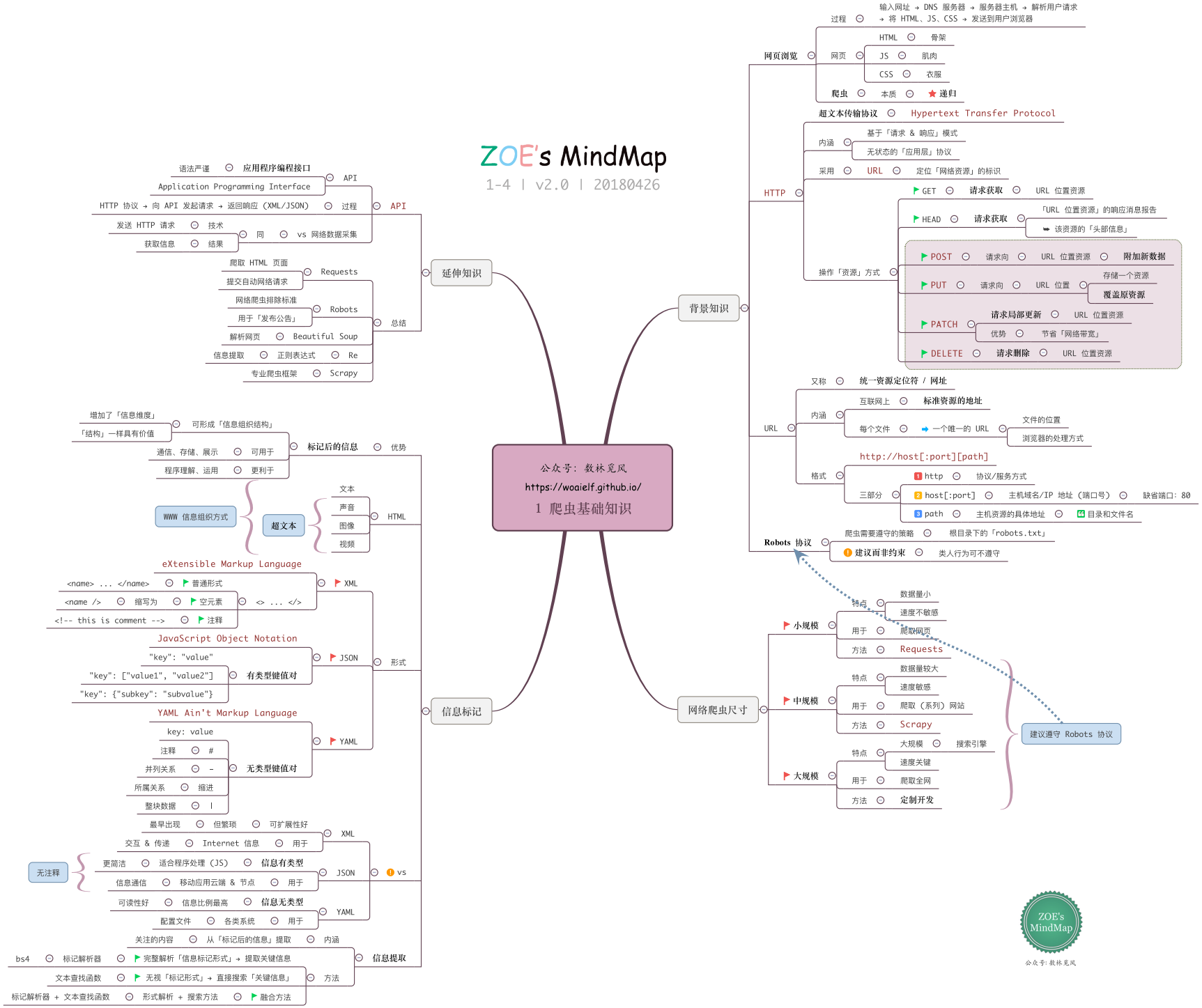

ZOE's MindMap
1-4 | v2.0 | 20180426
公众号：数林觅风
https://woaielf.github.io/
1 爬虫基础知识
背景知识
网页浏览
过程
输入网址 → DNS 服务器 → 服务器主机 → 解析用户请求 → 将 HTML、JS、CSS → 发送到用户浏览器
网页
HTML 骨架
JS 肌肉
CSS 衣服
爬虫 本质 递归
HTTP
超文本传输协议 Hypertext Transfer Protocol
内涵
基于「请求 & 响应」模式
无状态的「应用层」协议
采用 URL 定位「网络资源」的标识
操作「资源」方式
GET 请求获取 URL 位置资源
HEAD 请求获取 「URL 位置资源」的响应消息报告
该资源的「头部信息」
POST 请求向 URL 位置资源 附加新数据
PUT 请求向 URL 位置 存储一个资源 覆盖原资源
PATCH 请求局部更新 URL 位置资源
优势 节省「网络带宽」
DELETE 请求删除 URL 位置资源
URL
又称 统一资源定位符 / 网址
内涵
互联网上 标准资源的地址
每个文件 一个唯一的 URL 文件的位置 浏览器的处理方式
格式
http://host[:port][path]
三部分
1 http 协议/服务方式
2 host[:port] 主机域名/IP 地址（端口号） 缺省端口：80
3 path 主机资源的具体地址 目录和文件名
Robots 协议
爬虫需要遵守的策略 根目录下的「robots.txt」
建议而非约束 类人行为可不遵守
网络爬虫尺寸
小规模
特点 数据量小 速度不敏感
用于 爬取网页
方法 Requests
中规模
特点 数据量较大 速度敏感
用于 爬取（系列）网站
方法 Scrapy
大规模
特点 大规模 搜索引擎 速度关键
用于 爬取全网
方法 定制开发
建议遵守 Robots 协议
信息标记
形式
HTML
WWW 信息组织方式
超文本
文本
