爬取贴吧获取html

- IWH
  - IS:爬取一定页数的HTML
  - W:爬虫入门第一步爬取HTML
  - H：使用request工具get去解析某个网页获得其uft8编码的文件，保准在本地
- 数据准备
  - url池
    - 用list存储
      - list是可变的类型
    - https://tieba.baidu.com/f?kw=XXXX&ie=utf-8&pn=0
      - kw=XXXX中XXXX为所要爬取的贴吧名字
      - pn=0中0表示第一页，第二页为50，第三页为100，以此类推
  - 爬取的页数
  - reuqest所需的headers，具体见request部分
- 使用工具
  - request
    - api
      - get
        - 参数
          - url，由url池提供
          - headers
            - 由chrome查看网页源代码->Network->Headers->Request Headers-> User-Agent
            - 以字典的方式存储，注意key和value均是字符串
              - 详见request官方文档->快速上手->定制请求头
        - 返回
          - class object 内有很多的方法，详见官方文档，这里以r表示
          - r.content
            - 二进制相应内容（bytes）
              - 写文件需要以str的形式，因此需要decode()
                - decode()默认参数encoding='utf-8'?

- 保存本地
  - with open as
    - python的文件读写 with open语句
    - 文件读写可能会遇到IOError，以但出错f.close就无法调用，可以使用try...except...finally,但是这样太麻烦，直接使用with open() as