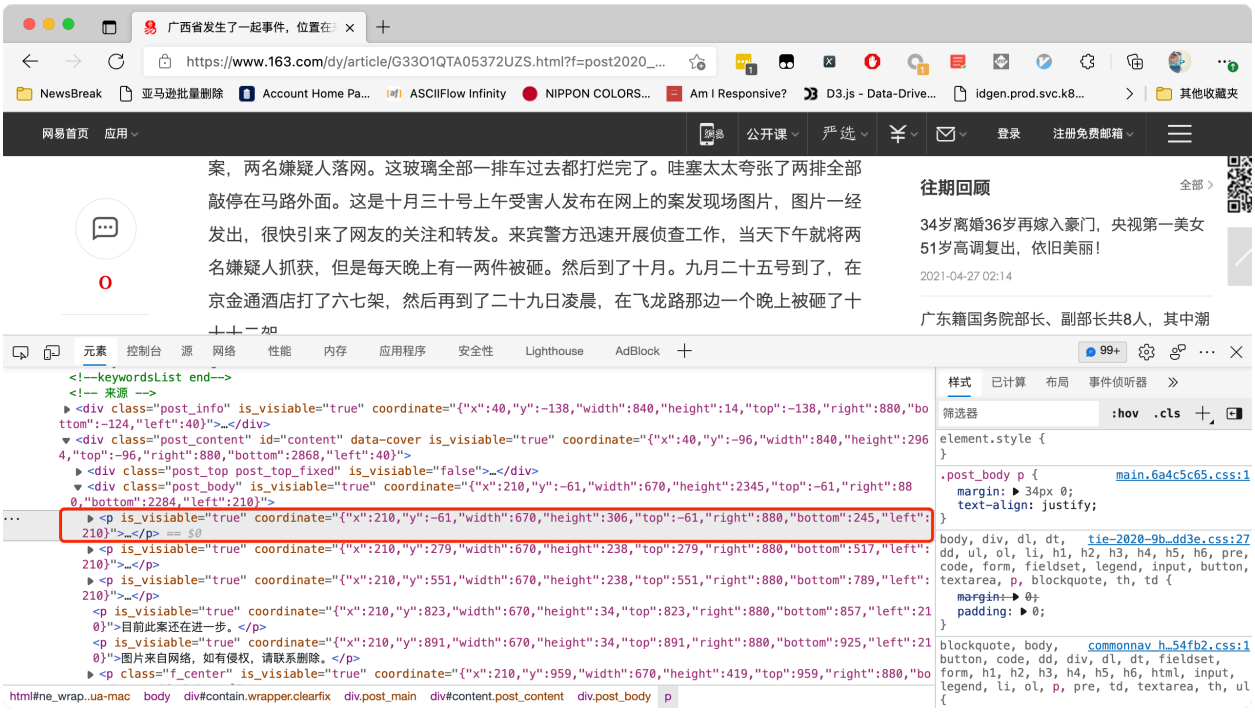
3. 提取列表页

```
>>> from gne import ListPageExtractor
>>> html = '''经过渲染的网页 HTML 代码'''
>>> list_extractor = ListPageExtractor()
>>> result = list_extractor.extract(html, feature='列表中任意元素的 XPath")
>>> print(result)
```

4. 基于可视化区域提高识别准确度（从 gne 0.3.0 版本加入）

请打开文件 `example_visiable.py` ，从这个文件里面，你可以看到 GNE 会从 `visiable_test` 文件夹中读取特殊的 HTML 源代码，并且在调用 `extractor.extract()` 方法的时候，会传入一个参数: `use_visiable_info=True` 。此时，GNE 会基于这些 HTML 中带有的节点坐标信息，更准确地提取正文。

这些特殊的 HTML 主要特点如图所示:



在 `body` 标签下面的所有节点，都有一个属性叫做 `is_visiable` ，它的值是字符串的 `true` 或者 `false` 。如果值为 `true` ，那么，还有一个属性叫做 `coordinate` 。它的值是一个 JSON 字符串，包含了这个节点的尺寸，坐标等信息。

那么，这些特殊的 HTML 是怎么生成的呢？其实只需要在网页上执行这样一段 js 代码就可以了:

```
function insert_visiability_info() {
    function get_body() {
        var body = document.getElementsByTagName('body')[0]
        return body
    }

    function insert_info(element) {
        is_visiable = element.offsetParent !== null
        element.setAttribute('is_visiable', is_visiable)
        if (is_visiable) {
            react = element.getBoundingClientRect()
            coordinate = JSON.stringify(react)
            element.setAttribute('coordinate', coordinate)
        }
    }

    function iter_node(node) {
        children = node.children
        insert_info(node)
        if (children.length !== 0) {
            for(const element of children) {
                iter_node(element)
            }
        }
    }

    function sizes() {
        let contentWidth = [...document.body.children].reduce(
          (a, el) => Math.max(a, el.getBoundingClientRect().right), 0)
          - document.body.getBoundingClientRect().x;

        return {
          windowWidth:  document.documentElement.clientWidth,
          windowHeight: document.documentElement.clientHeight,
          pageWidth:    Math.min(document.body.scrollWidth, contentWidth),
          pageHeight:   document.body.scrollHeight,
          screenWidth:  window.screen.width,
          screenHeight: window.screen.height,
          pageX:        document.body.getBoundingClientRect().x,
          pageY:        document.body.getBoundingClientRect().y,
```

(下页继续)

(续上页)

