最近《歌手2018》比较火，要问郭郭我喜欢哪个歌手？当然是张韶涵啦，当年的公主小妹满满青春的回忆啦。说到歌手肯定会想到他们的成名曲啦，比如李圣杰的《痴心绝对》，汪峰的《怒放的生命》啦等等。

这次我要做的就是通过python爬虫爬取虾米的歌词信息，然后做成词云，直观得看出歌手们得一些信息。

比如像下面这样的



猜猜她是谁？

一.准备工作 语言：Python2

工具：Pycharm，firefox

技术:requests，BeautifulSoup，re，csv，jieba，

WordCloud，numpy

1. 流程概述 前往虾米爬取歌词信息

整理歌词

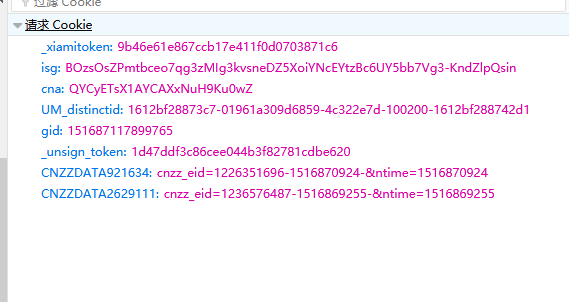
储存歌词

实现词云

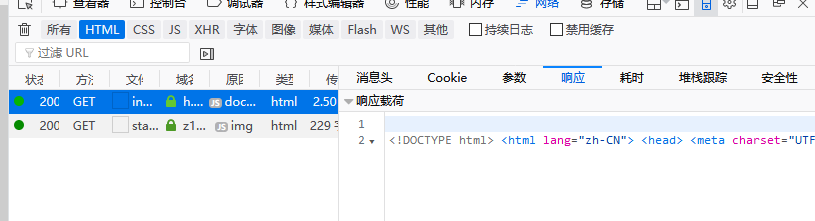
