



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

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/DS2025.html>



数据分析基础

2

□ 数据采集

Data Collection

□ 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



测量



数据库



监控



传统媒体

9/28/2025



数据采集

4

- 数据检索与开源
- 批量数据获取
 - 网络爬虫
- 数据生成



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
- 搜索引擎



9/28/2025



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

□ Google

Q GPT

上次更新日期 下载格式 使用权限 学科 提供方 免费

找到 100 多个数据集

gpt-dataset
huggingface.co
上次更新日期: Feb 22, 2023

kaggle GPT2 Output Dataset
kaggle.com
zip
上次更新日期: Nov 12, 2019

kaggle Paraphrased Articles using GPT-3
kaggle.com
zip
上次更新日期: Nov 12, 2019

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



数据采集：公开数据

7

□ 代表性公开数据集

- ImageNet : <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视频与社交网络: <http://netsg.cs.sfu.ca/youtubedata/>
- 用户评分MovieLens: <https://grouplens.org/datasets/movielens/>
- 网络数据-Large scale network: <https://snap.stanford.edu/data/>
- 教育数据:
 - ASSISTmentsData: <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

- 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





数据采集：网络爬虫

9

□ 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

存储

载入

发起请求(Request)
获得响应(Response)-类型多样

解析

解析内容
提取数据-类型多样

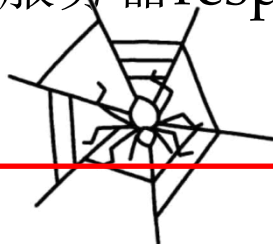


数据采集：网站数据

10

□ 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）



存储数据



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

11

□ 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/DS2025.html>
- 查看网页源代码

Introduction to Data Science 数据科学导论

课程代码：CS4023.01

学院：011计算机科学与技术学院

课程教师：黄振亚, 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：高新校区 GT-B212



数据采集：网站数据

12

□ 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/DS2023.html>
- 右键“检查”查看网页源代码



力换行 ☐

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
<head>

</head>
<body>
<font size="8px"><strong> <center>Introduction to Data Science </center></strong> <font>
<font size="8px"><strong> <center>数据科学导论 </center></strong> <font>
<br>
<font size="6px"><strong> <center>课程代码: CS4023.01 </center></strong> <font>
<font size="6px"><strong> <center>学院: 011计算机科学与技术学院 </center></strong> <font>
<font size="6px">
<strong>
  <center>课程教师:
  <a href="http://staff.ustc.edu.cn/~huangzhy/" target= "-parent" >黄振亚</a>,
  <a href="http://staff.ustc.edu.cn/~cheneh/" target= "-parent" >陈恩红</a>
  </center>
</strong>
<font>
<br><font size="6px">
  <strong>
    <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-B212 </center>
    </strong>
  </font>

<hr size="1px" noshade>
助教 (含作业提交):
<br><div><font size="5px">
  <li>程程, 于峻浩&nbsp;  <a href="mailto:ds_intro2024@163.com">data_science_2025@163.com</a></li><br>
  <li>QQ群: <a href="https://qm.qq.com/q/Fv9uQZVG8w">931835645</a> </li></font>
</div>