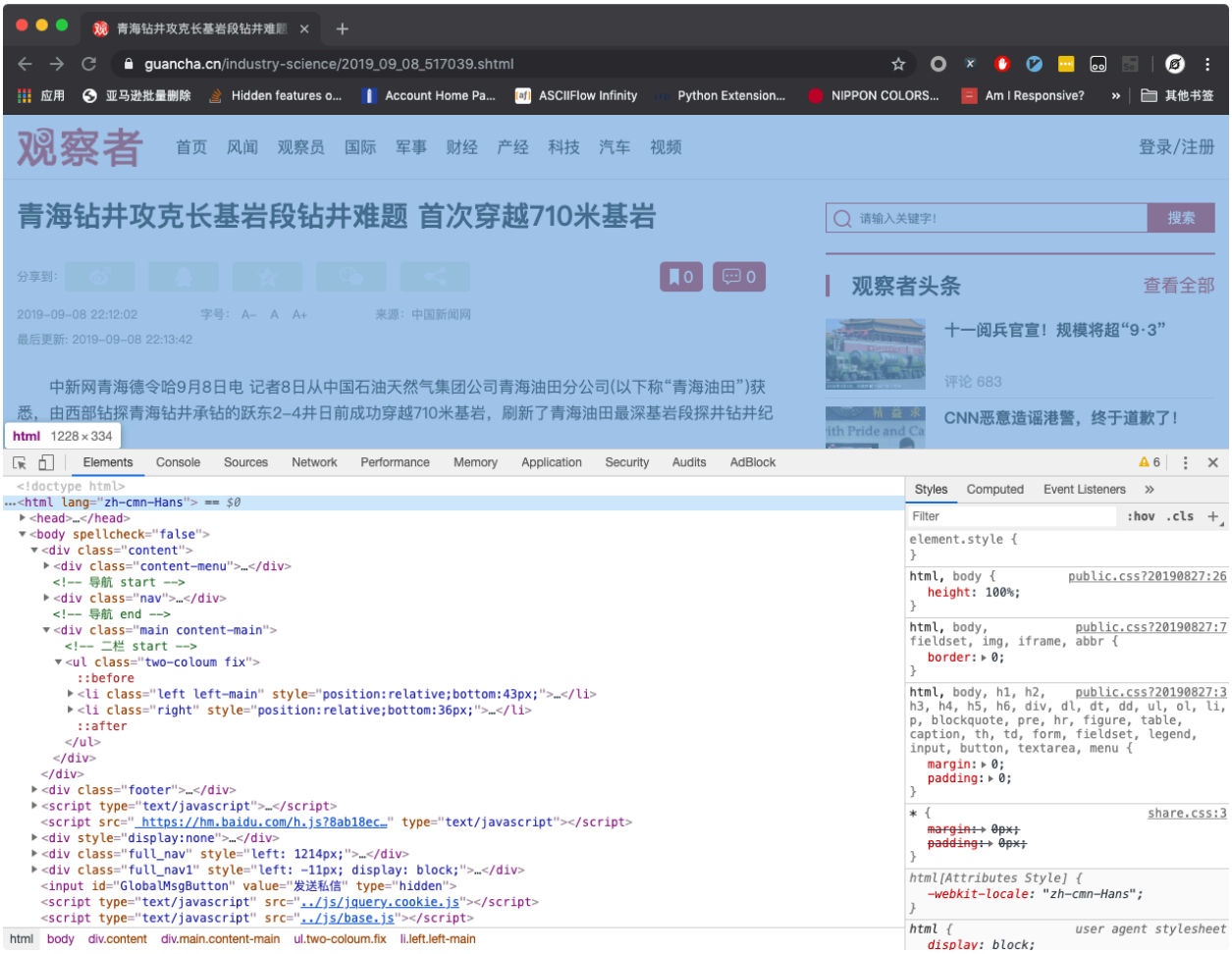
```
            screenX:      -window.screenX,
            screenY:      -window.screenY - (window.outerHeight-window.innerHeight),
        }
    }

    function insert_page_info() {
        page_info = sizes()
        node = document.createElement('meta')
        node.setAttribute('name', 'page_visiability_info')
        node.setAttribute('page_info', JSON.stringify(page_info))
        document.getElementsByTagName('head')[0].appendChild(node)
    }

    insert_page_info()
    body = get_body()
    iter_node(body)
}
insert_visiability_info()
```

我给出了一个使用 Puppeteer 生成这些特殊 HTML 的项目：GneRender 你可以阅读里面的 `render.js` 文件，就可以知道怎么做了。如果你使用的是 Selenium，其实原理是一样的。

