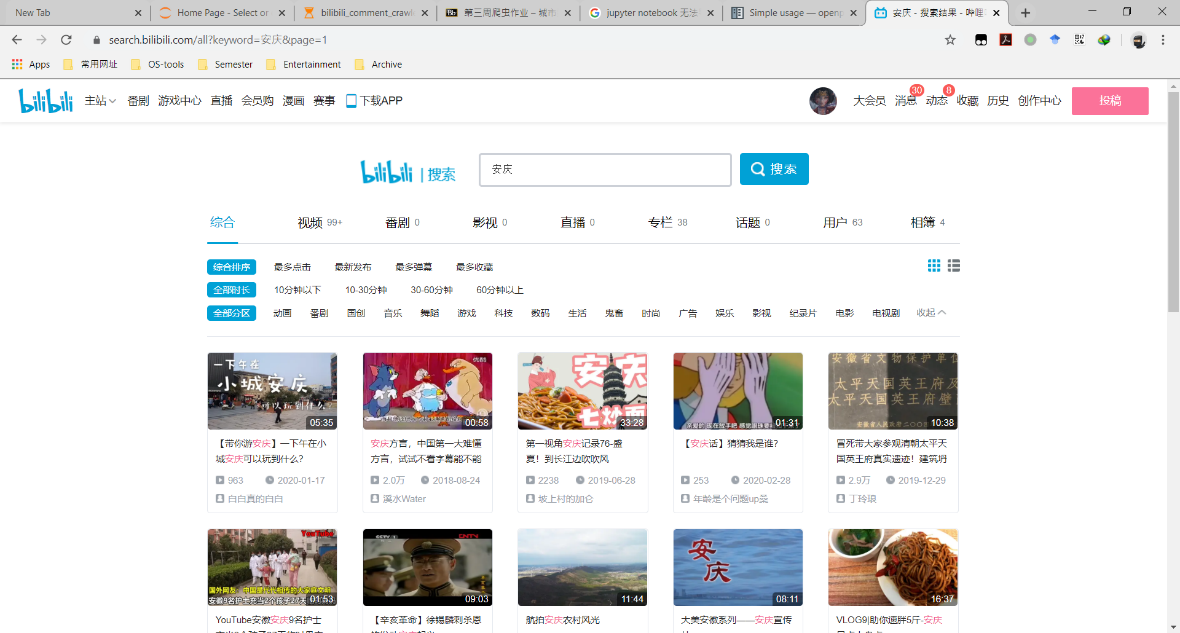
# 题目：Bilibili城市相关视频弹幕爬取

## 目的

通过对Bilibili上的和每个城市相关的视频的弹幕信息进行分析，获取城市的特点，可以进行城市评价

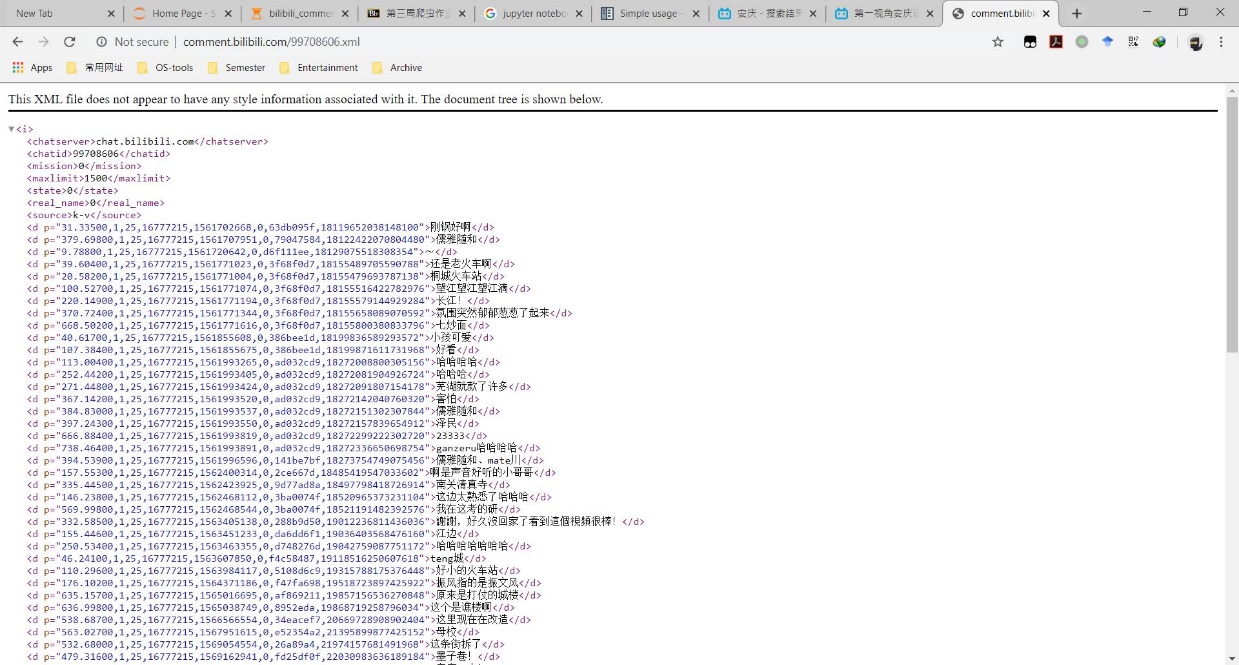
## 方法：

1. 通过Bilibili的搜索功能，检索城市名称，从而得到相关的所有视频，检索的网站为<https://search.bilibili.com/all?keyword=%E5%AE%89%E5%BA%86&page=1>。

* 其中，keyword参数为关键词，同时每个搜索结果页面只列了20个视频，可以通过另一个参数page=x指定显示哪一个页面。为了防止页面参数超出总页面数，可在页面底部的最后一页按钮上提取总页面数并对页面数作限制。
* 在搜索页面的每个图片元素可以获取对应视频的具体链接，使用选择器选择出所有的图片元素然后提取超链接属性即可。这里提取的超链接需要稍作修饰以符合网页地址格式。在本次提取中，每个城市最多只取20个搜索页面，考虑也就是最多只抓取前400个视频的信息。

1. 访问每个视频的具体页面，提取弹幕相关的ID和播放数量信息。

* 播放量信息在页面的json脚本里，如下图，可以直接获取。其中userInteractionCount就是播放量。
* 获取弹幕需要先提取出弹幕的ID信息，然后此ID可用于直接获得视频所有弹幕。通过在视频网页上搜索关键此cid，可以找到很多，但是经发现，只有一个找到的cid对饮于本视频的弹幕ID，其他的是推荐视频列表的cid。因此，通过在页面所有的script元素中按顺序搜索是否存在cid字符串，即可找到该ID。这里没有结构化的元素，因此直接采用字符串匹配的方式提取这个cid。

1. 通过弹幕专用的网页获取弹幕。Bilibili提供了一个网页[http://comment.bilibili.com/{cid}.xml](http://comment.bilibili.com/%7bcid%7d.xml)可以直接获取xml格式的弹幕信息，只需要传入对应的cid即可。如图即为一个示例。之后直接解析该页面即可。
2. 爬取所有的数据之后，保存为一个文件即可，使用pickle就可以简单的保存了。这里我尝试爬取了安徽省所有城市的信息，分别保存在不同的pickle文件里。

## 代码

见附件的notebook，需要以下包，环境Anaconda的python3.7。

requests

bs4

urllib

json

## 结果