

python实现CNKI知网爬虫

📅 2017-04-27

最近需要用到CNKI知网的爬虫爬取相关论文的信息，在网上搜索了一下，在GitHub上找到了之前大神写的[知网爬虫](#)，但是可能由于知网更新了接口，这个爬虫现在不能使用，于是打算自己动手写一个。因为是第一次写爬虫，全当是练手了。写的不对的地方，还请大家批评指正！

使用BeautifulSoup

要实现爬虫，首先要选取合适的库，当然也可以纯写，只不过有点费时费力。百度之后，发现BeautifulSoup库的相关文档很清楚，上手简单。而我在实现Cnki_Spider的过程中基本上就用到了这个库中的两三个方法，个人觉着在知网爬虫中，这个库已经足够使用。

在使用时只需要import即可，如果报错的话，需要使用pip命令安装一下。

```
1 from bs4 import BeautifulSoup
```

找到合适的知网接口

找到爬虫工具之后的重中之重就是找到一个适合我们使用的接口，找了一些其他人写的爬虫后发现，[这个接口](#)非常好用，他不会像知网主页上面把所有的搜索结果都嵌套在一个 `iframe` 中，而且搜索结果与知网相差无几。

[中药益气养阴方诱导白血病细胞凋亡机理的研究](#)

联合化疗是目前治疗急性**白血病**的主要方法，临床上常用的肿瘤化疗药物均能诱导敏感肿瘤细胞的凋亡。但由于化疗药物的副作用、抗药性等，常使患者不能坚持完成化疗或生存质量下降、易于复发等，导致患者难以长期生存而死亡。中药近年来在抗**白血病**方面有其独到之处，取得了较好的临床疗效。但由于中药成分及作用机理复杂，研...

山东中医药大学 硕士论文 2002年

下载次数 (348) | 被引次数 (0)

[三丁酸甘油酯诱导白血病细胞增殖抑制、分化、凋亡及其机理的研究](#)

目的：探讨三丁酸甘油酯(Tributyrin, TB)对人急性**白血病**细胞株NB4、K562、SHI-1以及**白血病**原代培养细胞的增殖抑制、诱导分化、凋亡作用及其可能的作用机制。 方法：利用台盼蓝拒染法，MTT比色法、细胞形态学观察，NBT还原率，联苯胺染色法，检测细胞表面抗...

苏州大学 硕士论文 2003年

下载次数 (122) | 被引次数 (1)

[激发型CD40单抗5C11对白血病来源树突状细胞的诱导作用及其细胞生物学特性研究](#)

树突状细胞(DC)是目前发现的功能最强的抗原递呈细胞，其功能独特之处在于能激活初始型T细胞。DC在机体的抗肿瘤免疫应答中起着重要的作用。业已表明，DC表面持续性表达CD40分子，且在成熟过程中呈上调性表达，激发型CD40mAb能激发CD40功能，促进DC成熟并使其具有直接激发CD8⁺T细胞的能力...

苏州大学 硕士论文 2003年

下载次数 (55) | 被引次数 (0)

[P27在急、慢性白血病中表达的研究](#)

〔目的〕研究细胞周期蛋白依赖激酶抑制物P27在急性**白血病**、慢性髓细胞性**白血病**(AL、CML)患者中的表达，探讨P27在**白血病**中的作用机制及其临床意义。〔方法〕应用免疫细胞化学SP法检测43例AL患者、27例CML患者及10例对照P27的表达，并用蛋白印迹(Western-blot) (DA...

昆明医学院 硕士论文 2004年

下载次数 (51) | 被引次数 (0)