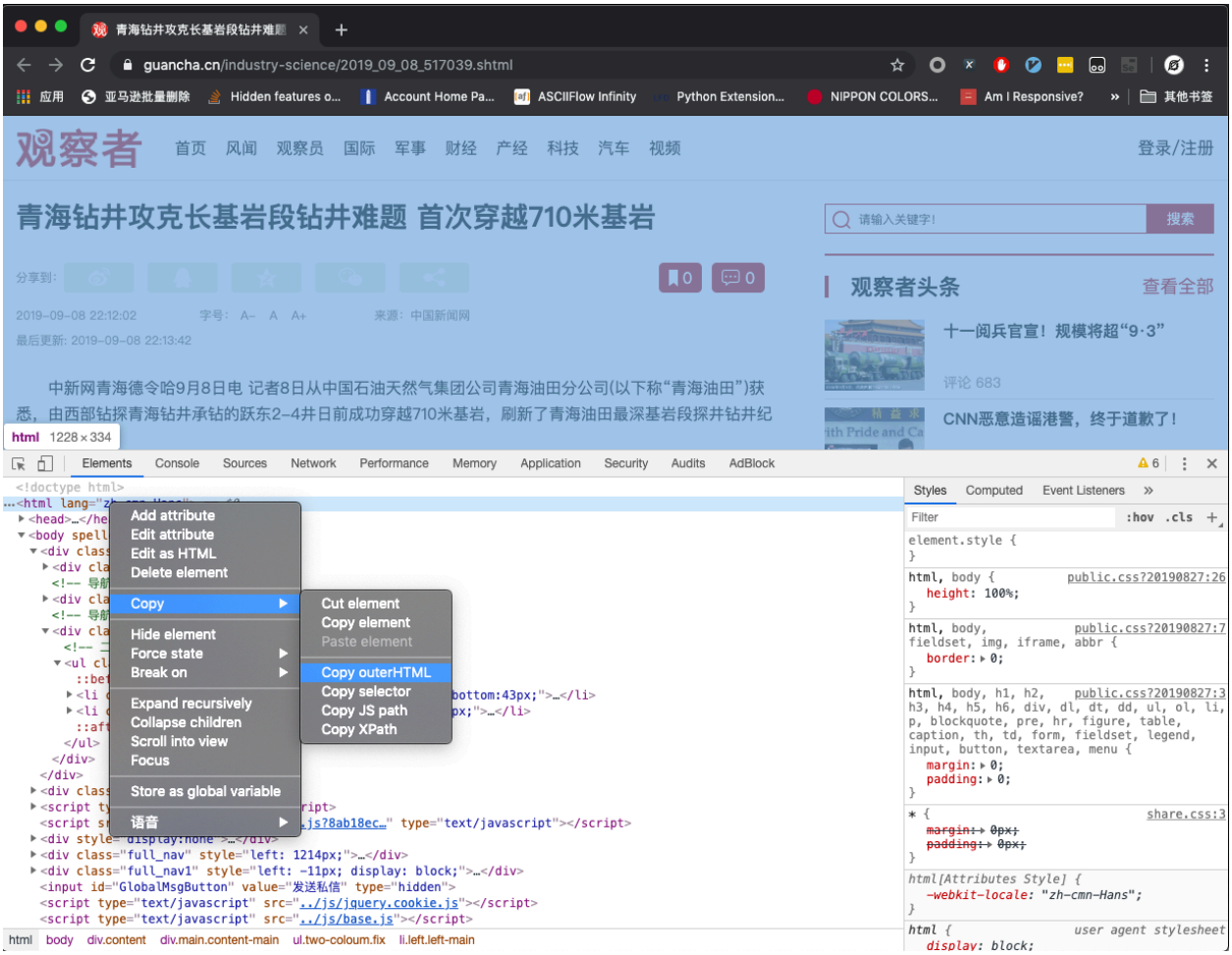
CHAPTER 2

注意事项

- 本项目的输入 HTML 为经过 JavaScript 渲染以后的 HTML，而不是普通的网页源代码。所以无论是后端渲染、Ajax 异步加载都适用于本项目。
- 如果你要手动测试新的目标网站或者目标新闻，那么你可以在 Chrome 浏览器中打开对应页面，然后开启 *开发者工具*，如下图所示：



在 Elements 标签页定位到 <html> 标签，并右键，选择 Copy - Copy OuterHTML ，如下图所示



- 当然，你可以使用 Puppeteer/Pyppeteer、Selenium 或者其他任何方式获取目标页面的 JavaScript 渲染后的源代码。
- 获取到源代码以后，通过如下代码提取信息：

```
from gne import GeneralNewsExtractor

extractor = GeneralNewsExtractor()
html = '你的目标网页正文'
result = extractor.extract(html)
print(result)
```

- 如果标题自动提取失败了，你可以指定 XPath：

```
from gne import GeneralNewsExtractor

extractor = GeneralNewsExtractor()
html = '你的目标网页正文'
```

(下页继续)

(续上页)

```
result = extractor.extract(html, title_xpath='//h5/text()')
print(result)
```

