我们想爬以下几个网站：

搜狗微信搜索：https://weixin.sogou.com/

新浪新闻搜索：https://search.sina.com.cn/

白鲸出海：https://www.baijingapp.com/

未来智库：https://www.vzkoo.com/index.php?s=/Wxwap/Index/index.html

乐晴智库：http://www.767stock.com/

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

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

1. 根据以上信息筛选出符合条件的页面，并在当前路径创建新建文件夹，用来存储以下文件：

* 包含文章文本内容的PDF文件（以文件id命名）
  + 路径example: sougouweixin/1419820.pdf
* 包含文章信息的json/txt文件
  + 路径example: sougouweixin/1419820.json

案例：

A screenshot of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generated

JSON/txt文件中包含以下信息：

1. 文章ID （在所爬取的网页的id）
2. 发表日期
3. 标题
4. 页数/字数
5. 作者/机构 （e.g. xx券商/xx证券)
6. 网站来源 （白鲸出海/未来智库）
7. 资料种类 （研报：externalReport/咨询：news）