声音
图像
视频
优势
标记后的信息
增加了「信息维度」
「结构」一样具有价值
可形成「信息组织结构」
通信、存储、展示 可用于
程序理解、运用 更利于
XML
eXtensible Markup Language
<name> ... </name> 普通形式
<name /> 缩写为 空元素 <> ... </>
<!-- this is comment --> 注释
JSON
JavaScript Object Notation
"key": "value"
"key": ["value1", "value2"]
"key": {"subkey": "subvalue"}
有类型键值对
YAML
YAML Ain't Markup Language
key: value
注释 #
并列关系 -
所属关系 缩进
整块数据 |
无类型键值对
vs
无注释
XML
最早出现 但繁琐 可扩展性好
交互 & 传递 Internet 信息 用于
JSON
更简洁 适合程序处理（JS） 信息有类型
信息通信 移动应用云端 & 节点 用于
YAML
可读性好 信息比例最高 信息无类型
配置文件 各类系统 用于
信息提取
内涵
关注的内容 从「标记后的信息」提取
方法
bs4 标记解析器 完整解析「信息标记形式」→ 提取关键信息
文本查找函数 无视「标记形式」→ 直接搜索「关键信息」
标记解析器 + 文本查找函数 形式解析 + 搜索方法 融合方法
延伸知识
API
语法严谨 应用程序编程接口
Application Programming Interface
HTTP 协议 → 向 API 发起请求 → 返回响应（XML/JSON） 过程
发送 HTTP 请求 技术
获取信息 结果
同 vs 网络数据采集
总结
Requests
爬取 HTML 页面
提交自动网络请求
Robots
网络爬虫排除标准
用于「发布公告」
Beautiful Soup
解析网页
Re
信息提取 正则表达式
Scrapy
专业爬虫框架
ZOE's MindMap
公众号：数林觅风