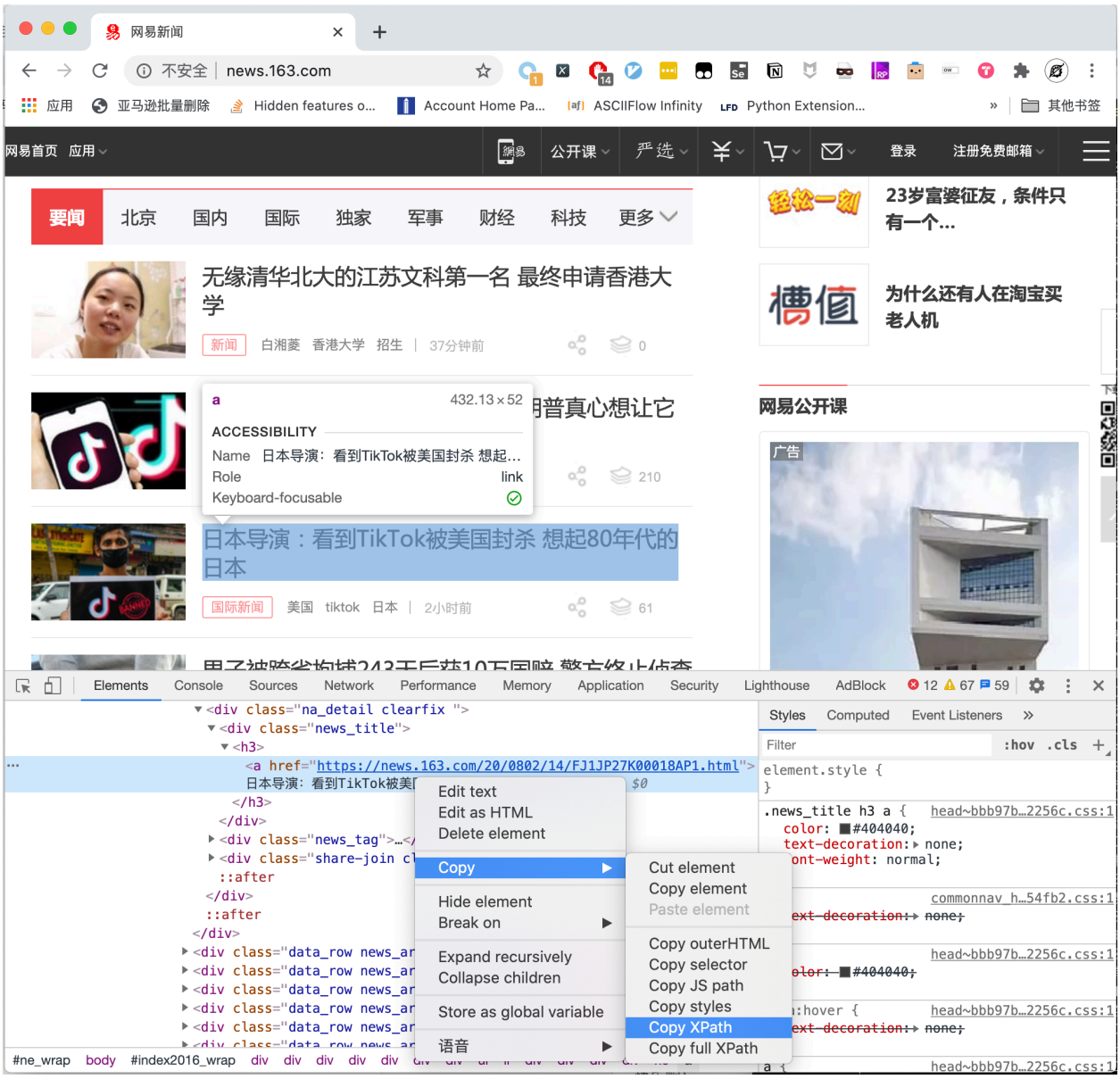
对大多数新闻页面而言，以上的写法就能够解决问题了。

但某些新闻网页下面会有评论，评论里面可能存在长篇大论，它们会看起来比真正的新闻正文更像是正文，因此 `extractor.extract()` 方法还有一个默认参数 `noise_node_list` ，用于在网页预处理时提前把评论区域整个移除。`noise_mode_list` 的值是一个列表，列表里面的每一个元素都是 XPath，对应了你需要提前移除的，可能会导致干扰的目标标签。

例如，观察者网下面的评论区域对应的 Xpath 为 `//div[@class="comment-list"]` 。所以在提取观察者网时，为了防止评论干扰，就可以加上这个参数：

```
result = extractor.extract(html, noise_node_list=['//div[@class="comment-list"]'])
```

- 提取新闻列表页的功能是测试功能，请勿用于生产环境。你可以通过 Chrome 浏览器开发者工具中的 `Copy XPath` 来复制列表中任意一项的 XPath，如下图所示。

