目标：

我们想要通过搜索特定关键词，并在添加一系列的筛选机制之后，爬取网页反馈的文章内容及创建存储文章信息的 json 文件

最终写出一个 run()函数，其中有以下参数：

1. search_keyword: 搜索关键词
2. min_page_count/min_word_count: 最少页数/字数
3. num_years: 要几年以内的文章

案例：

robo.run(search_keyword= '中芯国际', min_num_page=20, num_years=5)最后生成有关中芯国际的，至少 20 页的，近五年内的研报/资讯

爬虫流程：

1. 搜索关键词（例如：中芯国际）

2. 筛选机制：
   - 最少页数/字数（默认大于 10 页）
     - 只爬取页数大于 10 页的文章
     - 如果文章没有分页，则爬取文章字数大于 3000 字的文章
   - 在某特定时间段发表（默认三年内）
     - 三年内： 2018-01-01 – 今天
     - 五年内：2016-01-01 – 今天

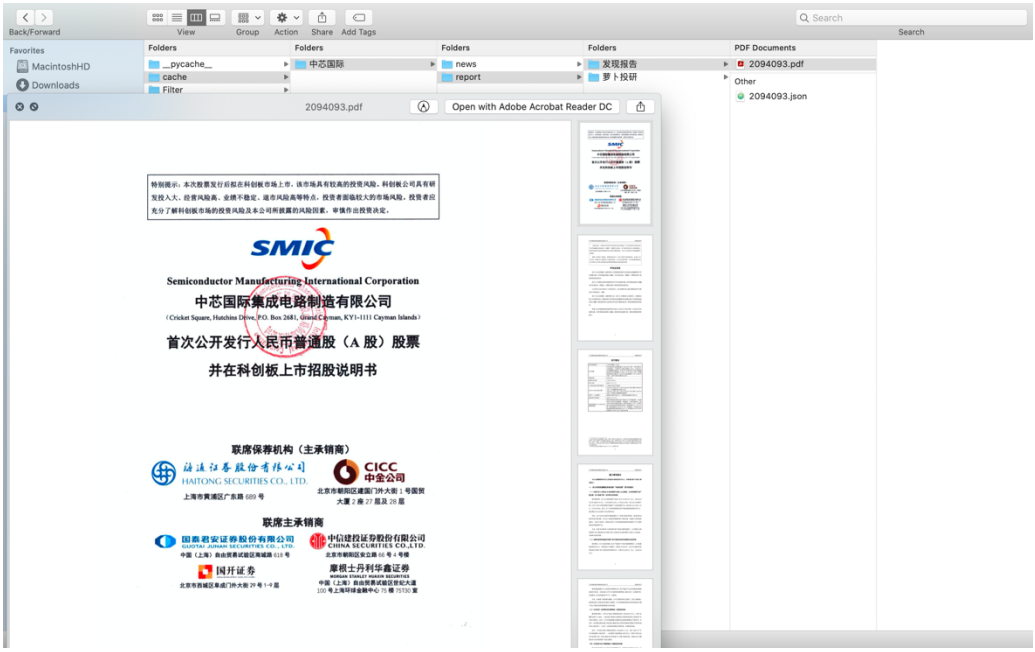
3. 根据以上信息筛选出符合条件的页面，根据页面数据类型分为以下三种情况：
   a. 目标网站提供 pdf 直接下载途径
   b. 目标网站不提供 pdf 下载途径，文章数据为文本格式
   c. 目标网站提供图片及文本格式的文章内容