ZOE's MindMap

2-4 | v2.0 | 20180426

公众号：数林觅风

https://woaielf.github.io/

2 Requests 库

异常

- 类别
  - .ConnectionError
    - e.g.
      - DNS 查询失败
      - 拒绝连接
    - 网络连接异常
  - .HTTPError
    - HTTP 错误异常
  - .URLRequired
    - URL 缺失异常
  - .TooManyRedirects
    - 重定向异常
    - 超过最大重定向次数
  - .ConnectTimeout
    - 超时异常
    - 连接远程服务器
  - .Timeout
    - 超时异常
    - 请求 URL 超时
- 函数
  - r.raise_for_status()
    - != 200
    - HTTPError

步骤

1. 构造「Request 对象」
   - 向服务器请求资源
2. 返回「Response 对象」
   - 包含服务器资源
3. 属性
   - .status_code
     - 成功 200
     - 失败 404
   - .text
     - 字符串形式
     - 响应内容
   - .encoding
     - 显示网页内容
     - r.encoding
     - 根据
     - 编码方式
     - 猜测
     - 从 HTTP header
     - ISO-8859-1
     - 当 header 不存在 charset
   - .apparent_encoding
     - 编码方式
     - 从内容分析
   - .content
     - 二进制形式
     - 响应内容
   - .url/headers

★ 通用代码框架

```
def getHTMLText(url):
  try:
    r = requests.get(url, timeout=30)
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    return r.text
  except:
    return '产生异常'
```

主要方法

- .request()
  - 构造一个请求
    - method
      - 'GET'
    - url
  - ★ 基础方法
    - 支撑以下方法
- .get()
  - 获取网页
- .head()
  - 获取网页头信息
  - r.headers
- .post()
  - 提交 POST 请求
- .put()
  - 提交 PUT 请求
- .patch()
  - 提交局部修改请求
- .delete()
  - 提交删除请求

request() 衍生方法

控制访问参数

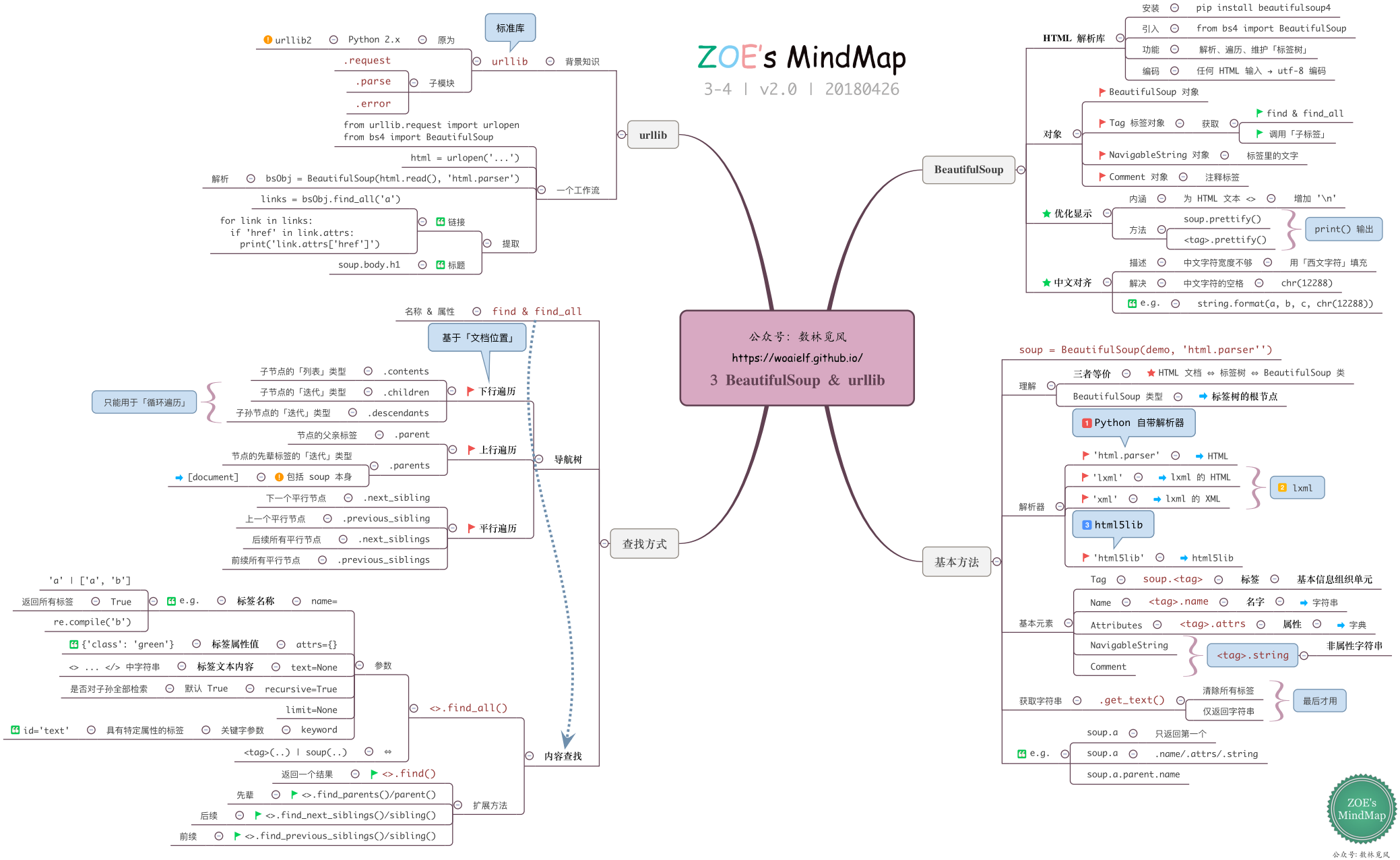
- params
  - 字典/字节序列
  - 作为参数添加到 url
- data
  - 字典/字节序列/文件对象
  - e.g.
    - 字典
      - 自动编码为 form
    - 字符串
      - 自动编码为 data
- json