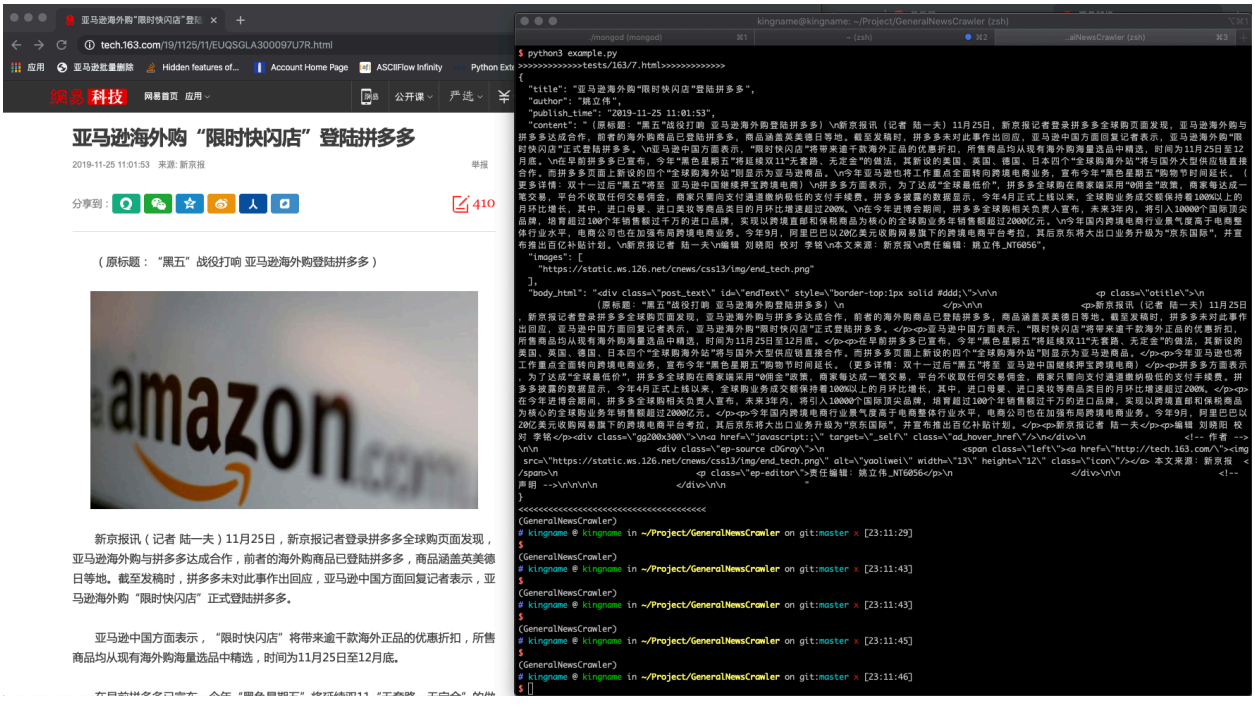


GNE 会根据这一项的 XPath，自动找到这个列表里面其他行的数据。

CHAPTER 3

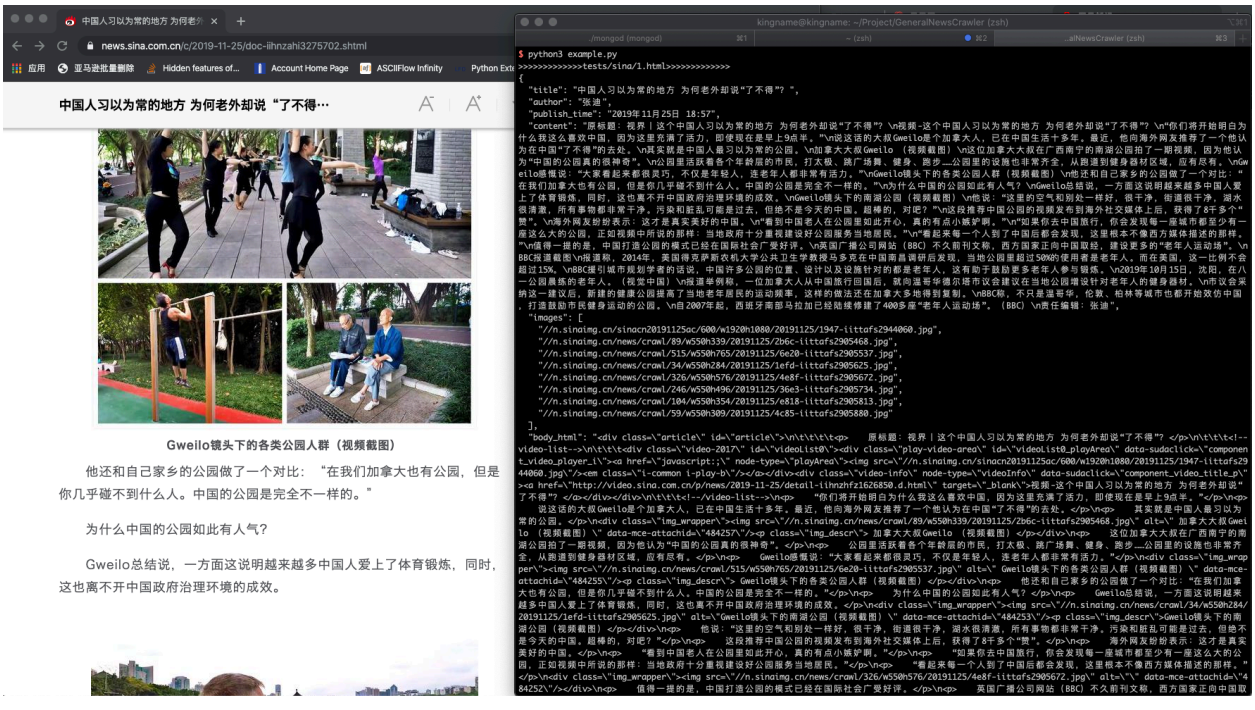
运行截图

3.1 网易新闻

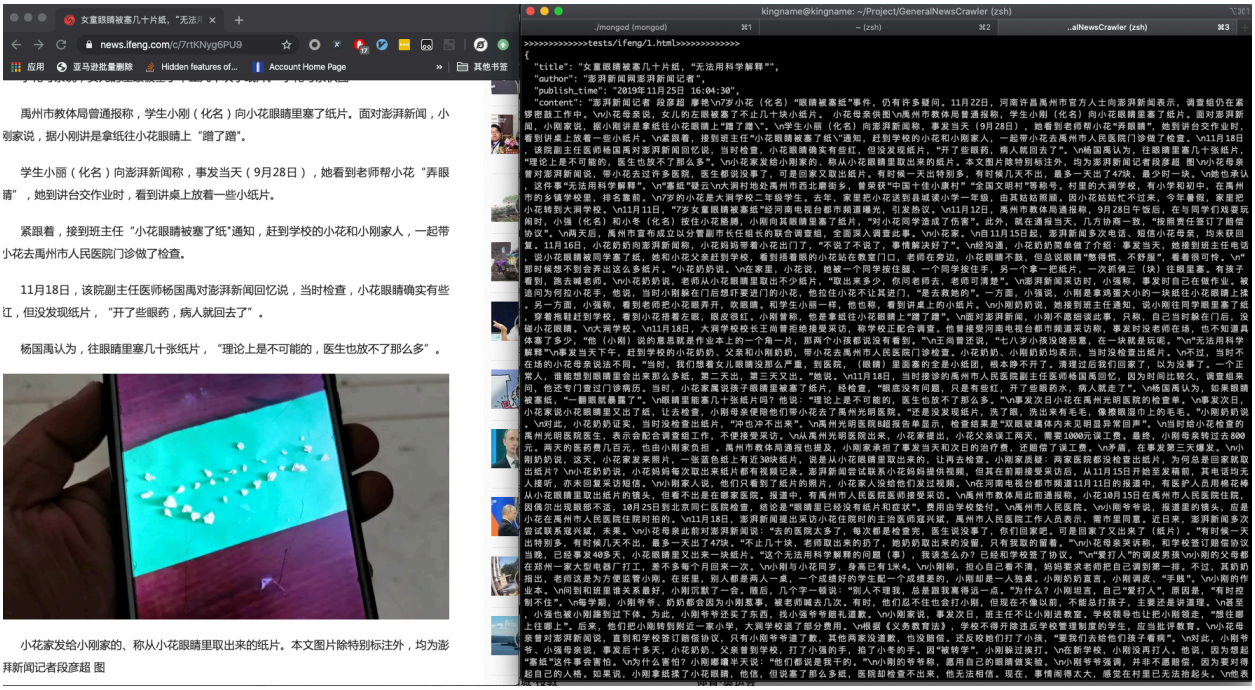


3.2 今日头条

3.3 新浪新闻



3.4 凤凰网