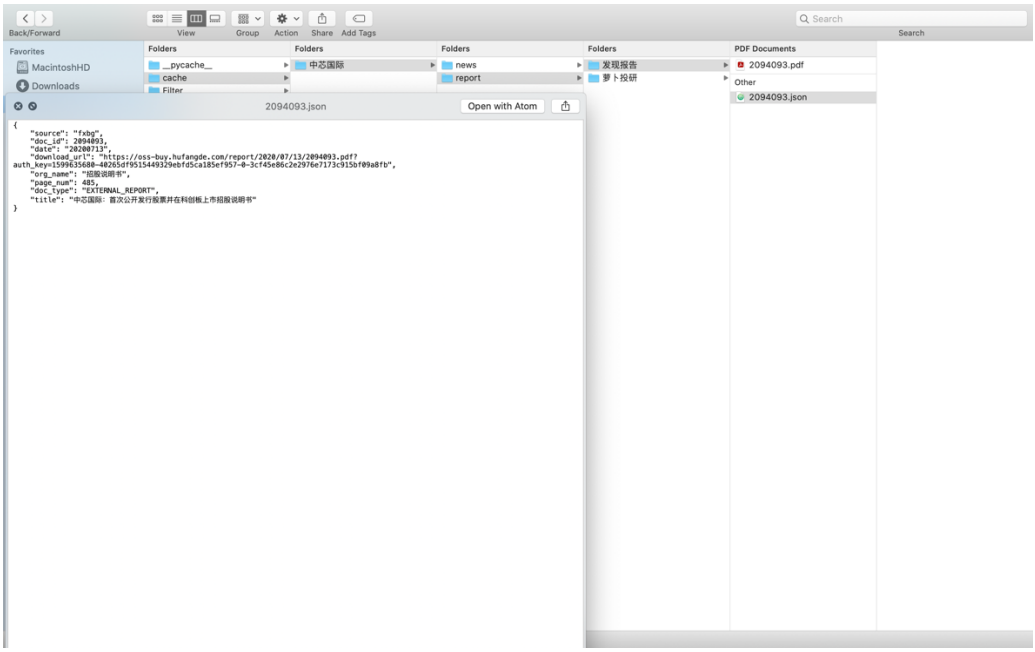
4. 根据以上所列举的页面的不同情况，以及网站为资讯(news)或是研报(report)类别，在 cache 文件的 news/report 文件夹内新建文件夹，用来存储以下文件：

   a. 当目标网站提供 pdf 直接下载途径，则下载：

   - 包含文章文本内容的 PDF 文件（以文件 id 命名）
     - 路径案例: cache/中芯国际/news/sougouweixin/1419820.pdf
   - 包含文章信息的 json 文件
     - 路径案例: cache/中芯国际/news/sougouweixin/1419820.json

   案例：





b. 当目标网站不提供 pdf 下载途径，文章数据为文本格式，则下载：

- 包含文章文本内容的 html 文件（以文件 id 命名）
    - 路径案例: cache/中芯国际/news/sougouweixin/1419820.html
- 包含文章信息的 json 文件
    - 路径案例: cache/中芯国际/news/sougouweixin/1419820.json

html 案例：

不要有多余的文本，如页面上方的标头，广告，推荐页面等，只要标题、作者、日期、内容。

c. 当目标网站提供图片及文本格式的文章内容，则将图片用 ocr 转换成文本后储存在 json 文件中，同时存储 html 文件：

- 若图片数量小于五张
- 若图片数量为 5 - 10 张
- 若图片数量多于 10 张，将文件储存至一个默认的文件夹里

总结：

	JSON 文件	*HTML 文件*	*PDF 文件*
目标网站提供 pdf 直接下载途径	要	不要	要

目标网站不提供 pdf 下载途径，文章数据为文本格式	要	要	不要
目标网站提供图片及文本格式的文章内容	要	要	不要

文件格式：

JSON/txt 文件中包含以下信息，所有 key 和 value 的格式请按照我们要求，编码为 utf-8，并过滤掉任何多余字符（如 '\r' '\b'等）：

1. 网站来源 "source"（白鲸出海 "bjch" /未来智库 "wlzk" ……）
2. 文章 ID "doc_id"（在所爬取的网页的 id）
3. 发表日期 "date"（格式：20200129，8 位数字：yyyymmdd）
4. 下载 url "download_url"
5. 作者/机构 "org_name"（如 xx 券商/xx 证券)
6. 页数/字数 "page_num"/ "word_count"，如果有 pdf 文件就用页数，没有就用字数
7. 资料种类 "doc_type"（研报："EXTERNAL_REPORT"/咨询："NEWS"）
8. 文章标题 "title" ("中芯国际：首次公开发行股票并在科创板上市招股说明书")
9. *ocr 文本内容 "ocrtext"若网站只提供图片，则将图片转换成的文本储存至此