作为 Request 的内容

- headers
  - 定制 HTTP 头信息
- cookie | auth
- files
  - 字典
  - 传输文件
- timeout
- proxies
  - 设定访问代理服务器
- allow_redirects/stream/verify/cert

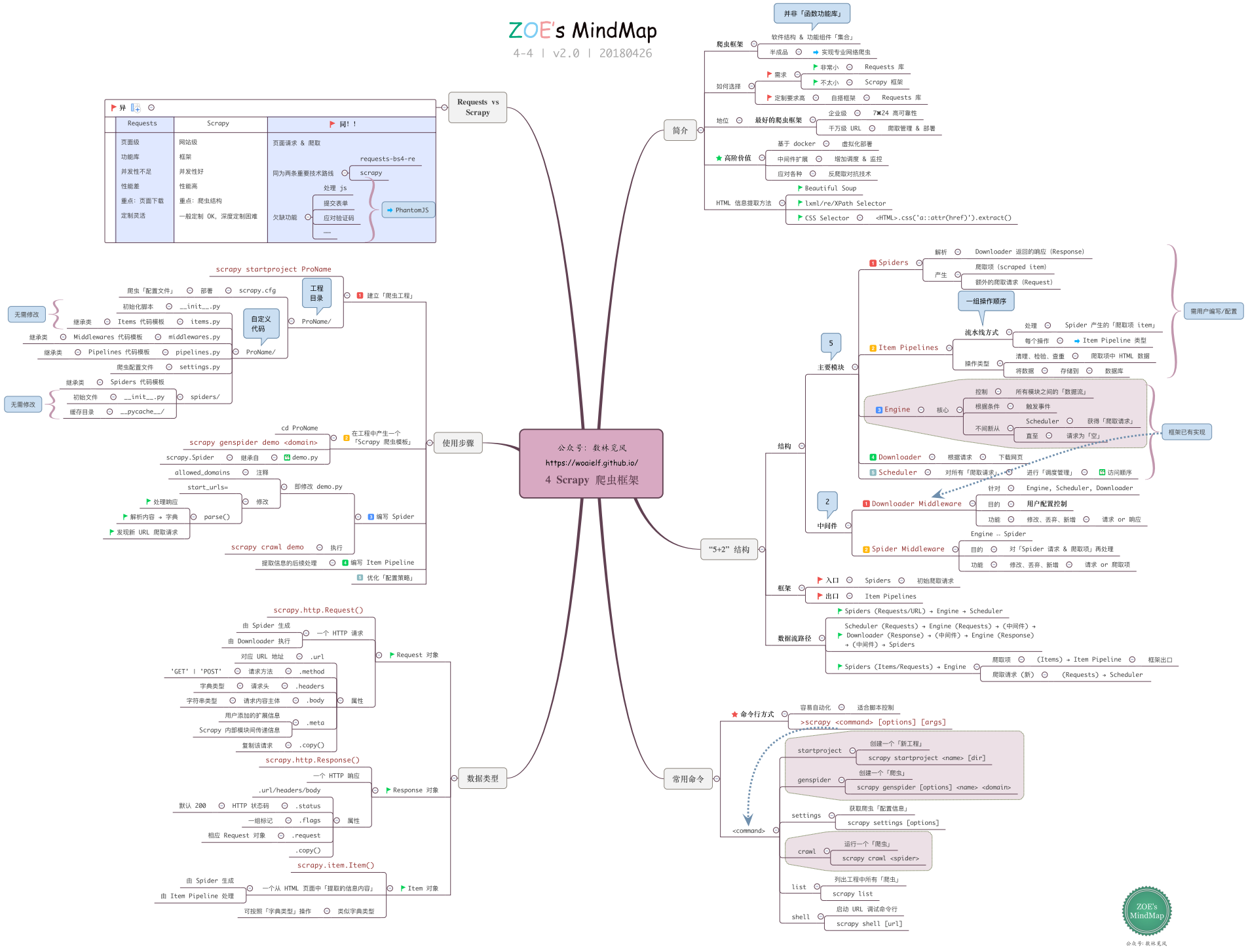
ZOE's MindMap

公众号：数林觅风


ZOE's MindMap
3-4 | v2.0 | 20180426
公众号：数林觅风
https://woaielf.github.io/
3 BeautifulSoup & urllib
urllib
标准库
urllib
背景知识
urllib2
Python 2.x
原为
.request
.parse
.error
子模块
from urllib.request import urlopen
from bs4 import BeautifulSoup
html = urlopen('...')
解析
bsObj = BeautifulSoup(html.read(), 'html.parser')
一个工作流
links = bsObj.find_all('a')
for link in links:
if 'href' in link.attrs:
print('link.attrs['href']')
链接
提取
soup.body.h1
标题
查找方式
find & find_all
名称 & 属性
基于「文档位置」
导航树
下行遍历
子节点的「列表」类型
.contents
子节点的「迭代」类型
.children
子孙节点的「迭代」类型
.descendants
只能用于「循环遍历」
上行遍历
节点的父亲标签
.parent
节点的先辈标签的「迭代」类型
.parents
[document]
包括 soup 本身
平行遍历
下一个平行节点
.next_sibling
上一个平行节点
.previous_sibling
后续所有平行节点
.next_siblings
前续所有平行节点
.previous_siblings
内容查找
<>.find_all()
参数
'a' | ['a', 'b']
返回所有标签
True
e.g.
标签名称
name=
re.compile('b')
{'class': 'green'}
标签属性值
attrs={}
<> ... </> 中字符串
标签文本内容
text=None