1. 具体实现

进入www.xiami.com照例查看网页结构

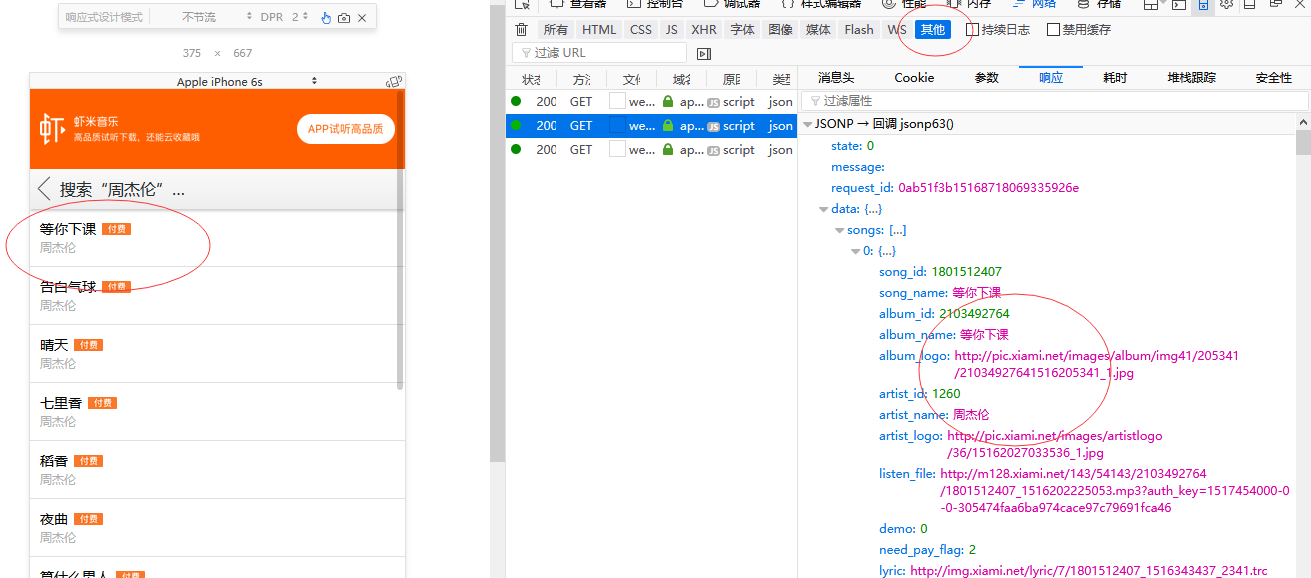




HTML里一堆乱码以及cookie验证瞬间吓退了我。不过我是不会死心的，大公司的网站肯定有H5版本的，于是转战https://h.xiami.com



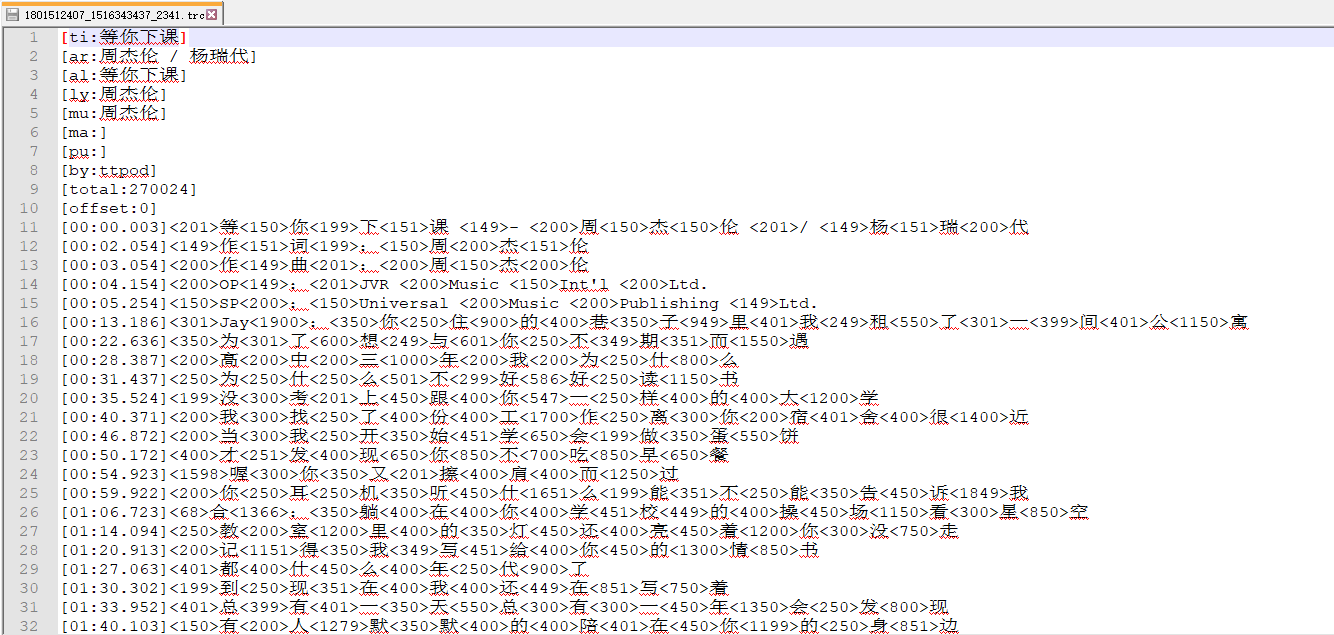
查看html以后发现网页里面啥都没，猜测是用ajax动态生成的，查看了其他形式的返回时，我看到了这个。。。



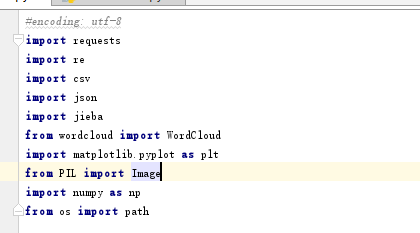
哈哈哈，功夫不有心人，原来在返回的json字符串里发现了“秘密”，而且说json，json在日常后端编写中就极为常用，而且遵循一定格式，比杂乱的HTML标记好处理多了。



