



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

数据科学导论

Introduction to Data Science

第二章 数据分析基础

黄振亚，陈恩红

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html>



数据分析基础

2

- 数据采集 Data Collection
- 数据存储 Data Storage
- 数据预处理 Data Preprocessing
- 特征工程 Feature Engineering



数据采集

3

- 无时无刻产生数据，获得数据的方式多种多样



网页



数据库



监控



测量



传统媒体



数据采集

4

- 数据检索
- 公开数据
- 批量数据获取
 - 网络爬虫
- 数据筛选



数据采集：数据检索

5

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎
 - 百度：适合于搜索中文信息
 - Google：更适合搜索英文信息

Baidu 百度 手写识别 数据集

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约696,000个

优秀的手写识别数据集_百度文库

★★★★★ 评分: 4/5 2页

2012年4月11日 - 优秀的手写识别数据集_IT\计算机_专业资料。手写识别能够使用户按照最自然、最方便的输入方式进行文字输入,易学易用,可取代键盘或者鼠标。手写识别属...
wenku.baidu.com/link?url=...

开源标准数据集——mnist(手写字符识别) - 计算机科学与艺术 - ...

2016年4月25日 - 开源标准数据集 ——mnist(手写字符识别) 2016-04-25 18:45 823人阅读 评论(2) 收藏 举报 分类: 机器学习(109) 作者同美文汇X ...
blog.csdn.net/fanchunh... - 百度快照 · 1588条评论

为您提供: notepad++ 第一报关网 zinnia

如何写一个简单的手写识别算法? - 编程 - 知乎

2015年11月21日 - ocr问题的重点是怎么选择特征,比如知名的uci 数据集就有以下这些特征: 1...图形的识别来说,knn写起来不用几分钟就完成,其中判断相似性可以直接将手写的图形...

www.zhihu.com/question... - 百度快照

SMALL
LARGE
MIXED

开源标准数据集——mnist (手写字符识别)

2016-04-25 18:45 895人阅读

分类: 机器学习 (111)

版权声明: 本文为博主原创文章,未经博主允许不得转载。

下载地址: [mnist.pkl.gz](http://www.fanchunhao.com/mnist.pkl.gz)

0 4 1 9 2 1 3 1 4 3



数据采集：数据检索

6

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎
 - Google: 更适合搜索英文信息
 - Google Scholar, DBLP

三 Google 学术搜索 EKT: Exercise-aware

文章	找到约 82 条结果 (用时0.04秒)
时间不限	Ekt: Exercise-aware knowledge tracing for student performance prediction
2021以来	Q Liu, Z Huang, Y Yin, E Chen, H Xiong... - IEEE Transactions on ..., 2019 - ieeexplore.ieee.org
2020以来	For offering proactive services (eg, personalized exercise recommendation) to the students in computer supported intelligent education, one of the fundamental tasks is predicting student performance (eg, scores) on future exercises, where it is necessary to track the ...
2017以来	
自定义范围...	☆ ⚡ 被引用次数: 77 相关文章 所有 12 个版本
按相关性排序	[HTML] Exercise Hierarchical Feature Enhanced Knowledge Tracing
按日期排序	H Tong, Y Zhou, Z Wang - International Conference on Artificial ..., 2020 - Springer
不限语言	... Eur. J. Psychol. Assess. 16(1), 3 (2000) CrossRef Google Scholar. 5. Huang, Z., et al.: Ekt:Exercise-aware knowledge tracing for student performance prediction. IEEE Trans. Knowl. Data Eng. (2019) Google Scholar. 6. Johnson, SC: Hierarchical clustering schemes ...
中文网页	☆ ⚡ 被引用次数: 4 相关文章 所有 6 个版本
简体中文网页	[PDF] Context-aware Knowledge Tracing Integrated with The Exercise

Index of /xml

Name	Last modified	Size	Description
Parent Directory		-	
CHANGES.txt	2019-11-22 21:20	3.5K	
README.txt	2019-11-22 21:20	3.5K	
dblp.dtd	2019-11-22 21:20	12K	
dblp.xml.gz	2021-09-14 02:20	624M	
dblp.xml.gz.md5	2021-09-14 02:20	46	
docu/	2018-03-01 16:43	-	
osd.xml	2020-12-18 16:26	1.5K	
release/	2019-08-20 15:57	-	

Apache/2.4.29 (Ubuntu) Server at dblp.uni-trier.de Port 443



数据采集：数据检索

7

□ 最简单、最灵活的数据获取方式就是依靠检索

The screenshot shows a search interface for datasets. At the top, there is a search bar with the query "GPT". Below the search bar are several filter options: "上次更新日期" (Last updated date), "下载格式" (Download format), "使用权限" (Usage rights), "学科" (Subject), "提供方" (Provider), and "免费" (Free). A large result card for the "gpt-dataset" by Haziqsayyed is displayed, showing its source as "huggingface.co" and the last update date as "Feb 22, 2023". Below this, two more dataset cards are shown: "kaggle GPT2 Output Dataset" from kaggle.com, last updated on Nov 12, 2019; and "kaggle Paraphrased Articles using GPT-3" from kaggle.com, last updated on Jan 21, 2020.

