工作任务二：数据采集

（一）认识数据

1、 确定数据类型是数据采集前的关键步骤。常见数据类型包括：

（数据的存在方式可以是文字、模型、多媒体、软件语言或特定学科规范的特定格式。在数据新闻的生产过程中，常见的数据格式包括**文本格式（.txt）**、**电子表格式（.xls或.xlsx）**、**逗号分隔符文件格式（.csv）**、**可扩展标记语言格式（XML）**、**轻量数据交换格式（.json）**、**地图数据格式（.shp）**等。

常见文本数据格式一览表

格式名称	扩展名	说明	打开方式
HTML	.htm、 .html、.shtml、.xhtml	带超本命格式的文本，网页	可以使用网页制作工具加Microsoft SharePoint Designer、SeaMonkey、Dreamweaver等打开，同时以源代码形式展示出来，并可编辑其内容。
TeX	.tex、ltx	TeX文件，可以是Plain TeX、LaTeX等	WinEdt、TeXworks等专用编辑器，Vim、Emacs、Gedit、Kate等通用文本编辑器
DOC	.doc、.docx	Word文档	Word、Docx需要Word2017、金山WPS Office、OpenOffice等可以部分兼容
PPT	.ppt、.pptx	PowerPoint演示文稿	PowerPoint、pptx需PowerPoint2010、金山WPS Office、OpenOffice等可以部分兼容
WPS	.wps	WPS文档	金山WPS Office的文字处理程序的文件格式，通MS Word的DOC结构一致，与Microsoft Office的Word的组件兼容，但是需要将扩展名修改为.doc
PDF	.pdf	可携式文件	Adoe Reader

（一）认识数据

数据就是数值，是通过观察、实验或计算得出的结果。数据只有在被使用的过程中才能转化为信息，并通过与故事结合呈现其意义。**大数据**是指那些规模巨大、复杂多变、难以用常规数据库和软件工具进行管理和处理的数据集合。

选用数据主要从以下角度出发：

1.用户需求：从用户角度出发，考虑他们可能关注的数据点，以及这些数据如何帮助他们思考和做出决策。

2.数据质量：考虑数据的准确性、完整性、一致性等因素，确保收集到的数据具有高质量。

3.合规性：确保数据收集和使用符合相关法律法规的要求，如隐私政策、数据保护法等。

（二）数据采集

1、在数据化社会中，数据确实无处不在，《数据新闻手册》列出的7种数据查找方法为我们提供了全方位的数据获取途径。

（二）数据采集

2、数据查找中最常用的工具是搜索引擎和各类开放数据：

（1）**搜索引擎：**在网络搜索框中输入想要查找的内容直接搜索，能帮助我们提供搜索效率，下面介绍几种常用的技巧。

使用准确的关键词：选择与你要查找的内容直接相关的关键词。使用具体的词汇代替模糊的词汇，例如“大学生饮食消费”而非“饮食”。注意避免使用常见的停用词，如“的”、“是”、“在”等，它们对搜索结果的影响很小。

在特定网站内搜索：如果你想在特定网站内搜索信息，可以在搜索框中输入“site:网站域名 关键词”。例如，在知乎上搜索关于Python的问题，可以输入“site:zhihu.com Python”。

（二）数据采集

3、数据查找中最常用的工具是搜索引擎和各类开放数据：

(2)**开放数据：**开放数据是一类可以被任何人免费使用、再利用、再分发的数据。 数据开放平台的建设不仅促进了政府数据资源的共享，也推动了市场和社会主体开发应用数据。截至2023年8月，我国已有226个地方政府上线了数据开放平台，数量增长迅速。

例如：**百度指数、微信指数、艾瑞咨询、艾媒咨询、36氪、CNNIC**等，提供关键词搜索热度趋势，适合做需求洞察、用户画像等。**国家统计局、工业和信息化部、中国人民银行、银监会**等官网数据平台，可提供垂直类数据信息。