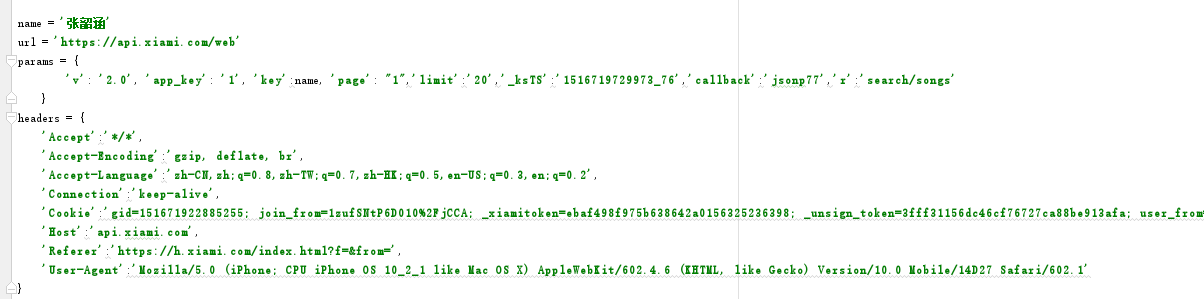
通过查看json报文我们看到了这个，歌词的英文不就是lyric嘛？我们复制这个链接在浏览器打开，是一个文件，下载以后使用记事本或者notepad打开，这不就是歌词嘛，不过要想办法去掉[]<>等妨碍查看的东西。



好了，我们已经知晓可行性以后，开始编码



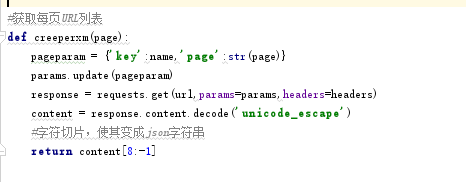
这是要导入的包，pip安装即可，词云wordcloud可能会有编译问题，py2直接会告诉你C++包的下载地址，py3可以去网上下载已经编译好的whl文件

由于虾米也是有反爬虫机制的，所以我们也要伪装成模拟器，

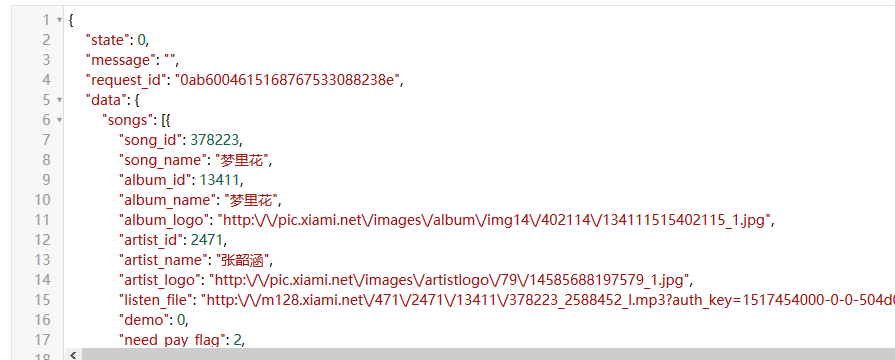
最简单的就是设置header头



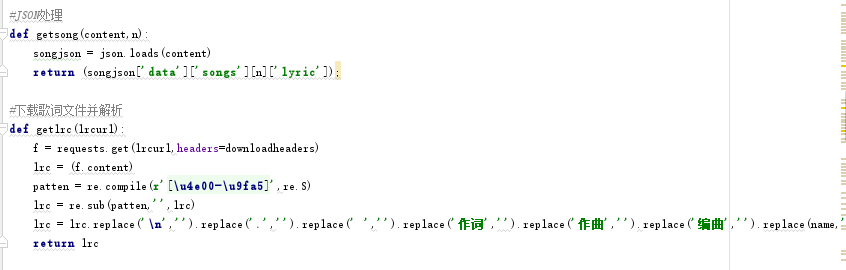
完全复制下即可

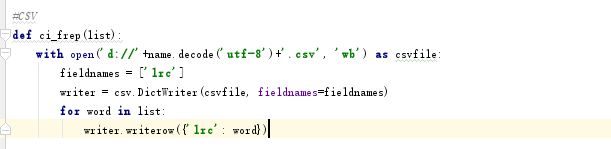


首先现要获取网页json，返回会有一些干扰字符，利用切片去除即可



以上就是json字符串，利用python的json包拿到指定数据，然后利用正则提取中文，再替换掉“作词/作曲”等常用词汇，其实可以用正则，结巴分词的停用词实现，这样代码更为优雅有效，时间关系，我只是基本处理了下。