找到 100 多个数据集

gpt-dataset
Haziqsayyed/gpt-dataset

探索网页: [huggingface.co](#)

73 篇学术文章引用了此数据集 ([在 Google 学术搜索中查看](#))

数据集上次更新日期
Feb 22, 2023

作者
Haazique Sayyed

说明
Haziqsayyed/gpt-dataset dataset hosted on Hugging Face and contributed by the HF Datasets community

kaggle GPT2 Output Dataset
kaggle.com

上次更新日期: Nov 12, 2019

zip

kaggle Paraphrased Articles using GPT-3
kaggle.com

上次更新日期: Jan 21, 2020

zip



数据采集：公开数据

8

- 代表性公开数据集
 - 1400万的图像数据
 - <http://www.image-net.org/>
 - Amazon从2008年开始就为开发者提供几十TB的开发数据
 - <http://aws.amazon.com/datasets>
 - YouTube视频的统计与社交网络数据
 - <http://netsg.cs.sfu.ca/youtubedata/>



数据采集：公开数据

9

□ 代表性公开数据集

- 用户评分MovieLens: <https://grouplens.org/datasets/movielens/>
- 文本数据-头条: <https://github.com/aceimnorstuvwxyz/toutiao-text-classification-dataset>
- 网络数据-Large scale network: <https://snap.stanford.edu/data/>
- 教育数据:
 - ASSISTmentsData-学业: <https://sites.google.com/site/assistmentsdata/home/>
 - BASEGroup: <https://github.com/bigdata-ustc/EduData>
- 阿里天池数据-数据平台: <https://tianchi.aliyun.com/dataset/>
- 公开大数据竞赛的数据: KDDCup, NeurIPS Challenge



数据采集：批量数据获取

10

- 大量数据的获取难以手动实现，需借助**爬虫程序**
 - 也有可能通过交易（购买）“数据”而得
- 网络爬虫是一个自动在网上抓取数据的程序
 - 爬虫本质上就是**下载**特定网站网页的HTML/JSON/XML数据，并对数据进行**解析、提取与存储**
 - 通常先定义一组**入口URL**，根据页面中的其他URL，**深度优先或广度优先**的遍历访问，逐一抓取数据

热点要闻 | 个性推荐 | 进入推荐版



习近平对智利进行国事访问 发署名文章 专题
同智利总统会晤 中拉媒体领袖峰会 为印度总理莫迪访问 治理政策
李克强：打造改革创新开放的新标杆
“致中拉朋友 希望在新的历史条件下加强交流合作”会见国防部长
纸币要消失了？央行筹备数字货币
数字货币是否安全？能带来什么好处？纸币不会不作废？

国家卫计委：全国性公共场所控烟条例有望今年出台
新华社正式签署军事情报保护协定 共享军情情报
习近平出席中拉媒体领袖峰会开幕式
新华社正式签署军事情报保护协定 共享军情情报
揭秘：神十一返航的开舱 香港都带回了什么
父子同日被终止 人大代表资格 被移送司法机关
北京三审“气象灾害防洪条例”征求意见稿 入法
大连一载3万吨燃油油船突然起火 周围近百余艘渔船
韩媒：韩日签署《军情协定》为深化双边军事合作起步
特朗普：纽约时报对我非常粗暴 不看此报能多活20多年
e租宝案件8项指控罪名 目前已追缴涉案资产逾百亿元
湖南浏阳1岁小孩被一句诗“叫你”累死 景观工程
中国首次成为资本净输出国 流量反超日本居全球第二
2020年东京奥运会会徽征集数约2万件
父亲半道弃女儿遗体于垃圾桶内 当事人接受调查
外交部回应达赖窜访：美方的错误做法损害中美关系
北京市食药监局：网传水体污染致淡水鱼下架不可信

热搜新闻词 HOT WORDS ↓

习近平同智利总统巴切莱特举行会谈	习近平出席中拉媒体领袖峰会开幕式	全国性公共场所控烟条例有望今年出台	央行筹备数字货币
蒋洁敏因受贿案被提起公诉	李连杰因吸毒被立案侦查	老虎牙人案当事人起诉动物园	香港国安法草案
蒋洁敏因受贿案被提起公诉	李连杰因吸毒被立案侦查	老虎牙人案当事人起诉动物园	香港国安法草案

百度百家 BAIDU.COM | 财经 | 文化 | 娱乐 | 体育 | 加入百家号

百度开放大批语音黑科技
拦车救狗：识别恐怖主义
方舟子：木槿能预防脑癌吗

PUBG
加入百家号
TOP
HOT
杨丽娟
杨丽娟
方舟子



数据采集：网络爬虫

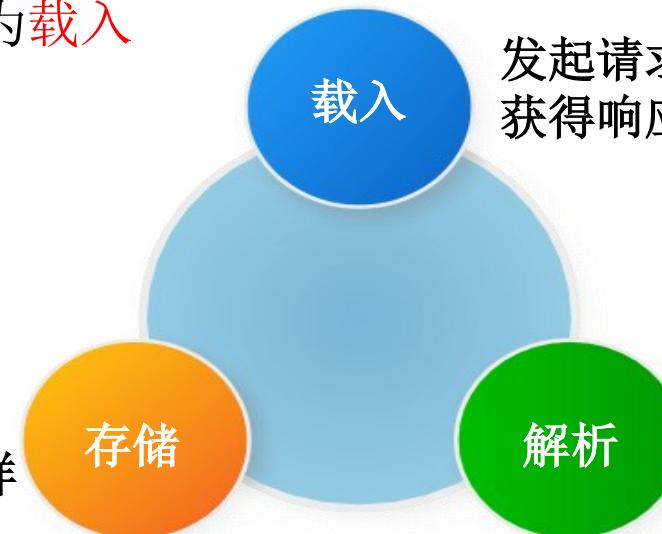
11

□ 网络爬虫是什么？

- 网络爬虫（又被称为网页蜘蛛，网络机器人，网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。
 - 请求网站并提取数据的自动化程序
- 爬虫的行为可以划分为：载入、解析、存储，
 - 最复杂的部分为载入



