201810307:

1. 伪造head头请求信息。
2. 根据特定的网页源码，获取标签信息，筛选目标。
3. <a href="http://so.csdn.net/so/search/s.do?q=Poop类&t=blog" target="\_blank">Poop类</a>

获取href的值：href = a["href"] # 获取a标签的链接

4．从后往前截取字符串：

name = img\_url[-9 : -4] # 截取

1. 多线程爬取：

# 线程池个数，电脑CPU个数

pool = Pool(4)

for i in range(4):

pool.apply\_async(download, args = (all\_url, root\_path, num) ) # apply\_async非阻塞且支持结果返回进行回调

pool.close()

pool.join()

20180308:

1. 正则表达式匹配替换特殊字符：

str = "哈？哈\*哈?哈|哈：哈:哈$哈@哈&哈<>哈“”哈哈''哈lx/j哈哈\哈1111111111"

result = re.sub(r'\\*|\?|\？|\:|\：|\||\&|\$|\@|\>|\<|\""|\'\'|\“”|\\|\/', "", str)

print(result)

对每个要筛选的元素进行匹配，用符号：| 进行分割，匹配字符串最后不加 | ，否则作为匹配左边或右边，右边为null，则在原str中每个字符之间都匹配替换成（re.sub()函数）目标字符串。