是否对子孙全部检索
默认 True
recursive=True
limit=None
id='text'
具有特定属性的标签
关键字参数
keyword
<tag>(..) | soup(..)
⇔
扩展方法
返回一个结果
<>.find()
先辈
<>.find_parents()/parent()
后续
<>.find_next_siblings()/sibling()
前续
<>.find_previous_siblings()/sibling()
BeautifulSoup
HTML 解析库
安装
pip install beautifulsoup4
引入
from bs4 import BeautifulSoup
功能
解析、遍历、维护「标签树」
编码
任何 HTML 输入 → utf-8 编码
对象
BeautifulSoup 对象
Tag 标签对象
获取
find & find_all
调用「子标签」
NavigableString 对象
标签里的文字
Comment 对象
注释标签
优化显示
内涵
为 HTML 文本 <>
增加 '\n'
方法
soup.prettify()
<tag>.prettify()
print() 输出
中文对齐
描述
中文字符宽度不够
用「西文字符」填充
解决
中文字符的空格
chr(12288)
e.g.
string.format(a, b, c, chr(12288))
基本方法
soup = BeautifulSoup(demo, 'html.parser')
理解
三者等价
HTML 文档 ⇔ 标签树 ⇔ BeautifulSoup 类
BeautifulSoup 类型
标签树的根节点
1 Python 自带解析器
解析器
'html.parser'
HTML
'lxml'
lxml 的 HTML
'xml'
lxml 的 XML
2 lxml
3 html5lib
'html5lib'
html5lib
基本元素
Tag
soup.<tag>
标签
基本信息组织单元
Name
<tag>.name
名字
字符串
Attributes
<tag>.attrs
属性
字典
NavigableString
Comment
<tag>.string
非属性字符串
获取字符串
.get_text()
清除所有标签
仅返回字符串
最后才用
e.g.
soup.a
只返回第一个
soup.a
.name/.attrs/.string
soup.a.parent.name
ZOE's MindMap
公众号：数林觅风