（二）数据采集

问卷中的问题序列会影响到被调查者对问题的回答，甚至影响到调查的顺利进行。安排好问题的顺序一般有以下常用的规则：

1.个人背景资料可放前面，也可放后面。

2.按照一定的逻辑顺序排列问题。

3.把简单易答的问题放在前面，把复杂难答的问题放在后面。

4.把能引起被调查者兴趣的问题放前面，把容易引起他们紧张、顾虑的问题放在后面。

5.把行为方面的问题放在前面，把态度、意见方面的问题放在后面。

6.把开放式问题放在后面。

（二）数据采集

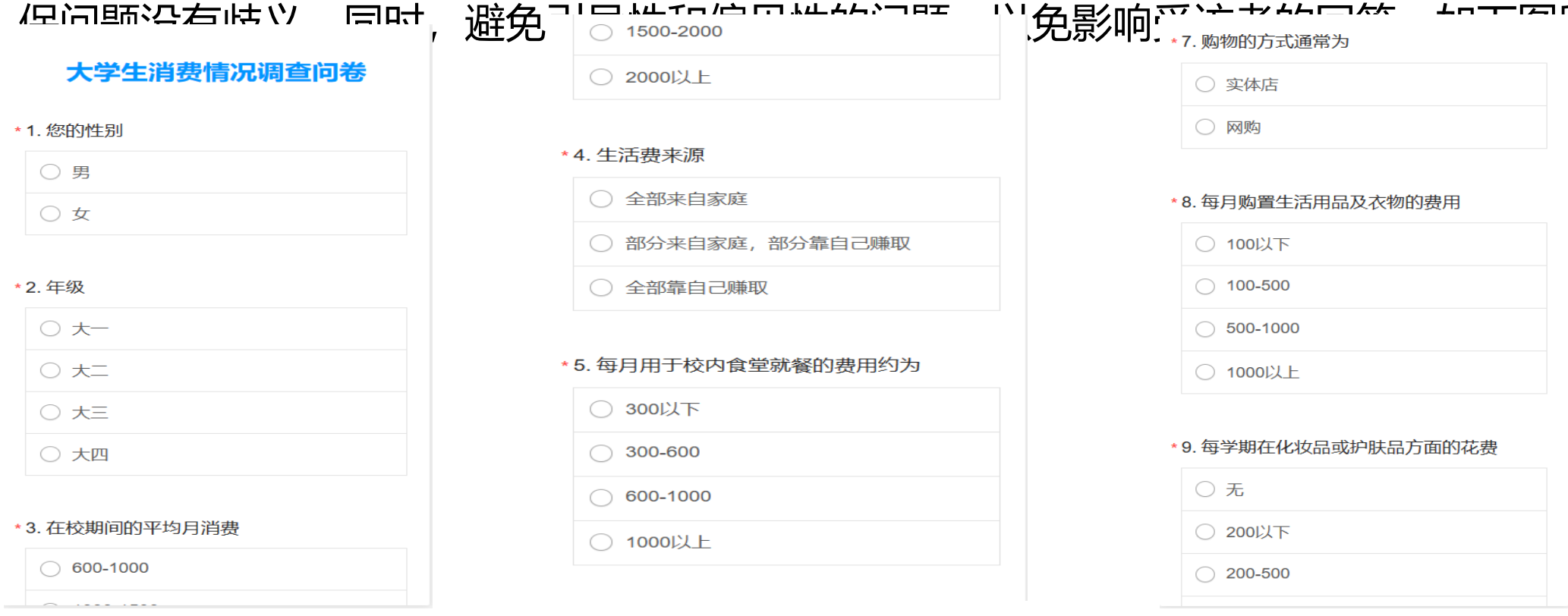
第一步，注册。登录“问卷星”官网，完成注册。

第二步，创建问卷。在登录后，网页左边为“创建问卷”操作区。可单击创建问卷，根据自己需求设计好问卷内容

（二）数据采集

例如创建一个关于“大学生消费情况”的调查问卷。在创建问卷时，我们需要明确调查的目的，比如了解大学生的**消费结构、消费观念、消费习惯等**。这将有助于设计问卷的具体问题和内容。