禹州市教体局曾通报称，学生小刚（化名）向小花眼睛里塞了纸片。面对澎湃新闻，小刚家说，据小刚讲是拿纸往小花眼睛上“蹭了蹭”。

学生小丽（化名）向澎湃新闻称，事发当天（9月28日），她看到老师帮小花“弄眼睛”，她到讲台交作业时，看到讲桌上放着一些小纸片。

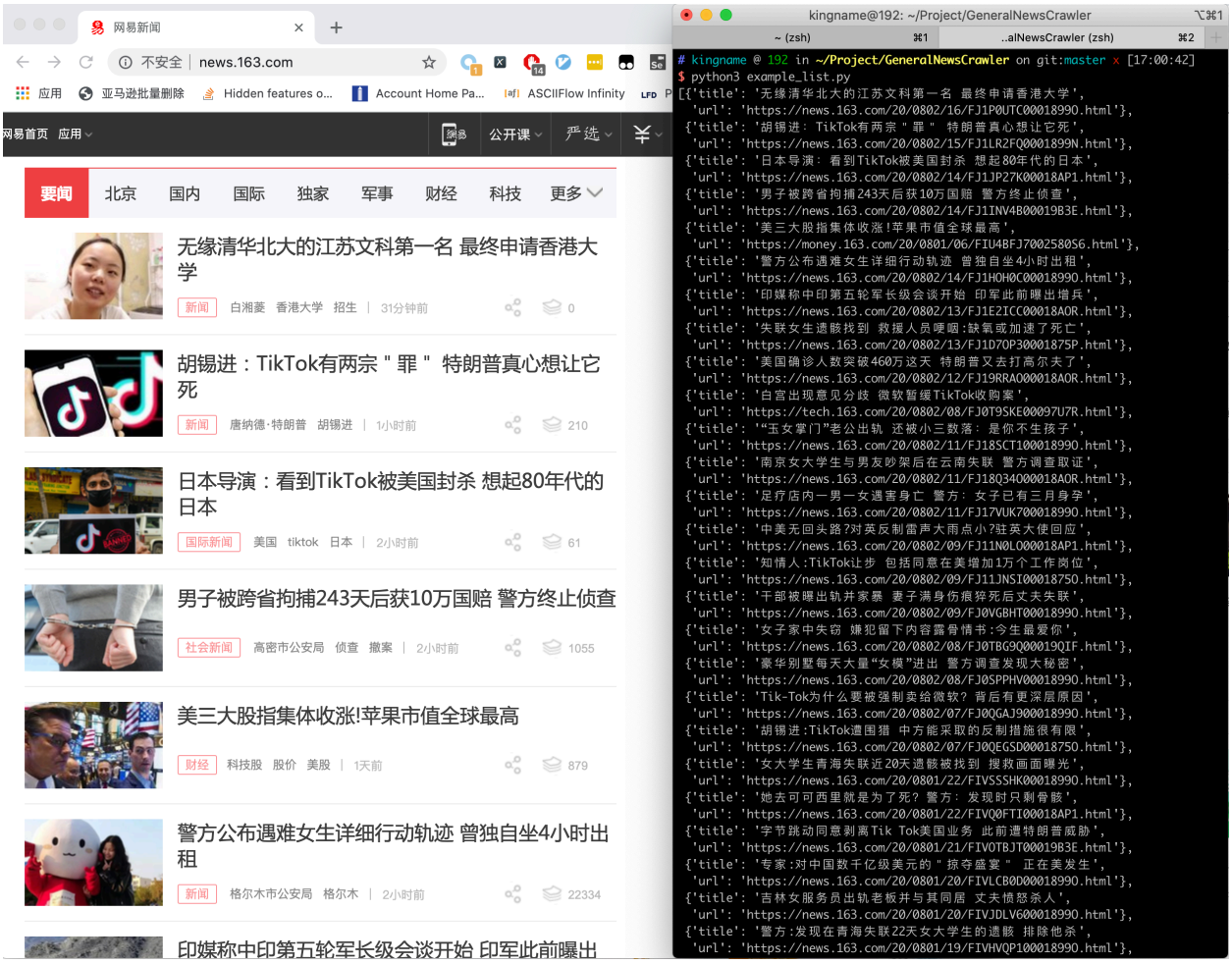
紧跟着，接到班主任“小花眼睛被塞了纸”通知，赶到学校的小花和小刚家人，一起带小花去禹州市人民医院门诊做了检查。