ZOE's MindMap

4-4 | v2.0 | 20180426



ZOE's MindMap

1-1 | v2.1 | 20180402

公众号：数林觅风

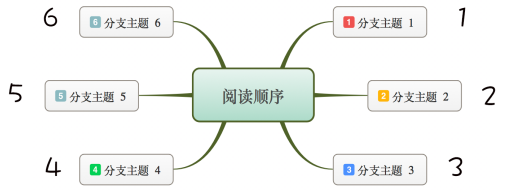
https://woaielf.github.io/

ZOE 思维导图规则

1 阅读顺序

顺时针方向！

自右→向下→向左→向上



2 ★ 图标含义

- 重点标注
  - ★ 核心内容
  - ★ 进阶内容/巧妙应用/Tips
  - ★ 了解内容
- 平行层级/项目符号
  - 一级
  - 二级
  - 三级
  - 递进/交叠
- 专有名词 — 名词定义，需重点理解
- 警告 — 警告注意
- 补充 — 资源
- 举例 — 范例，有助于理解
- 继续
  - 输出内容
  - 补充说明
  - 附加内容/内容递进
- 下一行继续内容（用于代码）
- 等价于 ⇔
- 表示「对比」 vs

3 基本规则

编号-系列总数 | 版本号 | 更新日期

1-1 | v2.0 | 20180323

- 释义
  - 本系列共「1」幅思维导图
  - 当前为第「1」幅图
  - 第二版本 — 当前版本
  - 2018-03-23 — 最后更新日期
- 字体设置
  - Mshtakan — 主标题 + 一级标题
  - Monaco — 普通文本
  - 加粗、颜色、更大
  - Monaco — 英文
  - Mshtakan — 中文
  - 强调文本

4 关于作者

- ★「最新版本高清 PDF」下载
- ZOE
  - zoemindmap — 数林觅风 — 公众号
  - https://woaielf.github.io/ — 博客
  - 数林觅风 — 新浪微博
  - 数林觅风——数据科学思维导图笔记 — 知乎专栏
- “关键词”
  - I'm a GTDer
  - 思维导图
  - 数据科学
  - 生物信息
  - 精进思考

图例

- 编号1
- 编号2
- 编号3
- 编号4
- 一级
- 二级
- 三级
- 核心理解
- 进阶内容
- 简单了解
- 输出内容
- 补充前面
- 注意
- 附加/递进
- 资源
- 举例说明
- 爱心
- 名词定义

ZOE's MindMap

公众号：数林觅风

获取更多 ZOE 思维导图

https://woaielf.github.io/

数林觅风 | zoemindmap

@数林觅风

ZOE 酱

思维导图 | 数据科学 | 编程语言 | 读书笔记 | 精进思考