1 2 3 4 5 6 7 8 9 10 下一页

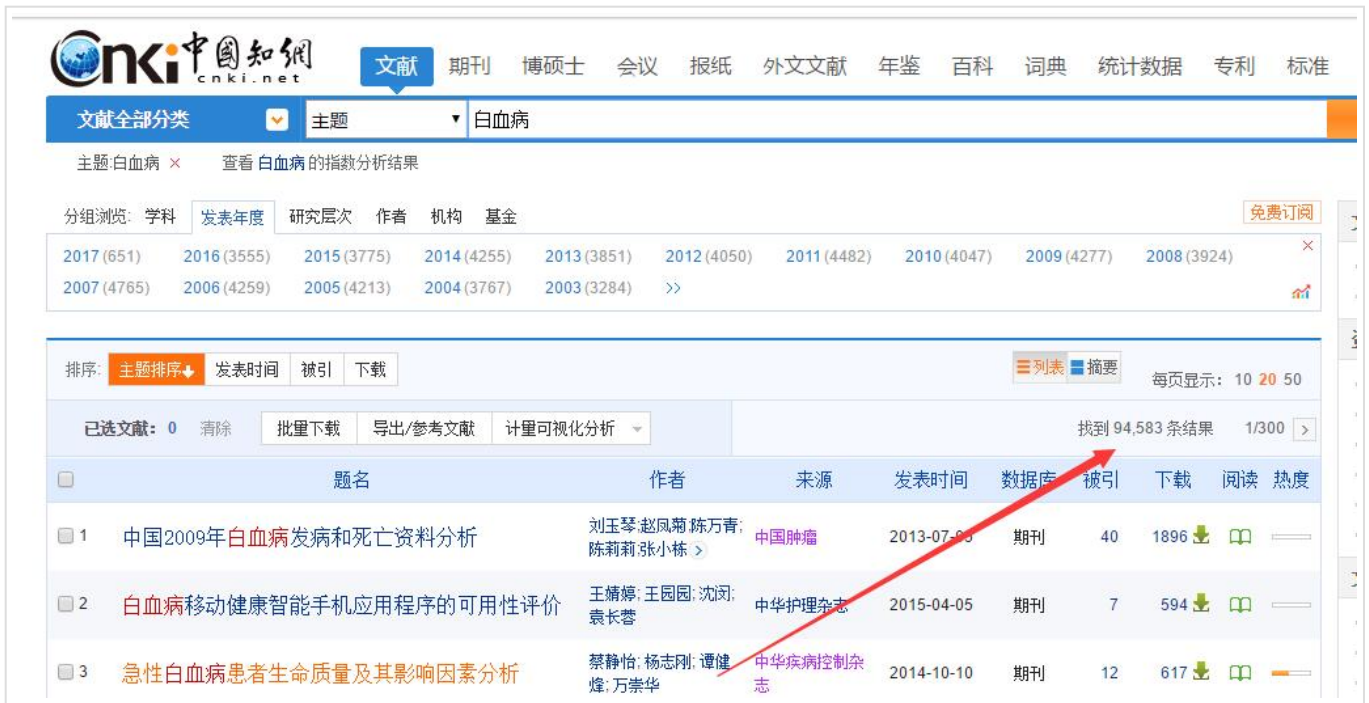
共找到相关记录94719条

相关搜索： 儿童白血病 白血病治疗 类白血病 亚白血病
红白血病 白血病基因 牛白血病 抗白血病
白血病，牛 小鼠白血病

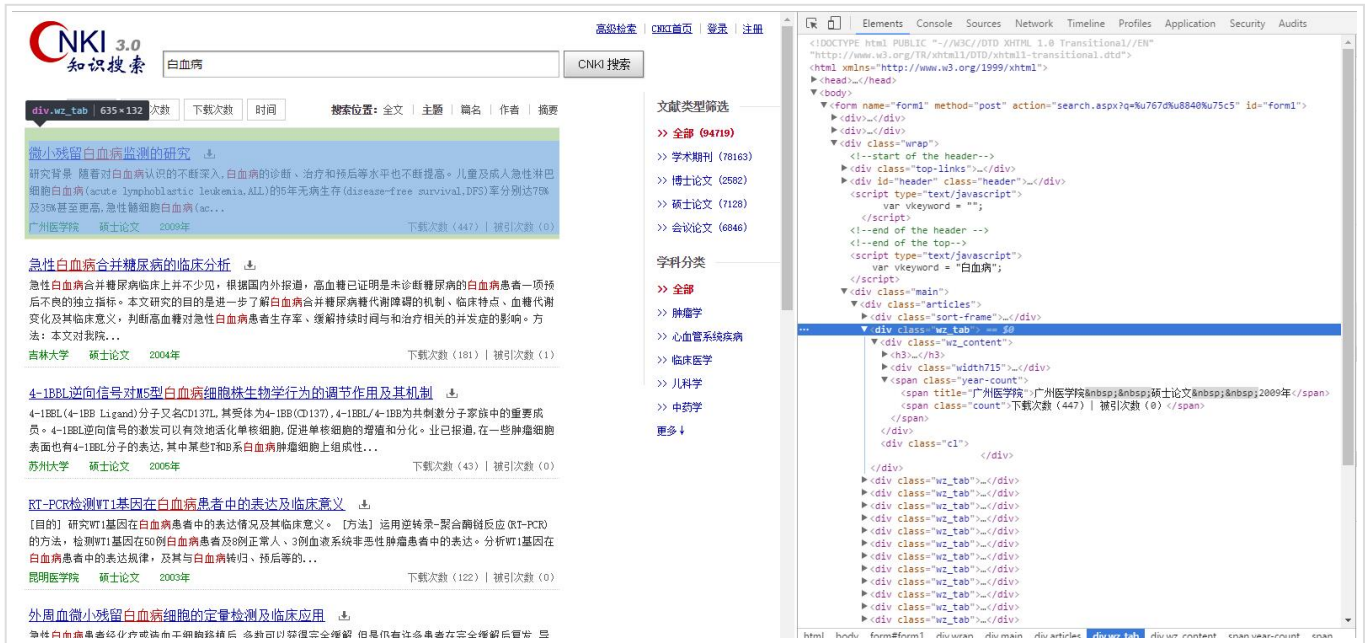
白血病

CNKI 搜索

[高级检索](#)