<hr size="1px" noshade>
Recommended Textbooks:
  <br><div><font size="5px">
    <li><a href="http://book.douban.com/subject/5377669/" target= "-parent" >《Introduction to Data Mining》</a></li>
    <li><a href="http://book.douban.com/subject/10590856/" target= "-parent" >《统计学习方法》</a></li>
    <li><a href="https://book.douban.com/subject/26708119/" target= "-parent" >《机器学习》</a></li>
    <li><a href="https://item.jd.com/31931077735.html" target= "-parent" >《大数据导论》</a></li></font>
  </div>
```



网络爬虫：载入

14

- 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

15

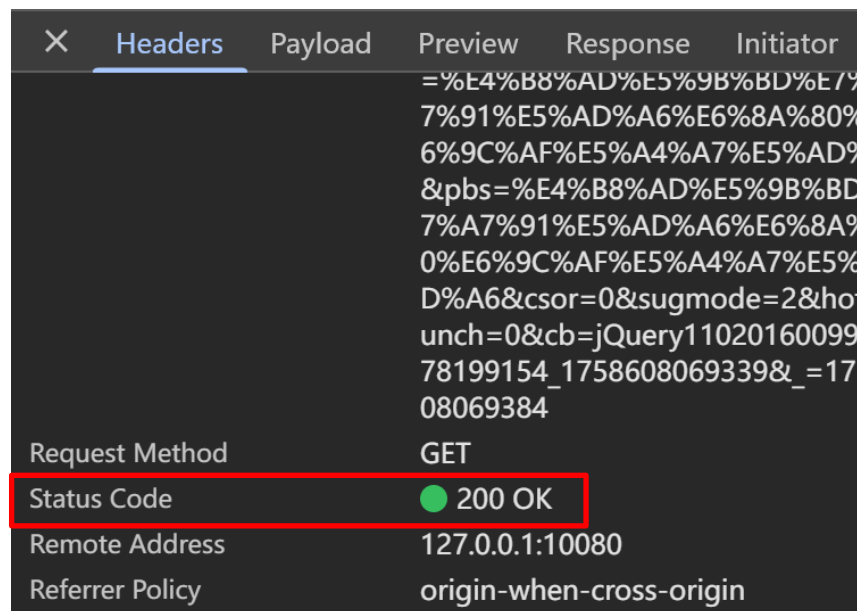
□ 载入：将目标网站数据下载到本地

□ 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?ie=[中国科学技术大学]`



9/28/2025



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

■ **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 用户名密码登录

手机号/用户名/邮箱

密码

登录

忘记密码?

扫码登录

立即注册

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: ddfac7e17ce70dc6187ce33dffee73ed

tpl: mn

username:

password:

JyR31zjQ6WKlg10sW0jf12!

3jz15ofHQtMTeUneEHAbKW

9/28/2025



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
 - 获得服务器的响应：Response，即获取网页源代码

响应头

▼ Response Headers View source

Bdpagetype: 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

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
<head>
</head>
<body>
<font size="8px"><strong> <center>Introduction to Data Science </center></strong> <font>
<font size="8px"><strong> <center>数据科学导论 </center></strong> <font>
<br>
<font size="6px"><strong> <center>课程代码: CS4023.01 </center></strong> <font>
<font size="6px"><strong> <center>学院: 011计算机科学与技术学院 </center></strong> <font>
<font size="6px">
<strong>
<center>课程教师:
<a href="http://staff.ustc.edu.cn/~huangzhy/" target="-parent">黄振亚</a>,
<a href="http://staff.ustc.edu.cn/~cheneh/" target="-parent">陈恩红</a>
</center>
</strong>
<font>
<br><font size="6px">
<strong>
<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-B212 </center>
</strong>
<font>
<br><font size="1px" noshade>
助教(含作业提交):
<br><dir><font size="5px">
<li>程, 于峻浩<br>
<li>QQ群: <a href="https://qq.qq.com/q/fv9uqZVG8w">931835645</a> </li></font>
</dir>
<br>
<font size="1px" noshade>
Recommended Textbooks:
<br><dir><font size="5px">
<li><a href="http://book.douban.com/subject/5377669/" target="-parent">《Introduction to Data Mining》</a></li>
<li><a href="http://book.douban.com/subject/10590856/" target="-parent">《统计学习方法》</a></li>
<li><a href="https://book.douban.com/subject/26708119/" target="-parent">《机器学习》</a></li>
<li><a href="https://item.jd.com/31931077735.html" target="-parent">《大数据导论》</a></li></font>
</dir>
```



网络爬虫：载入

18

- 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

19



□ 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

20

- **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

您的帐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

请输入六位数字验证码

重新发送(57)

确定

9/28/2025



网络爬虫：解析(对比JSON与XML)

21

```
{
  "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：抓取微博评论

22

**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zI**
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; ALbktAKLUXNkW1un7fu00CXjkppVYn1wGjJ3knF4g.; SUBP=0p5NHD95Q0So5Re0.cS020Ws4Dqcjn-fHBxHzLxK-LB.eLBK5L505136002; M_WEIBOCN_PARAMS=featurecode%3D200003236170084375%26uicode%3D20000061%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

9/28/2025



网络爬虫： 抓取微博评论

23

获得评论的json格式