获得数据
存储数据-类型多样



发起请求(Request)
获得响应(Response)-类型多样

解析内容
提取数据-类型多样

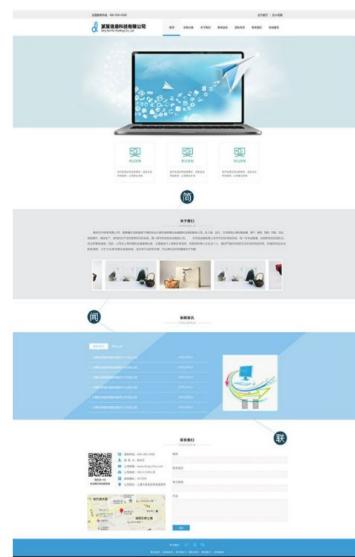
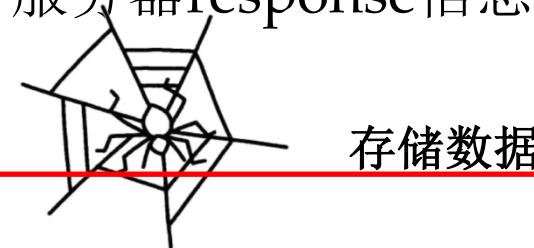


数据采集：网站数据

12

□ 访问网页示例

- 网站数据主要依托于[网页](#)（html, 超文本标记语言）展示
- 用户Request服务器，服务器response信息（html等）



浏览器解析
展示数据



发起请求(Request)



获得响应(Response)
网页数据



数据采集：网站数据

13

□ 网页示例

- 网站数据主要依托于网页（html, 超文本标记语言）展示
- 用户Request服务器，服务器response信息（html等）
- 课程主页<http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html>
- 右键“检查”查看网页源代码

Introduction to Data Science 数据科学导论

课程代码：CS4043.01

学院：011计算机科学与技术学院

课程教师：董振亚, 陈恩红

上课时间：(1-18周)每周二下午第8、9节：15:55-17:30, 教室：高新区 GT-B105



数据采集：网站数据

14



```
1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
2 <html>
3   <head>
4     </head>
5   <body>
6     <font size="8px"><strong> <center>Introduction to Data Science </center></strong> <font>
7       <strong> <center>数据科学导论 </center></strong> <font>
8         <br>
9         <font size="6px"><strong> <center>课程代码：CS4043.01 </center></strong> <font>
10        <strong> <center>学院：011计算机科学与技术学院 </center></strong> <font>
11        <font size="6px">
12          <strong>
13            <center>课程教师：
14              <a href="http://staff.ustc.edu.cn/~huangzhy/" target=" -parent" >黄振亚</a>,
15              <a href="http://staff.ustc.edu.cn/~cheneh/" target=" -parent" >陈恩红</a>
16            </center>
17          </strong>
18        </font>
19        <br><font size="6px">
20          <strong>
21            <center>上课时间：(1-18周)每周二下午第8、9节：15:55-17:30，教室：高新区 GT-B105 </center>
22          </strong>
23        </font>
24
25
26 <hr size="1px" noshade>
27 助教（含作业提交）：
28 <br><dir><font size="5px">
29   <li>肖桐, &nbsp; <a href="mailto:ds_intro2024@163.com">ds_intro2024@163.com</a></li><br>
30   <li>QQ群： <a href="https://qm.qq.com/q/Fv9uQZVG8w">483794507</a> </li></font>
31 </dir>
32
33 <hr size="1px" noshade>
34 Recommended Textbooks：
35   <br><dir><font size="5px">
36     <li><a href="http://book.douban.com/subject/5377669/" target=" -parent" >《Introduction to Data Mining》 </a></li>
37     <li><a href="http://book.douban.com/subject/10590856/" target=" -parent" >《统计学习方法》 </a></li>
38     <li><a href="https://book.douban.com/subject/26708119/" target=" -parent" >《机器学习》 </a></li>
39     <li><a href="https://item.jd.com/31931077735.html" target=" -parent" >《大数据导论》 </a></li></font>
40   </dir>
41 <br size="1px" noshade>
```



网络爬虫：载入

15

- 载入：将目标网站数据下载到本地
 - Html, HyperText Markup Language
 - 爬虫程序向服务器发送网络请求 Request，获取相应的网页
 - 网站常用网络协议：http, https
 - 数据常用请求方式：get, post
 - get：参数常放置在URL中
 - http://www.abc.com?p=1&q=2&r=3,
 - 问号后为参数
 - post：参数常放置在一个表单中（报文头（header））
 - 在向目标URL发送请求时，将参数放置在一个网络请求的报文头中
 - 更安全



网络爬虫：载入

16

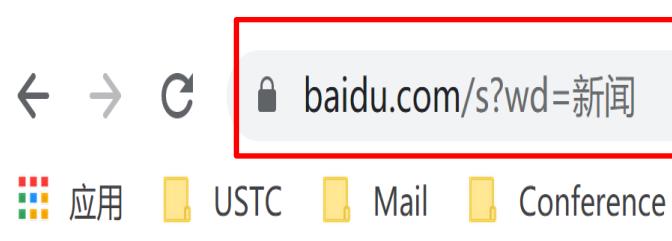
□ 载入：将目标网站数据下载到本地

□ 数据常用**请求方式**：get, post

■ **get**：参数常放置在URL中

■ http://www.abc.com?p=1&q=2&r=3，问号后为**参数**

■ 例如，https://www.baidu.com/s?wd=新闻



▼ General

Request URL: https://www.baidu.com/s?wd=%E6%96%
97%BB

Request Method: GET

Status Code: 200 OK

Remote Address: 127.0.0.1:9999

Referrer Policy: strict-origin-when-cross-origin

▼ Request Headers View source

请求头 Accept: text/html,application/xhtml+xml



网络爬虫：载入

17

- 载入：将目标网站数据下载到本地
 - 数据常用**请求方式**：get, post
 - **post**：参数常放置在一个表单中
 - 在向目标URL发送请求时，将参数放置在一个网络请求的报文头中
 - 相比于Get，多了Form Data部分（请求体）
 - 更安全：登录操作常用（不会放在URL后面）

Baidu 百度 · 用户名密码登录

手机号/用户名/邮箱

密码

登录

忘记密码?