在这张图里就可以看到知识搜索中的所有搜索结果都可以直观获取到。



使用如下语句就可以轻松愉快地获取该页面所有内容（Python版本3.x）。

```
1 html = urllib.request.urlopen(page_url).read()
2 soup = BeautifulSoup(html, 'html.parser')
```

爬虫架构

因为在搜索页面中没有作者信息，摘要内容也不全，所以我们只能在这页面中爬取文章链接、文章标题、文章出处、年份以及下载次数、引用次数信息。在爬取的过程中，大部分时间其实都是在做字符串处理。

例如使用 `all = soup.find_all('div', class_='wz_content')` 这条语句就可以获取搜索页面中全部的

将这页获取到的信息保存在文本文件中，方便我们进入文章详情页面进行爬取。因为一个搜索页面显示15条信息，所以我们可以自己设置爬取页面数量等信息。

爬取后的结果如下：

微小残留白血病监测的
急性白血病合并糖尿病
4-1BBL逆向信号对M¹
RT-PCR检测WT1基因
外周血微小残留白血病
同源盒基因A₁ (10)
白血病DcR3与c-myc
清毒饮和养正片对白₁
红细胞生成素受体在急
多重RT-PCR检测儿童

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	下载网址	标题	来源	引用	作者	作者单位	关键词	摘要	参考文献												
2	http://www.	羟基羧加干	(当代医学)	下载次数	廖春淑	重庆作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-LCXZ200301012.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-LCXZ200302002.htm	http://www.cnki.										
3	http://www.	慢性粒细胞	(护士进修)	下载次数	陈玉华	张作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHYX200102000.htm												
4	http://cdm.c	CaMK II γ	浙江大学	下载次数	郑维威	学位授予单	关键词:	[摘要]													
5	http://cdm.m	TOR靶点	河北医科大学	下载次数	李杰	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHYX200407001.htm												
6	http://cdm.S	T3GALA家	大连医科大学	下载次数	李岩	学位授予单	关键词:	S [摘要]													
7	http://cdm.	慢性粒细胞	河北医科大学	下载次数	李岩	学位授予单	关键词:	S [摘要]													
8	http://cdm.S	hh信号通路	华中科技大学	下载次数	苏文霞	学位授予单	关键词:	J [摘要]													
9	http://cdm.	慢性粒细胞	三州大学	下载次数	王海鹰	学位授予单	关键词:	[摘要]													
10	http://www.m-	bcr-慢性粒	(张家口医	下载次数	薛海强	张作者单位	关键词:	F [摘要]													
11	http://cdm.	慢性粒细胞	北京中医药大学	下载次数	刘小鹏	学位授予单	关键词:	分 [摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-LNZY200110014.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-LCXZ199901011.htm	http://www.cnki.										
12	http://cdm.CD	4→CD4	中国大南大学	下载次数	周利	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDOTAL-BXBZ903301.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-GWYY200005007.htm	http://www.cnki.c										
13	http://www.bcr-	abl融合点	(中国病理)	下载次数	陈奕玉	王作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-ZBL5200004016.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHYX200008001.htm											
14	http://cdm.	微流控芯片	天津医科大学	下载次数	曹旭东	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-AIHZ200309003.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-KXTB201121006.htm	http://www.cnki.c										
15	http://www.	慢性粒细胞	(中国组学)	下载次数	李继红	黄作者单位	关键词:	[摘要]													
16	http://cdm.	自体来源的	吉林大学	下载次数	李杰	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHYX903304.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-ZLSW200001007.htm	http://www.cnki.c										
17	http://cdm.N	otch信号	第四军医大	下载次数	付伟	学位授予单	关键词:	N [摘要]													
18	http://cdm.	初治慢性粒	安徽医科大学	下载次数	夏雷鸣	学位授予单	关键词:	[摘要]													
19	http://cdm.TP	A逆转录	郑州大学	下载次数	林全德	学位授予单	关键词:	S [摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-GWZL199001008.htm												
20	http://www.	伊马替尼片	(当代医学)	下载次数	(徐玉秀	孙作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHYX200406005.htm												
21	http://www.	慢性粒细胞	(中国组学)	下载次数	(朱希山	宋作者单位	关键词:	[摘要]													
22	http://cdm.	慢性粒细胞	河北医科大学	下载次数	孟艳梅	学位授予单	关键词:	S [摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-AIHZ200904009.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-BXBZ200301001.htm	http://www.cnki.										
23	http://www.	慢性粒细胞	(中国病理)	下载次数	蒋元强	学位授予单	关键词:	[摘要]													
24	http://www.	慢性粒细胞	(临床医学)	下载次数	王新林	王作者单位	关键词:	[摘要]													
25	http://cdm.N	otch1对慢	第四军医大	下载次数	尹郁丹	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-ZLSW199902000.htm												
26	http://www.	急性巨核病	(首都医药)	下载次数	王建萍	作者单位	关键词:	[摘要]													
27	http://cdm.	慢性粒细胞	河北医科大学	下载次数	韩玉萍	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-XYSY200405008.htm	http://cdmd.cnki.com.cn/Article/CDMD-10089-1014248037.htm											
28	http://cdm.rn	血小板减少	山东中医药大学	下载次数	曹芳	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFD												
29	http://www.	骨髓基因D	(中国实验	下载次数	李芳	杨作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDOTAL-GWSQ201103014.htm												
30	http://www.	骨髓在慢性	广东省下	下载次数	赖运东	舒作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDOTAL-BXBZ200104006.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-JYYL200703029.htm	http://www.cnki										
31	http://www.	体外培养骨	(中国组学	下载次数	黄东	李作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-GWSX200003001.htm												
32	http://cdm.Cr	kL调控大	大连医科大	下载次数	施国平	学位授予单	关键词:	[摘要]													
33	http://www.	慢性粒细胞	(护士进修)	下载次数	林郁清	作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDOTAL-FSJX201001011.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-LCFC200512031.htm	http://www.cnki.										
34	http://cdm.x	旋复蜜素	延安大学	下载次数	李树香	学位授予单	关键词:	左 [摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-SHYH201411024.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-XCYZ201108005.htm											
35	http://www.	124例慢性	(当代医学)	下载次数	韩丽晶	作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-BATE200801019.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-DDYJ201024074.htm											
36	http://www.	丹参配合化	(四川中医	下载次数	韩明	吴作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDOTAL-ZWXH200001019.htm												
37	http://www.	中药治疗骨	(实用现代	下载次数	陈博香	汪作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-GSZY200610027.htm	http://www.cnki.com.cn/Article/CJFDOTAL-GXZY200705022.htm	http://www.cnki										
38	http://cdm.	慢性粒细胞	山东大学	下载次数	李梅	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDOTAL-ZGLZ201209011.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHAY201402003.htm											
39	http://www.	慢性粒细胞	(世界最新	下载次数	赵敬	作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-LCYJ200003015.htm												
40	http://www.	α-干扰素与	(军医进修	下载次数	杨清明	作者单位	关键词:	[摘要]													
41	http://www.	疑难病诊断	(湖南师范	下载次数	杨作成	作者单位	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDTOTAL-SYYZ200503000.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHEK9014208.htm											
42	http://cdm.I	马替尼片	苏州大学	下载次数	杨光东	学位授予单	关键词:	[摘要]	http://www.cnki.com.cn/Article/CJFDOTAL-LCZL200004019.htm	http://www.cnki.com.cn/Article/CJFDTOTAL-ZHYX200504006.htm											