设计问题时，要确保问题的表述清晰、简洁、易于理解。避免使用过于专业或复杂的术语，确保问题没有歧义。同时，避免[illegible]免影响[illegible]所示：

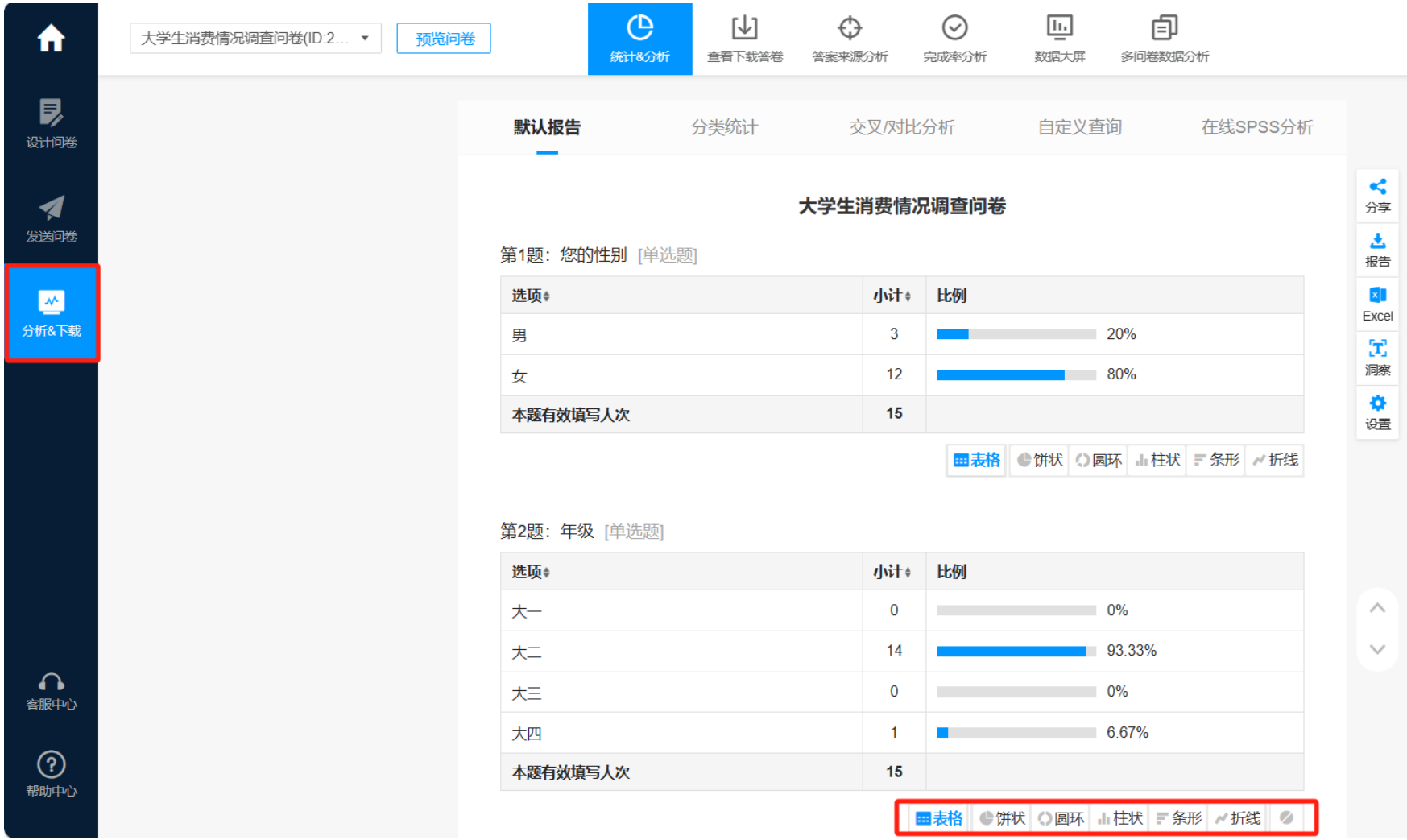


（二）数据采集

第三步，完成问卷的创建之后，发出问卷链接。单击“发送问卷”，菜单栏中有发送“问卷链接与二维码”，将问卷链接或二维码推送给被调查对象即可。