```
京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": {
        "id": 4142016554789113,
        "name": "陈赫",
        "avatar": "http://img1.img-cdn.dcloud.net.cn/avatar/20150818/08/46/4142016554789113.jpg"
      },
      "text": "回复
```

解析出需要的
字段

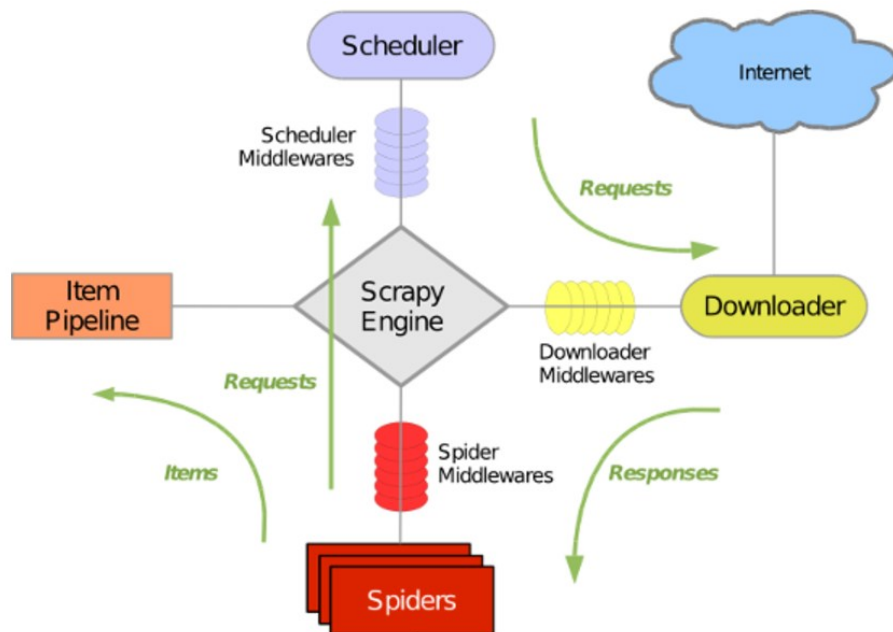
用户id	时间	内容
陈赫	08-18	天霸
邓超	08-18	我们都很好，谢谢大家
邓超	08-18	我也不知道
贼亮zl	08-17	迪丽热巴
.....		



网络爬虫：现有技术

24

- 基于Java的工具
 - HttpClient
 - Jsoup
- 基于Python的工具
 - **Scrapy**
 - BeautifulSoup



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



数据采集

25

- 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, 以及我们为您提供的访问、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）

26

- 写一个爬虫，爬取一个感兴趣的网站

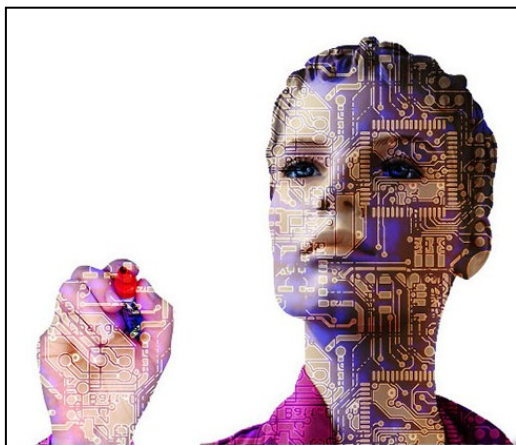


数据采集：自动生成数据

27

➤ AI：1993 – 现在

- ❑ 信息检索、互联网搜索
- ❑ 机器学习、自然语言处理
- ❑ 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