立即注册

x Headers Preview Response »

General

Request URL: https://passport.baidu.com

Request Method: POST

Status Code: 200 OK

Remote Address: 127.0.0.1:9999

Referrer Policy: unsafe-url

请求体

Form Data view source view URL-encoded

staticpage: https://www.baidu.com/cache/user/html/v3Jump.html
charset: UTF-8
token: dd fac7e17ce70dc6187ce33dfee73ed
tpl: mn

username:
password: JyR31zjQ6WKlg1OsW0jf12!
3jz15ofHQtMTeUneEHAbKW0



网络爬虫：载入

18

- 载入：将目标网站数据下载到本地
 - 数据常用**请求方式**：get, post
 - 获得服务器的响应：Response，即获取网页源代码

响应头

▼ Response Headers View source

Bdpage type: 2

Bdqid: 0xdabb3fd4000073b3

Headers	Preview	Response	Initiator	Timing
▼ General				
Request URL:	http://staff.ustc.edu.cn/~huangzhy/			
Request Method:	GET			
Status Code:	200 OK (from disk cache)			
Remote Address:	127.0.0.1:10001			
Referrer Policy:	strict-origin-when-cross-origin			

响应状态：200, 404等

HTTP状态码

响应体，即网页源代码

```
1<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
2<html>
3<head>
4
5</head>
6<body>
7<font size="8px"><strong> <center>Introduction to Data Science </center></strong> <font>
8<font size="8px"><strong> <center>数据科学导论 </center></strong> <font>
9<br>
10<font size="6px"><strong> <center>课程代码: CS4043.01 </center></strong> <font>
11<font size="6px"><strong> <center>学院: 011计算机科学与技术学院 </center></strong> <font>
12<font size="6px">
13<strong>
14<center>课程教师:</center>
15<a href="http://staff.ustc.edu.cn/~huangzhy/" target=" -parent" >黄振亚</a>,
16<a href="http://staff.ustc.edu.cn/~cheneh/" target=" -parent" >陈恩红</a>
17</center>
18</strong>
19<font>
20<br><font size="6px">
21<strong>
22<center>上课时间: (1-18周)每周二下午第8、9节: 15:55-17:30, 教室: 高新校区 GT-B105 </center>
23</strong>
24<font>
25<hr size="1px" noshade>
26助教 (作业提交) :
27<br><dir><font size="5px">
28<li>肖桐, &nbsp; <a href="mailto:ds_intro2024@163.com">ds_intro2024@163.com</a></li><br>
29<li>QQ群: <a href="https://qm.qq.com/g/Fv9uQZVG8w">483794507</a> </li></font>
30</dir>
31<br>
32<hr size="1px" noshade>
33Recommended Textbooks:
34<br><dir><font size="5px">
35<li><a href="http://book.douban.com/subject/5377669/" target=" -parent" >《Introduction to Data Mining》</a></li>
36<li><a href="http://book.douban.com/subject/10590856/" target=" -parent" >《统计学习方法》</a></li>
37<li><a href="https://book.douban.com/subject/26708119/" target=" -parent" >《机器学习》</a></li>
38<li><a href="https://item.id.com/31931077735.html" target=" -parent" >《大数据导论》</a></li><font>
39</dir>
40<br><hr size="1px" noshade>
```



网络爬虫：载入

19

- 实际操作：抓取一个静态网页步骤
 - 首先确定URL，例如：<http://www.baidu.com>
 - 其次确定请求的方式以及相关参数：
 - 直接用浏览器实现：chrome, firefox浏览器抓包工具，详见
 - <http://jingyan.baidu.com/article/3c343ff703fee20d377963e7.html>
 - 或者抓包工具：charles等，详见
 - <http://blog.csdn.net/jiangwei0910410003/article/details/41620363/>
 - 最后在代码中按照特定的请求方式（get, post）向URL发送参数，即可收到网页的结果



网络爬虫：载入

□ 但部分页面的数据是动态加载的

□ Ajax异步请求

- 网页中的部分数据需要**浏览器渲染**(JavaScript调用接口获取数据)
- 用户的某些点击、下拉的**操作**触发才能获得

□ 解决方案：

- 借助抓包工具，分析Ajax某次操作所触发的请求，通过代码实现相应的请求

- 有技术难度，但抓取速度快。

- 利用智能化的工具：`selenium webdriver`

- 用**程序控制驱动浏览器**，模拟浏览器

- 可以**模拟实现人的所有操作**

- 操作简单，但是速度慢

- 因为爬虫需要启动浏览器，浏览器需要渲染页面，所以速度比较慢

- 其他：`Splash`, `Pyv8`等





网络爬虫：载入

21

- **反爬虫**: 随着网络爬虫对目标网站访问频率的加大，网站禁止爬虫程序继续访问
- 常见反爬手段：
 - 出现用户登录界面，需要验证码
 - 禁止某个固定帐号或ip一段时间内访问网站
 - 更有甚者，直接返回错误的无用数据
- 应对措施：
 - 优化爬虫程序，尽量减少访问次数，尽量不抓取重复内容
 - 使用多个cookie（网站用来识别用户的手段，每个用户登录会生成一个cookie）
 - 使用多个ip（可以用代理实现）

安全验证

您的账号可能存在安全风险，为了确保为您本人操作，请先进行安全验证。

发送成功

验证方式

186*****23手机

请输入六位数验证码 重新发送(57)