11月18日，该院副主任医师杨国禹对澎湃新闻回忆说，当时检查，小花眼睛确实有些红，但没发现纸片，“开了些眼药，病人就回去了”。

杨国禹认为，往眼睛里塞几十张纸片，“理论上是不可能的，医生也放不了那么多”。

小花家发给小刚家的、称从小花眼睛里取出来的纸片。本文图片除特别标注外，均为澎拜新闻记者段彦超 图

3.5 网易新闻首页列表


网易新闻
不安全 news.163.com
网易首页 应用
公开课
严选
要闻 北京 国内 国际 独家 军事 财经 科技 更多
无缘清华北大的江苏文科第一名 最终申请香港大学
新闻 白湘菱 香港大学 招生 | 31分钟前
胡锡进：TikTok有两宗＂罪＂ 特朗普真心想让它死
新闻 唐纳德·特朗普 胡锡进 | 1小时前
日本导演：看到TikTok被美国封杀 想起80年代的日本
国际新闻 美国 tiktok 日本 | 2小时前
男子被跨省拘捕243天后获10万国赔 警方终止侦查
社会新闻 高密市公安局 侦查 撤案 | 2小时前
美三大股指集体收涨!苹果市值全球最高
财经 科技股 股价 美股 | 1天前
警方公布遇难女生详细行动轨迹 曾独自坐4小时出租
新闻 格尔木市公安局 格尔木 | 2小时前
印媒称中印第五轮军长级会谈开始 印军此前曝出
kingname@192: ~/Project/GeneralNewsCrawler
# kingname @ 192 in ~/Project/GeneralNewsCrawler on git:master x [17:00:42]
$ python3 example_list.py
[{'title': '无缘清华北大的江苏文科第一名 最终申请香港大学',
'url': 'https://news.163.com/20/0802/16/FJ1P0UTC00018990.html'},
{'title': '胡锡进：TikTok有两宗＂罪＂ 特朗普真心想让它死',
'url': 'https://news.163.com/20/0802/15/FJ1LR2FQ0001899N.html'},
{'title': '日本导演：看到TikTok被美国封杀 想起80年代的日本',
'url': 'https://news.163.com/20/0802/14/FJ1JP27K00018AP1.html'},
{'title': '男子被跨省拘捕243天后获10万国赔 警方终止侦查',
'url': 'https://news.163.com/20/0802/14/FJ1INV4B00019B3E.html'},
{'title': '美三大股指集体收涨!苹果市值全球最高',
'url': 'https://money.163.com/20/0801/06/FIU4BFJ7002580SG.html'},
{'title': '警方公布遇难女生详细行动轨迹 曾独自坐4小时出租',
'url': 'https://news.163.com/20/0802/14/FJ1HOH0C00018990.html'},
{'title': '印媒称中印第五轮军长级会谈开始 印军此前曝出增兵',
'url': 'https://news.163.com/20/0802/13/FJ1E2ICC00018AOR.html'},
{'title': '失联女生遗骸找到 救援人员哽咽:缺氧或加速了死亡',
'url': 'https://news.163.com/20/0802/13/FJ1D7OP30001875P.html'},
{'title': '美国确诊人数突破460万这天 特朗普又去打高尔夫了',
'url': 'https://news.163.com/20/0802/12/FJ19RRAO00018AOR.html'},
{'title': '白宫出现意见分歧 微软暂缓TikTok收购案',
'url': 'https://tech.163.com/20/0802/08/FJ0T9SKE00097U7R.html'},
{'title': '"玉女掌门"老公出轨 还被小三数落：是你不生孩子',
'url': 'https://news.163.com/20/0802/11/FJ18SCT100018990.html'},
{'title': '南京女大学生与男友吵架后在云南失联 警方调查取证',
'url': 'https://news.163.com/20/0802/11/FJ18Q34O00018AOR.html'},
{'title': '足疗店内一男一女遇害身亡 警方：女子已有三月身孕',
'url': 'https://news.163.com/20/0802/11/FJ17VUK700018990.html'},
{'title': '中美无回头路?对英反制雷声大雨点小?驻英大使回应',
'url': 'https://news.163.com/20/0802/09/FJ11N0LO00018AP1.html'},
{'title': '知情人:TikTok让步 包括同意在美增加1万个工作岗位',
'url': 'https://news.163.com/20/0802/09/FJ11JNSI00018750.html'},
{'title': '干部被曝出轨并家暴 妻子满身伤痕猝死后丈夫失联',
'url': 'https://news.163.com/20/0802/09/FJ0VGBHT00018990.html'},
{'title': '女子家中失窃 嫌犯留下内容露骨情书:今生最爱你',
'url': 'https://news.163.com/20/0802/08/FJ0TBG9Q00019QIF.html'},
{'title': '豪华别墅每天大量"女模"进出 警方调查发现大秘密',
'url': 'https://news.163.com/20/0802/08/FJ0SPPHV00018990.html'},
{'title': 'Tik-Tok为什么要被强制卖给微软？背后有更深层原因',
'url': 'https://news.163.com/20/0802/07/FJ0QGAJ900018990.html'},
{'title': '胡锡进:TikTok遭围猎 中方能采取的反制措施很有限',
'url': 'https://news.163.com/20/0802/07/FJ0QEGSD00018750.html'},
{'title': '女大学生青海失联近20天遗骸被找到 搜救画面曝光',
'url': 'https://news.163.com/20/0801/22/FIVSSSHK00018990.html'},
{'title': '她去可可西里就是为了死？警方：发现时只剩骨骸',
'url': 'https://news.163.com/20/0801/22/FIVQ0FTI00018AP1.html'},
{'title': '字节跳动同意剥离Tik Tok美国业务 此前遭特朗普威胁',
'url': 'https://news.163.com/20/0801/21/FIVOTBJT0001983E.html'},
{'title': '专家:对中国数千亿级美元的＂掠夺盛宴＂ 正在美发生',
'url': 'https://news.163.com/20/0801/20/FIVLCB0D00018990.html'},
{'title': '吉林女服务员出轨老板并与其同居 丈夫愤怒杀人',
'url': 'https://news.163.com/20/0801/20/FIVJDLV600018990.html'},
{'title': '警方:发现在青海失联22天女大学生的遗骸 排除他杀',
'url': 'https://news.163.com/20/0801/19/FIVHVQP100018990.html'},

