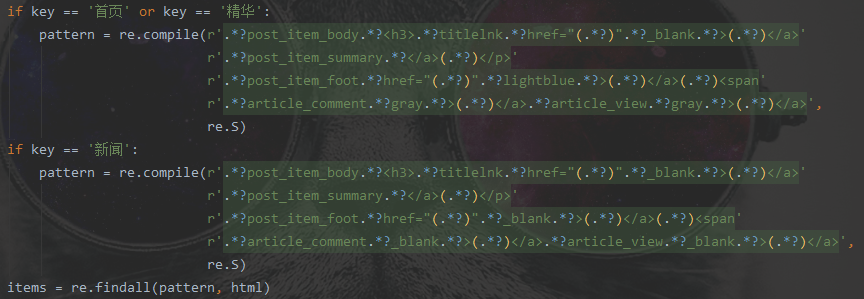
Python爬虫02 – 爬取博客园文章信息

1. 爬取目标
   1. 博客园首页、精华、新闻页面的文章的详细信息，比如：文章标题、链接、作者主页链接，发布时间，文章简介等。
   2. 数据处理为TXT文本格式，便于后期做数据可视化分析。
   3. 打包exe程序，每次运行程序，获取当天最新文章信息。
2. 解释

爬取的范围仅限于首页、精华、新闻页面是因为其包含了有用的信息，其他栏目下，非目标爬取内容。

1. 库的使用
   1. Requests
   2. Re
   3. Fake\_useragent 随机请求头
   4. Colorama 字体颜色
   5. Numba 加速器
   6. Time
   7. Tqdm 进度显示条
2. URL分析
   1. 首页：<https://www.cnblogs.com/>，点击下一页，URL为：<https://www.cnblogs.com/#p2>。因为翻页规律显而易见，只需要修改p后面的数字即可。
   2. 精华：<https://www.cnblogs.com/pick/>，点击下一页，URL为：<https://www.cnblogs.com/pick/#p2>。翻页规律同上。
   3. 新闻：<https://www.cnblogs.com/news/>，翻页规律同上。
3. 正则匹配

通过分析发现，首页、精华和新闻页面的格式基本相同，由于新闻页面的class属性为post\_item\_foot的div标签下略有差异，所以将第一个正则做小改动后便可适用于新闻页面的匹配，匹配截图如下：



1. 数据存储

前面在爬取目标中已经谈到存储为文本格式，便于后期做数据可视化分析。

1. 其他
   1. 打包用到的是pyinstaller第三方库，执行pip install pyinstaller进行安装，pyinstaller两个参数：
      1. -F – 指定打包后只生成一个exe格式的文件
      2. –i – 改变生成程序的icon图标（图片必须是ico格式）
      3. 打包命令为：pyinstaller –F –i New.ico New.py
   2. 为了方便打包为exe程序，将源代码复制并重命名为了New.py，icon图标名称也是New.ico
   3. 生成的.exe程序包含在dist文件夹下，点击运行的结果会直接创建在该目录下（三个文本文件）如图：