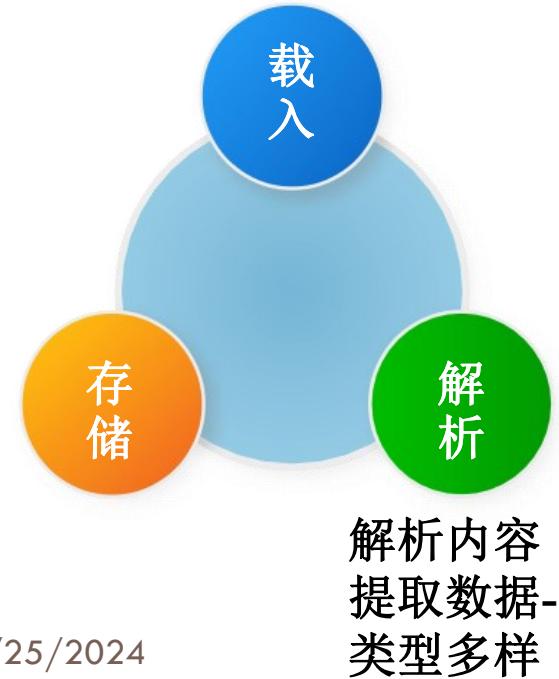
确定



网络爬虫：解析

22

- 解析：在载入的结果中**抽取特定的数据**，载入的结果主要分成三类html、json、xml
 - html
 - Java工具包：jsoup等
 - Python工具包：beautifulSoup等
 - json
 - Java工具包：json-lib、org-json、jackson等
 - Python工具包：json、demjson等
 - Xml
 - Java工具包：dom4j等
 - Java工具包：xml、libxml2等





网络爬虫：解析(对比JSON与XML)

23

```
{  
    "name": "中国",  
    "province": [  
        {"name": "黑龙江",  
         "cities": {  
             "city": ["哈尔滨", "大庆"]  
         }  
     },  
     {  
         "name": "广东",  
         "cities": {  
             "city": ["广州", "深圳", "珠海"]  
         }  
     },  
     ....  
    ]}
```

对象，成员：键值对

```
<?xml version="1.0" encoding="utf-8"?>  
<country>  
    <name>中国</name>  
    <province>  
        <name>黑龙江</name>  
        <cities>  
            <city>哈尔滨</city>  
            <city>大庆</city>  
        </cities>  
    </province>  
    <province>  
        <name>广东</name>  
        <cities>  
            <city>广州</city>  
            <city>深圳</city>  
            <city>珠海</city>  
        </cities>  
    </province>  
.....  
</country>
```



网络爬虫：解析(对比JSON与XML)

24

- 可读性
 - Json简洁， XML规范， xml比较好
- 可扩展性
 - 均很好
- 数据体积
 - Json数据量少， 传输快。 Xml数据量大， 传输慢
- 编码解码
 - Json容易， xml复杂（树结构， 父子节点）
- 数据描述
 - Xml数据描述更好
- 数据交互
 - Json与JavaScript交互更方便， 易于解析。 XML更适合跨平台共享



网络爬虫：抓取微博评论

25

邓超 8-17 20:49 来自 iPhone 7 Plus
跑男最新名单.....

344900 303031

转发 344900 评论 303031

陈赫 08-18 天霸

邓超 08-18 我们都很好，谢谢大家❤️

邓超 08-18 我也不知道🐱

贼亮z1 08-17 迪丽热巴😍😍

抓包工具
获取请求

▼ General

Request URL: <https://m.weibo.cn/api/comments/show?>
Request Method: GET
Status Code: 200 OK
Remote Address: 123.125.106.67:443
Referrer Policy: no-referrer-when-downgrade

► Response Headers (14)

▼ Request Headers [view source](#)

Accept: application/json, text/plain, */*
Accept-Encoding: gzip, deflate, br
Accept-Language: zh-CN,zh;q=0.8,en;q=0.6
Connection: keep-alive
Cookie: _T_WM=d9a7dba4dd130f79eaecac13c8906050; AL_bktAKLUXNkW1un7fu00CXjkppVYn1wGjJ3knF4g..; SUBP=0_p5NHD95Q0So5Re0.cS020Ws4Dqcjn-fHBxHzLxK-LB.eLBK5L505136002; M_WEIBOCN_PARAMS=featurecode%3D200003236170084375%26uicode%3D2000061%26fid%3D414183617
Host: m.weibo.cn
Referer: <https://m.weibo.cn/status/4141836170084375>
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 12.113 Safari/537.36)
X-Requested-With: XMLHttpRequest



网络爬虫：抓取微博评论

26

获得评论的json格式

皖ICP备15025187号-1 邮箱: service@json.cn

```
"mod_type": "mod/pagelist",
"previous_cursor": "",
"next_cursor": "",
"card_group": [
    {
        "id": 4142016554789113,
        "created_at": "08-18 08:46",
        "source": "柔光自拍vivo X7",
        "user": {"object": {...}},
        "text": "回复<a href="/n/%E9%82%93%E8%B6%85">邓超</a>:不管是谁,请大家记住陈赫的话,他们很好,感情都很好。恳请各家粉丝不要戏太多就好<i class="face face_1 icon_1">[微笑]</i>没准你们那么嫌弃骂的那么难听,人家正主还是感情好的时不时去吃火锅呢,你们不累吗?别用自己对他的爱去给他造成困扰",
        "reply_id": 4142015488402958,
        "reply_text": "<a href='/u/'5187664653>@邓超</a>: 我也不知道<i class="face face_1 icon_20">[doge]</i>",
        "like_counts": 10811,
        "liked": false,
        "mod_type": "mod/single/infoBox"
    }
]
```