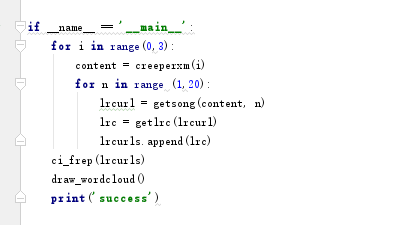


只有一行的csv很简单，并不需要特殊处理



结巴分词+词云，具体API说明注释有，我们只需要选个字体以及好的背景图也可以，color\_mask是背景图，如果不用就是默认方形图显示。

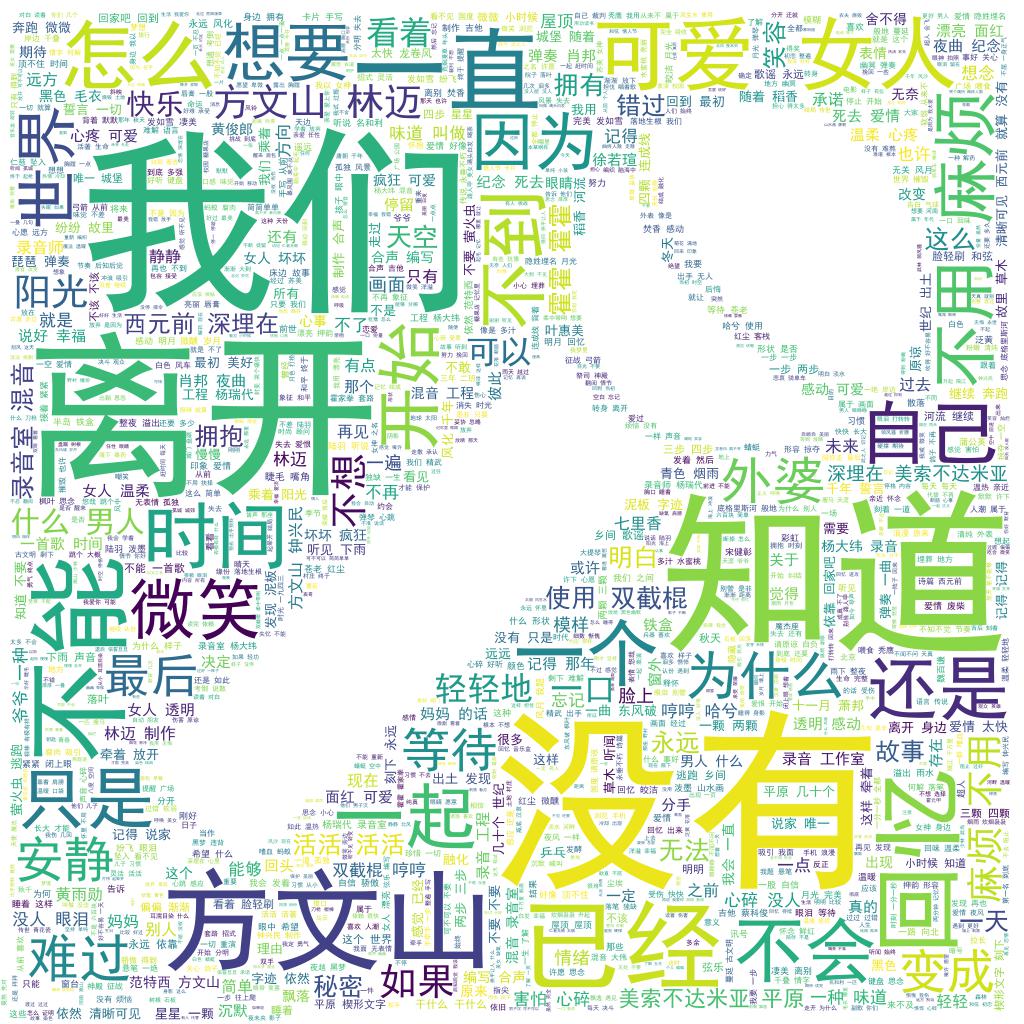
总流程



然后看看效果



医生很喜欢用黄伟文和林夕啊



方文山果然是杰伦的御用词人

然后看了几个以后我发现。。。。。。明星唱歌最常出现的词是。。。。。。。。。。你也写个看看呗