9/28/2025



数据采集：自动生成数据

28



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

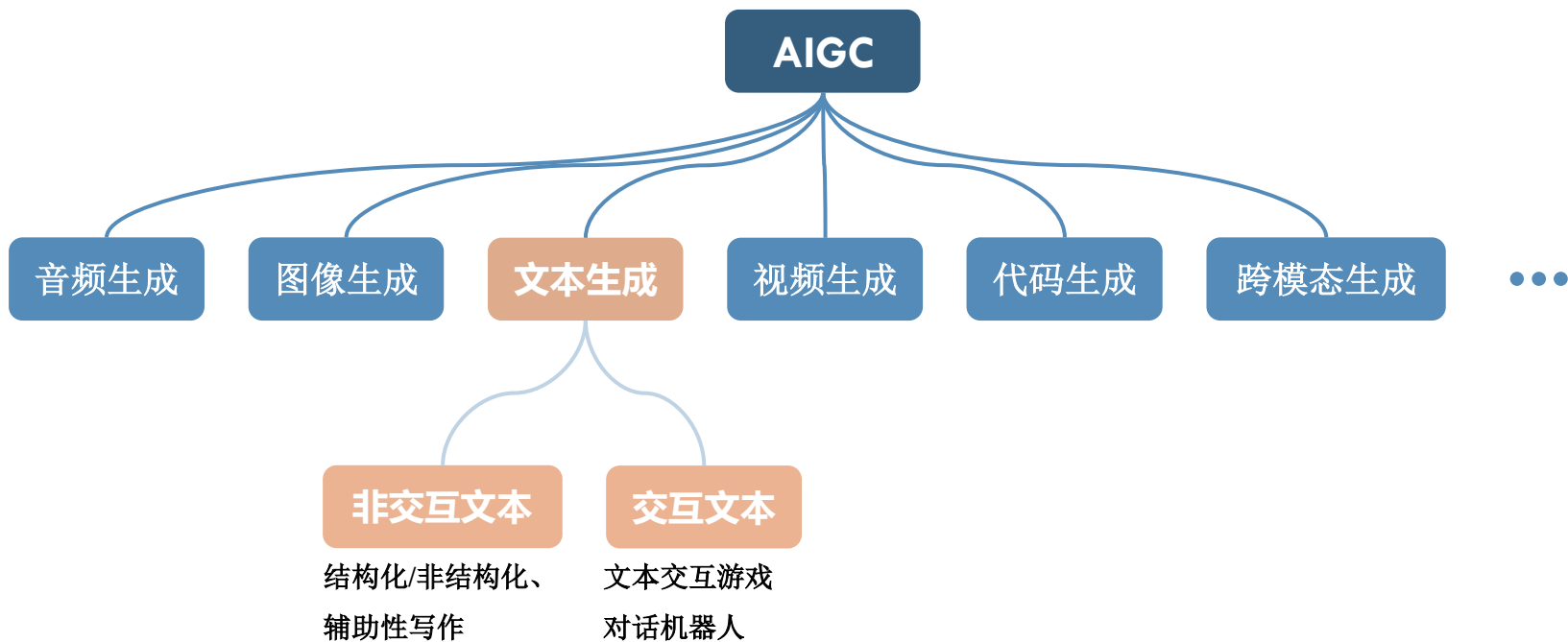
国内大模型从大模型多重能力构建，多模态语义理解等角度快速发展，在多个领域超过ChatGPT。



数据采集： 自动生成数据

29

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，可以定制化生成需要的数据。





数据采集：自动生成数据

30

ChatGPT 4



文本数据生成



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：

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，春天并不只是季节的轮回，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

31

创建智能体

新建对话

星火对话

讯飞智文

星火科研助手

文章润色高级助手

智能体中心

讯飞智文

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

32

科大讯飞
iFLYTEK

讯飞星火
XUNFEI XINGHUO

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



绘画大师

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

我已经生成你所要的内容，希望你合意！

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