解析出需要
的字段

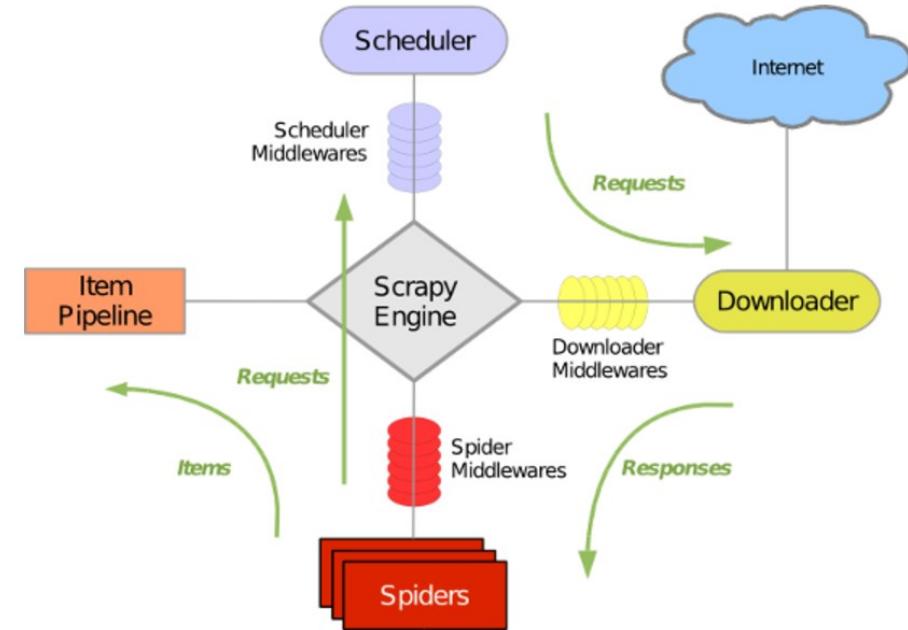
用户id	时间	内容
陈赫	08-18	天霸
邓超	08-18	我们都很好, 谢谢大家
邓超	08-18	我也不知道
贼亮zI	08-17	迪丽热巴
.....



网络爬虫：现有技术

27

- 基于Java的工具
 - HttpClient
 - Jsoup
- 基于Python的工具
 - Scrapy
 - Beautiful-soup



现有的爬虫框架很成熟，能够合理的控制爬取的过程，并有效的处理爬取过程中出现的各种异常，推荐使用Scrapy



网络爬虫：现有技术

28

■ ItSucks工具

- 支持通过下载模板和正则表达式来定义下载规则
- 提供swing GUI操作界面

■ Spidernet工具

- 以递归树为模型的多线程web爬虫程序
- 存储于sqlite数据文件

■ 完整解决方案

- 基于用户浏览器的爬虫（插件）
- 八爪鱼
- 火车采集器



火车采集器
网页数据采集利器



数据采集

29

- 注意网站规定
- 注意法律规定
 - 2021年6月1日，《中华人民共和国数据安全法》
- 注意数据使用规范
- etc

课达编程 (CODIA) · 隐私政策

引言

课达编程重视用户的隐私。您在使用我们的服务时，我们可能会收集和使用您的相关信息。我们希望通过本《隐私政策》向您说明，在使用我们的服务时，我们如何收集、使用、储存和分享这些信息，以及我们为您提供访问、更新、控制和保护这些信息的方式。本《隐私政策》与您所使用的课达编程服务息息相关，希望您仔细阅读，在需要时，按照本《隐私政策》的指引，作出您认为适当的选择。本《隐私政策》中涉及的相关技术词汇，我们尽量以简明扼要的表述，并提供进一步说明的链接，以便您的理解。

您使用或继续使用我们的服务，即意味着同意我们按照本《隐私政策》收集、使用、储存和分享您的相关信息。

如对本《隐私政策》或相关事宜有任何问题，请通过 codia-support@bdaa.pro 与我们联系。



数据采集：课后练习（不考核）

30

- 写一个爬虫，爬取一个感兴趣的网站
 - 如，课程主页：
<http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html>

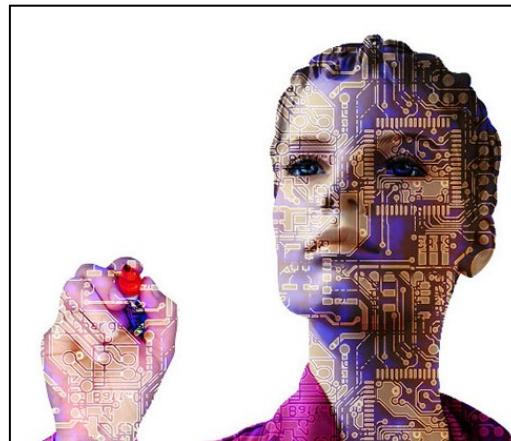


数据采集：自动生成数据

31

➤ AI: 1993 – 现在

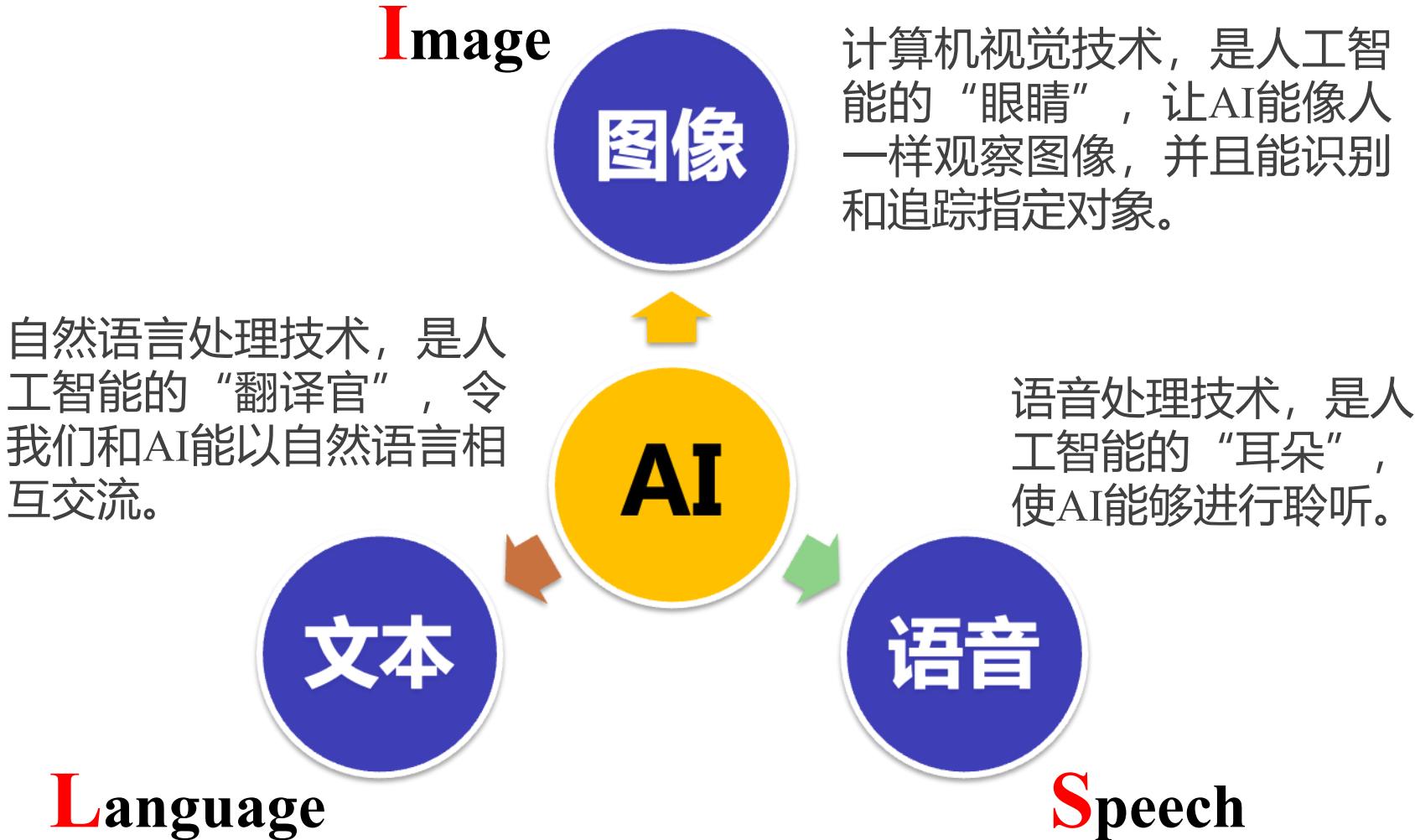
- 信息检索、互联网搜索
- 机器学习、自然会语言处理
- 数据挖掘、大数据分析
- “深蓝” 战胜人类国际象棋冠军卡斯帕罗夫