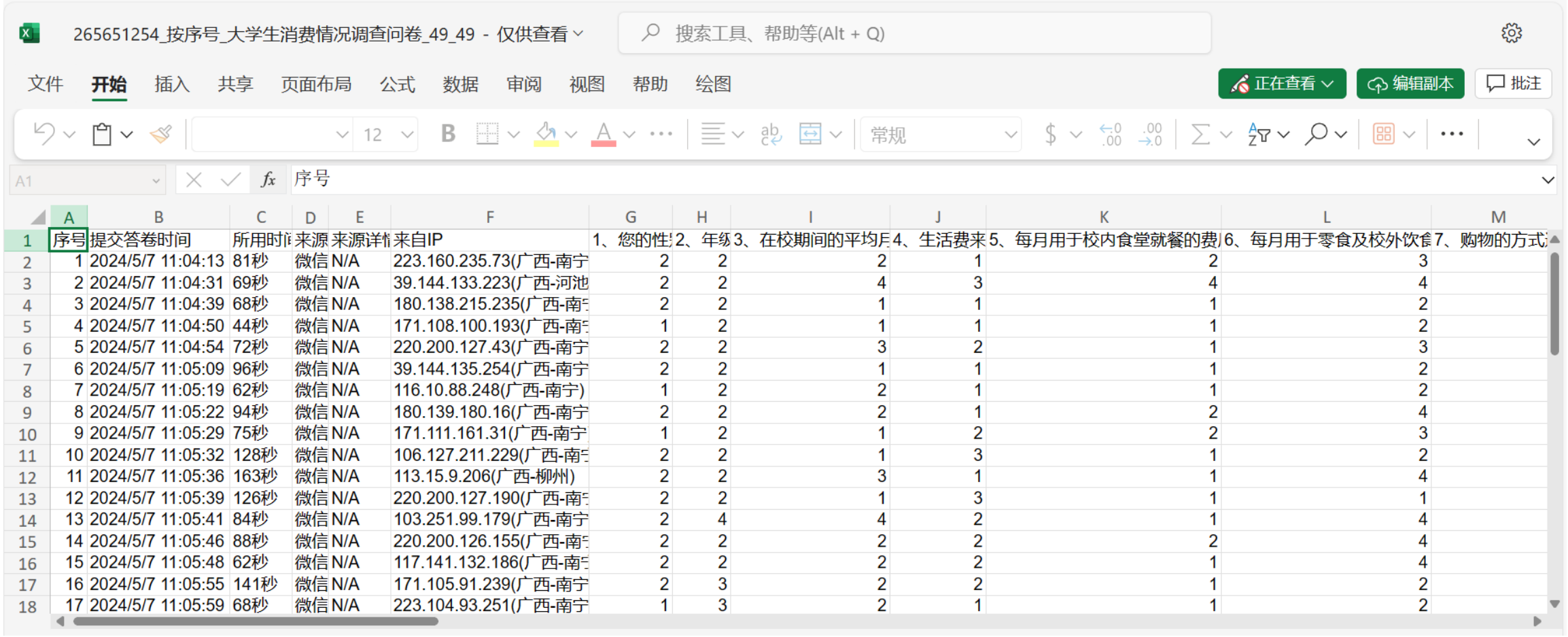
（二）数据采集

第四步，回收答卷。点击左边菜单栏的“分析下载”，单击它即可看到分析结果。



（二）数据采集

第五步，点击“查看下载问卷”。如图所示，我们可以下载Excel文件获得调查结果。



265651254_按序号_大学生消费情况调查问卷_49_49 - 仅供查看

搜索工具、帮助等(Alt + Q)