- 需要加入错误重试机制，在网络出现问题后能够及时重试。
- 对爬取结果进行重新优化显示，输出到excel文件中。
- 尝试加入文章下载功能，但是知网对频繁下载有限制，不一定会成功。
- 考虑将爬取到的数据直接写入数据库中，方便存储以及后续操作。

增加的功能

增加了对作者单位、文章关键词的爬取，增加了对参考文献的爬取，如果需要爬取共引文献或其他文献，只需要将参考文献换为需要爬取的其他类别即可。

github下载点击[这里](#)，谢谢支持!

本文作者： Qiu Qingyu

版权声明： 本博客所有文章除特别声明外，均采用CC BY-NC-SA 3.0 CN许可协议。转载请注明出处!


本文永久链接： <http://qiuqingyu.cn/2017/04/27/python实现CNKI知网爬虫/>

python # 爬虫

◀ 配置Nginx对网页进行加密

微信分享出现错误并且闪退 ▶

0



所谓的高傲： _____

来吐槽吐槽~~

还没有评论，快来抢沙发吧!

邱庆羽的博客正在使用畅言

热评话题

python实现CNKI知网爬虫 | Qupid and M
Hexo博客自动备份到Coding | Qupid and
Django中使用celery报错ValueError: I/
学习笔记--如何选择合适的网络 | Qupid
hexo中禁止渲染文件的方法 | Qupid and

python使用matplotlib绘图 | Qupid and
hadoop集群搭建 | Qupid and Monkey's
部署hexo | Qupid and Monkey's Blog
配置Nginx对网页进行加密 | Qupid and
将hexo部署到腾讯云上 | Qupid and Mon

© 2015 – 2017 ♥ Qiu Qingyu

蜀ICP备17009365号

由 Hexo 强力驱动 | 主题 – NexT.Mist