数据采集：自动生成数据

32

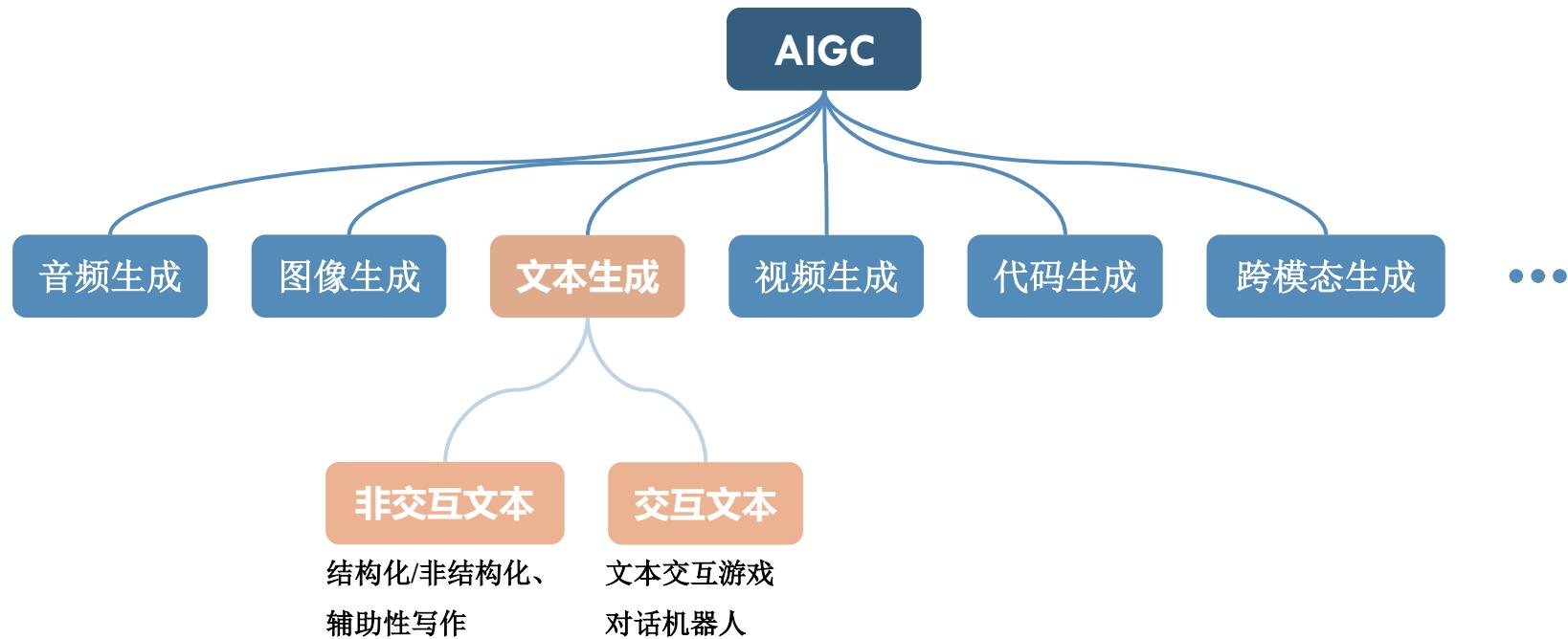




数据采集：自动生成数据

33

AI Generated Content (AIGC, 人工智能自动生成内容) 技术，是一种新型内容创作方式，可以在创意、表现力、迭代、传播、个性化等方面，充分发挥技术优势，打造新的数字内容生成与交互形态。





数据采集：自动生成数据

34

2022年11月30日，OpenAI 推出全新的对话式通用人工智能工具—**ChatGPT**。短短几天，其注册用户超过100万，引爆全网热议，成为历史上增长最快的消费者应用程序，掀起了人工智能领域大模型（参数量巨大的模型）技术巨浪。

ChatGPT 是一个聚焦于对话生成的大语言模型，其能够根据用户的文本描述，产生相应的智能回复。GPT通过学习大量网络已有文本数据，获得了像人类一样流畅对话的能力。

The image shows the official landing page for ChatGPT. On the left, there's a dark blue hexagonal icon containing a white OpenAI logo (a stylized brain). To its right, the word "ChatGPT" is written in a large, glowing blue font. Below this, a green button with the text "Try ChatGPT" and a small arrow is visible. To the right of the button is a section titled "Introducing ChatGPT" in white text. Underneath the title, a paragraph explains the model's conversational nature and its ability to handle follow-up questions, admit mistakes, challenge premises, and reject inappropriate requests. At the bottom of this section are two buttons: "Try ChatGPT" and "Read about ChatGPT Plus". The background features a dark gradient with horizontal stripes of varying shades of purple and pink.