文件 开始 插入 共享 页面布局 公式 数据 审阅 视图 帮助 绘图 正在查看 编辑副本 批注

A1 序号

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	序号	提交答卷时间	所用时间	来源	来源详情	来自IP	1、您的性别	2、年级	3、在校期间的平均月	4、生活费来	5、每月用于校内食堂就餐的费用	6、每月用于零食及校外饮食	7、购物的方式
2	1	2024/5/7 11:04:13	81秒	微信	N/A	223.160.235.73(广西-南宁	2	2	2	1	2	3	
3	2	2024/5/7 11:04:31	69秒	微信	N/A	39.144.133.223(广西-河池	2	2	4	3	4	4	
4	3	2024/5/7 11:04:39	68秒	微信	N/A	180.138.215.235(广西-南	2	2	1	1	1	2	
5	4	2024/5/7 11:04:50	44秒	微信	N/A	171.108.100.193(广西-南	1	2	1	1	1	2	
6	5	2024/5/7 11:04:54	72秒	微信	N/A	220.200.127.43(广西-南宁	2	2	3	2	1	3	
7	6	2024/5/7 11:05:09	96秒	微信	N/A	39.144.135.254(广西-南宁	2	2	1	1	1	2	
8	7	2024/5/7 11:05:19	62秒	微信	N/A	116.10.88.248(广西-南宁)	1	2	2	1	1	2	
9	8	2024/5/7 11:05:22	94秒	微信	N/A	180.139.180.16(广西-南宁	2	2	2	1	2	4	
10	9	2024/5/7 11:05:29	75秒	微信	N/A	171.111.161.31(广西-南宁	1	2	1	2	2	3	
11	10	2024/5/7 11:05:32	128秒	微信	N/A	106.127.211.229(广西-南	2	2	1	3	1	2	
12	11	2024/5/7 11:05:36	163秒	微信	N/A	113.15.9.206(广西-柳州)	2	2	3	1	1	4	
13	12	2024/5/7 11:05:39	126秒	微信	N/A	220.200.127.190(广西-南	2	2	1	3	1	1	
14	13	2024/5/7 11:05:41	84秒	微信	N/A	103.251.99.179(广西-南宁	2	4	4	2	1	4	
15	14	2024/5/7 11:05:46	88秒	微信	N/A	220.200.126.155(广西-南	2	2	2	2	2	4	
16	15	2024/5/7 11:05:48	62秒	微信	N/A	117.141.132.186(广西-南	2	2	2	2	1	4	
17	16	2024/5/7 11:05:55	141秒	微信	N/A	171.105.91.239(广西-南宁	2	3	2	2	1	2	
18	17	2024/5/7 11:05:59	68秒	微信	N/A	223.104.93.251(广西-南宁	1	3	2	1	1	2	

（三）数据转换

1、数据转化具有以下作用：

（1）**兼容性和可访问性：**不同的软件、应用程序或系统可能支持不同的文件格式。为了在不同平台或软件之间共享数据，通常需要将数据转换格式，以确保数据能够被正确地读取和解析。

（2）**数据分析和处理：**某些数据分析工具或软件可能仅支持特定的文件格式。为了使用这些工具进行分析，需要将数据转换为相应的格式。此外，不同的文件格式可能具有不同的数据结构和组织方式，将数据转换为更适合的格式可能使数据分析和处理更加高效。

不同类型的数据格式，需要用不同的数据数据工具完成转换。常见的如网页形式的文本型数据，需要用**open refine或者tabletools2**转换为表格；PDF形式的数据需要用tabula提取数据转换为逗号分隔符csv格式；纸质文本型数据，需使用文本提取工具OCR等等。