图像数据生成



数据采集：自动生成数据

33



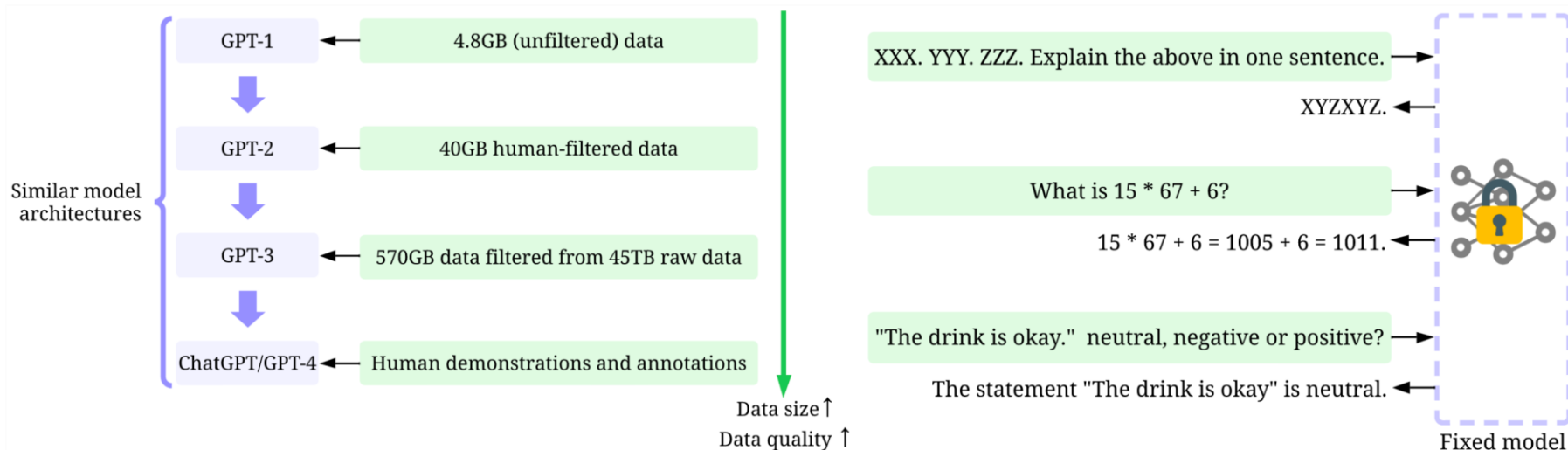
【提示词】逼真的特写视频展现了两艘海盗船在一杯咖啡中激烈交战，它们在咖啡液中航行。



数据科学基础

34

- 人工智能逐渐从以模型为中心过渡到以数据为中心
 - **GPT成功的数据基石：** GPT进化中，模型结构保持相似，训练数据的规模、质量得到极大提升
 - 模型对数据的需求急速上升





大模型的扩展定律

Scaling law

- 损失值随着模型规模、训练数据量、训练计算量的增加，呈现出幂律下降的趋势

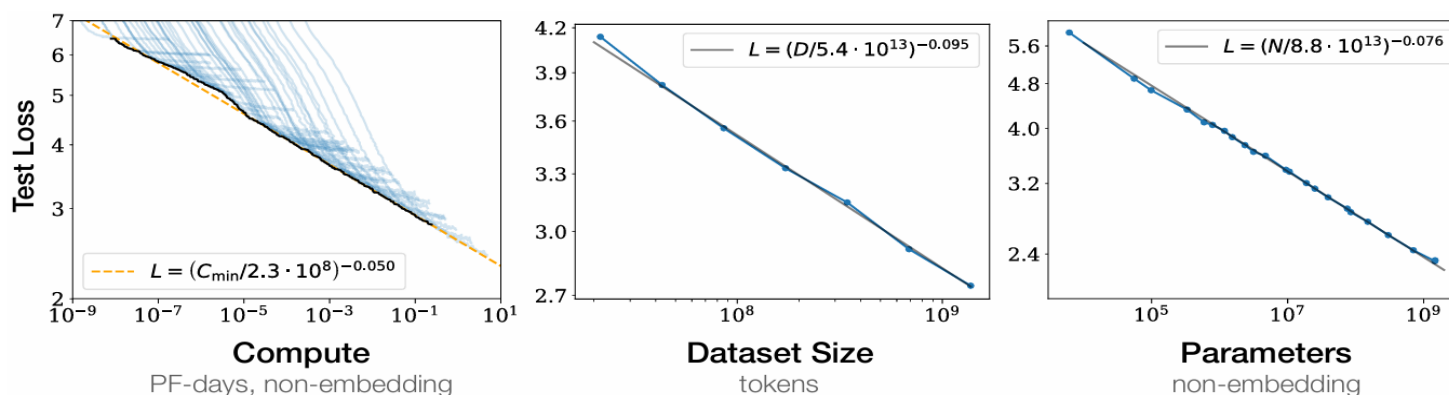


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$$



大模型的扩展定律

□ 通义千问：多样高质量训练数据

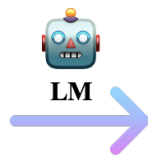
- **数据过滤**：使用大模型多维度数据分析和过滤
- **数据合成**：使用大模型合成和过滤高质量训练数据
- **数据混合**：平衡低质量社交、电商数据 vs 高质量的技术、学术类数据

增加数据量

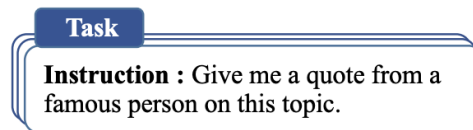
175 seed tasks with
1 instruction and
1 instance per task



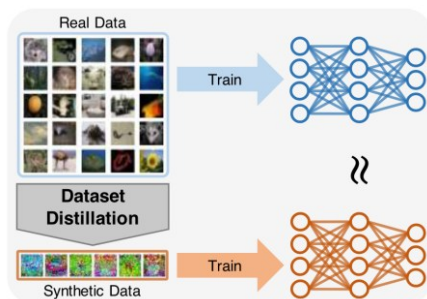
Task Pool



数据生成



改善数据质量



数据蒸馏



数据选择



数据分析基础

37

□ 数据采集

Data Collection

□ 数据预处理

Data Preprocessing

□ 特征工程

Feature engineering