CHAPTER 4

API

GeneralNewsExtractor 的函数原型为：

```
class GeneralNewsExtractor:
    def extract(self,
                html,
                title_xpath='',
                host='',
                author_xpath='',
                publish_time_xpath='',
                body_xpath='',
                noise_node_list=None,
                with_body_html=False,
                use_visiable_info=False)
```

各个参数的意义如下：

- **html(str)**：必填，目标网站的源代码
- **title_xpath(str)**：可选，新闻标题的 XPath，用于定向提取标题
- **host(str)**：可选，图片所在的域名，例如 `https://www.kingname.info`，那么，当 GNE 从新闻网站提取到图片的相对连接 "/images/123.png"时，会把 `host` 拼接上去，变成 "https://www.kingname.info/images/123.png"
- **body_xpath(str)**：可选，新闻正文所在的标签的 XPath，用于缩小提取正文的范围，降低噪音

- **noise_node_list(List[str])**: 可选，一个包含 XPath 的列表。这个列表中的 XPath 对应的标签，会在预处理时被直接删除掉，从而避免他们影响新闻正文的提取
- **with_body_html(bool)**: 可选，默认为 False，此时，返回的提取结果不含新闻正文所在标签的 HTML 源代码。当把它设置为 True 时，返回的结果会包含字段 `body_html`，内容是新闻正文所在标签的 HTML 源代码
- **author_xpath(str)**: 可选，文章作者的 XPath，用于定向提取文章作者
- **publish_time_xpath(str)**: 可选，文章发布时间的 XPath，用于定向提取文章发布时间
- **use_visiable_info(bool)**: 可选，HTML 是否带有节点坐标和可视化信息

ListPageExtractor 的函数原型为:

```
class ListExtractor:
    def extract(self, element: HtmlElement, feature)
```

各个参数的意义如下:

- **element(HtmlElement)**: 必填，经过 lxml.html.fromstring 处理后的 Dom 树对象
- **feature(str)**: 必填，列表中，任意一行的 XPath 或者内容。GNE 会根据这个 XPath 或者内容，自动找到它所在的列表，并返回该列表下面的全部内容。

CHAPTER 5

配置文件

API 中的参数 `title_xpath`、`host`、`noise_node_list`、`with_body_html`、`author_xpath`、`publish_time_xpath`、`body_xpath`、`use_visiable_info` 除了直接写到 `extract` 方法中外，还可以通过一个配置文件来设置。

请在项目的根目录创建一个文件 `.gne`，配置文件可以用 YAML 格式，也可以使用 JSON 格式。

- YAML 格式配置文件

```
title:
    xpath: //title/text()
host: https://www.xxx.com
noise_node_list:
    - //div[@class=\"comment-list\"]
    - //*[@style=\"display:none\"]
body:
    xpath: //div[@class="news-text"]
with_body_html: true
author:
    xpath: //meta[@name="author"]/@content
publish_time:
    xpath: //em[@id="publish_time"]/text()
use_visiable_info: false
```

- JSON 格式配置文件：

```
{
    "title": {
        "xpath": "//title/text()"
    },
    "host": "https://www.xxx.com",
    "noise_node_list": ["//div[@class=\"comment-list\"]",
                        "//*[@style=\"display:none\"]"],
    "body": {
        "xpath": "//div[@class=\"news-text\"]"
    },
    "with_body_html": true,
    "author": {
        "xpath": "//meta[@name=\"author\"]/@content"
    },
    "publish_time": {
        "xpath": "//em[@id=\"publish_time\"]/text()"
    },
    "use_visiable_info": false
}
```

这两种写法是完全等价的。

配置文件与 `extract` 方法的参数一样，并不是所有字段都需要提供。你只需要填写你需要的字段即可。

如果一个参数，既在 `extract` 方法中，又在 `.gne` 配置文件中，但值不一样，那么 `extract` 方法中的这个参数的优先级更高。

CHAPTER 6

已知问题

1. 目前本项目只适用于新闻页的信息提取。如果目标网站不是新闻页，或者是今日头条中的相册型文章，那么抽取结果可能不符合预期。
2. 可能会有一些新闻页面出现抽取结果中的作者为空字符串的情况，这可能是由于文章本身没有作者，或者使用了已有正则表达式没有覆盖到的情况。

CHAPTER 7

交流沟通

如果您觉得 GNE 对您的日常开发或公司有帮助，请加作者微信 mekingname（或扫描下方二维码）并注明”GNE”，作者会将你拉入群。

验证消息：GNE

如果你不用微信，那么可以加入 Telegram 交流群：https://t.me/joinchat/Bc5swww_XnVR7pEtDUl1vw

CHAPTER 8

目录

- genindex
- modindex
- search