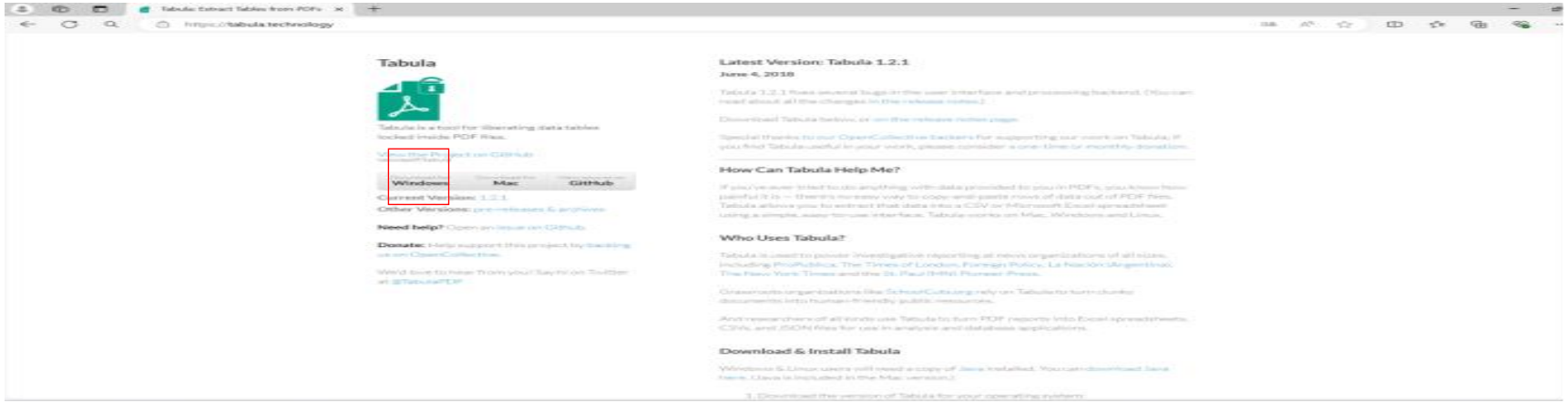
（三）数据转换

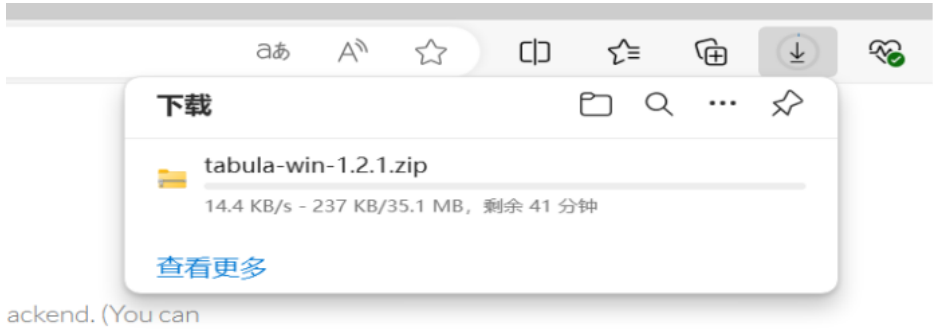
2、Tabula 是一款用于提取 PDF 文档中的表格的开源软件。

首先，先进入 Tabula 官网。

（三）数据转换

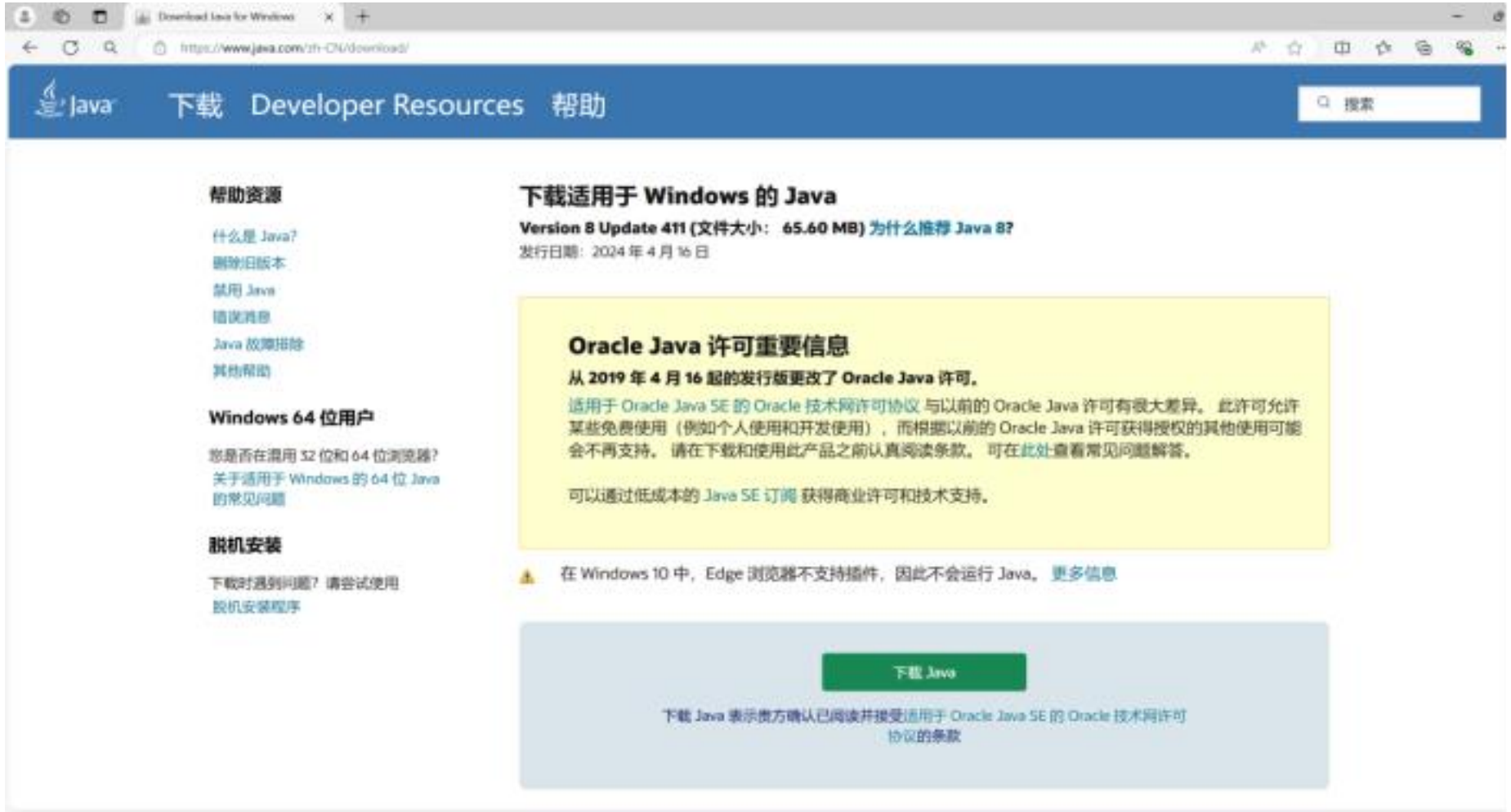
目前，本软件最新版本是 1.2.1，发布于 2018 年 6 月。Tabula 可运行在 Windows、macOS、Linux 等操作系统上。例如，如果当前你的系统是 Windows 操作系统，可点击官网的“Download for Windows”按钮，直接下载。





（三）数据转换

Tabula 依赖于 Java 运行。如果你的电脑上未安装 Java 运行时，则首次运行该软件时，Tabula 会自动打开 Java 官网的下载页面：https://www.java.com/zh-CN/download/