数据采集：自动生成数据

35

国内外大模型产业竞争目前呈现“激烈迅猛”之势



OpenAI-ChatGPT

OpenAI于2022年11月30日发布的文本模态聊天机器人，被社会广泛关注。



讯飞-星火大模型



百度-文心一言



阿里-通义千问



腾讯-混元大模型



智谱-智谱清言



百川-百川大模型

国内大模型从大模型多重能力构建，多模态语义理解等角度快速发展，在多个领域超过ChatGPT。



数据采集：自动生成数据

36

ChatGPT 4 ▾



SA You

请以鲁迅的写作风格，写一段对春天的描述

ChatGPT

鲁迅先生的文风犀利深刻，同时带有一丝讽刺，他对于社会现象的描写往往深入骨髓。试以他的风格，来描述春天：

文本数据生成

春天，这个听起来应当充满生机的季节，总是在无声无息中悄然到来。你看，那嫩绿的芽尖透过泥土，似乎是在告诉人们，即使经历了漫长冷冽的冬日，生命依旧顽强地挣扎着要站直。田野里，那些归来的燕子在空中划过弧线，似乎是在嘲笑地上那些还未觉醒的生灵，它们或许并不理解，即使春光明媚，人间的烦恼和苦楚并未因此减少半分。

在这看似温暖的阳光下，农夫的背影显得更加佝偻，他们在田间辛勤地耕作，希望收获的不仅仅是庄稼，更是对未来的一线希望。而城市里，那些街道和小巷，则被春雨洗刷得略显清新，却也掩盖不了那永恒的尘埃和疲惫。春风并未带来太多温柔，反倒是像是在嘲讽那些还奢望着彻底变革的愚人们。

于是，春天不只是季节的轮回，它更像是一面镜子，折射出每一个人内心的希望与绝望。在这复苏的季节里，人们的心却不见得能与大自然一样，得以更新。

< 2 / 2 > ⏪ ⏴ ⏵ ⏹



Message ChatGPT...





数据采集：自动生成数据

37

科大讯飞 | 讯飞星火

讯飞智文

创建智能体 新建对话

星火对话 星火智文 星火科研助手 文章润色高级助手 智能体中心 群聊

一键生成PPT, 智能文档AI助理

请为我写一份半年度总结PPT →

策划一份新员工培训PPT →

准备一份电子产品介绍PPT →

非结构化数据生成

半年度工作总结报告

全面回顾与前瞻性展望

作者: 智文
汇报时间: 2024/09/14

CONTENTS 目录

01 引言 02 销售与市场分析 03 成本效益分析 04 学习与成长 05 挑战与机遇
06 经验教训总结 07 下半年工作规划 08 结语



数据采集：自动生成数据

38

图像数据生成

科大讯飞 | 讯飞星火

绘画大师

创建智能体 新建对话

星火对话

绘画大师

讯飞智文

星火科研助手

文章润色高级助手

智能体中心 群聊

历史对话

188xxxx0538 日程待办

API接入 星火官网

输入主题及要求，一键为你创作图像

阳光海滩，壮丽，充满色彩，电影镜头，完美光影

由星火AI生成

我已生成你所要的内容，希望你会满意！

退出智能体

在此输入您想了解的内容，输入"/"可获取模版，Shift+Enter换行

发送

内容由讯飞星火大模型生成，仅供您参考 用户协议 | 隐私政策 - 网信算备340104764864601230021号



数据采集：自动生成数据

39



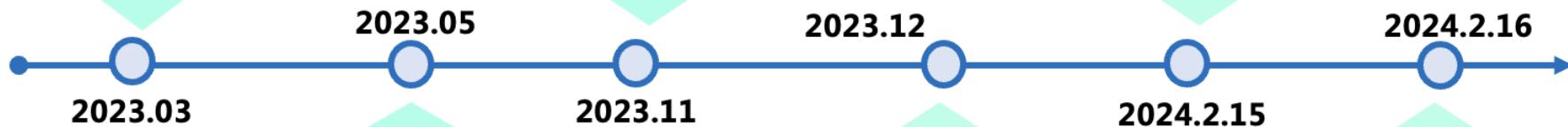
14日OpenAI发布多模态大模型
GPT-4，综合能力远超ChatGPT
OpenAI



OpenAI 7日凌晨发布GPT-4 Turbo，
六大升级，支持128k上下文输入，同时
推出OpenAI应用商店GPT Store



Google发布了新一代多模态大模型
Gemini 1.5 Pro，最高支持1000万上
下文长度，GSM8K评测全球第一



Meta
重磅发布LIMA 65B并于7月19日发布免费商
用开源模型Llama 2



Google
首次发布Gemini 1.0，多模态能力效果
惊艳，MMLU测试中首次超过人类专家



OpenAI
16日发布的文生视频模型，其效果显
著超越了业界现有同类产品的水平





数据采集：自动生成数据

40



【提示词】逼真的特写视频展现了两艘海盗船在一杯咖啡中激烈交战，它们在咖啡液中航行。



数据分析基础

42

- 数据采集 Data Collection
- 数据存储 Data Storage
- 数据预处理 Data Preprocessing
- 特征工程 Feature engineering