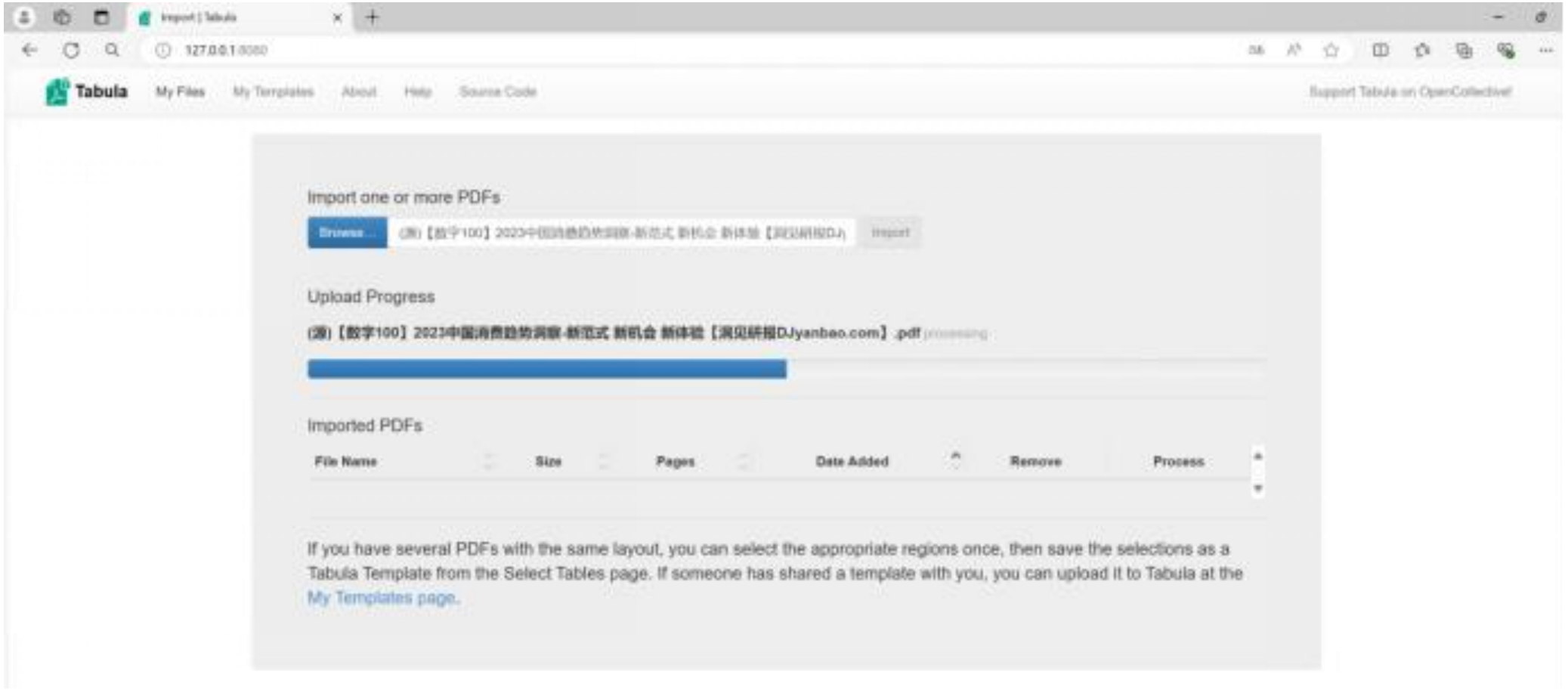
（三）数据转换

下载 Java 后，点击安装即可。

下载完毕后，解压压缩包tabula-win-1.2.1.zip，解压后直接点击应用程序tabula。电脑的默认浏览器会自动打开 http://127.0.0.1:8080/ 。如下图。

（三）数据转换

首先，点击 Browse...按钮选择要处理的PDF文件。点击Import 按钮即可完成文件导入。通过tabula进度条可以看到导入进度。

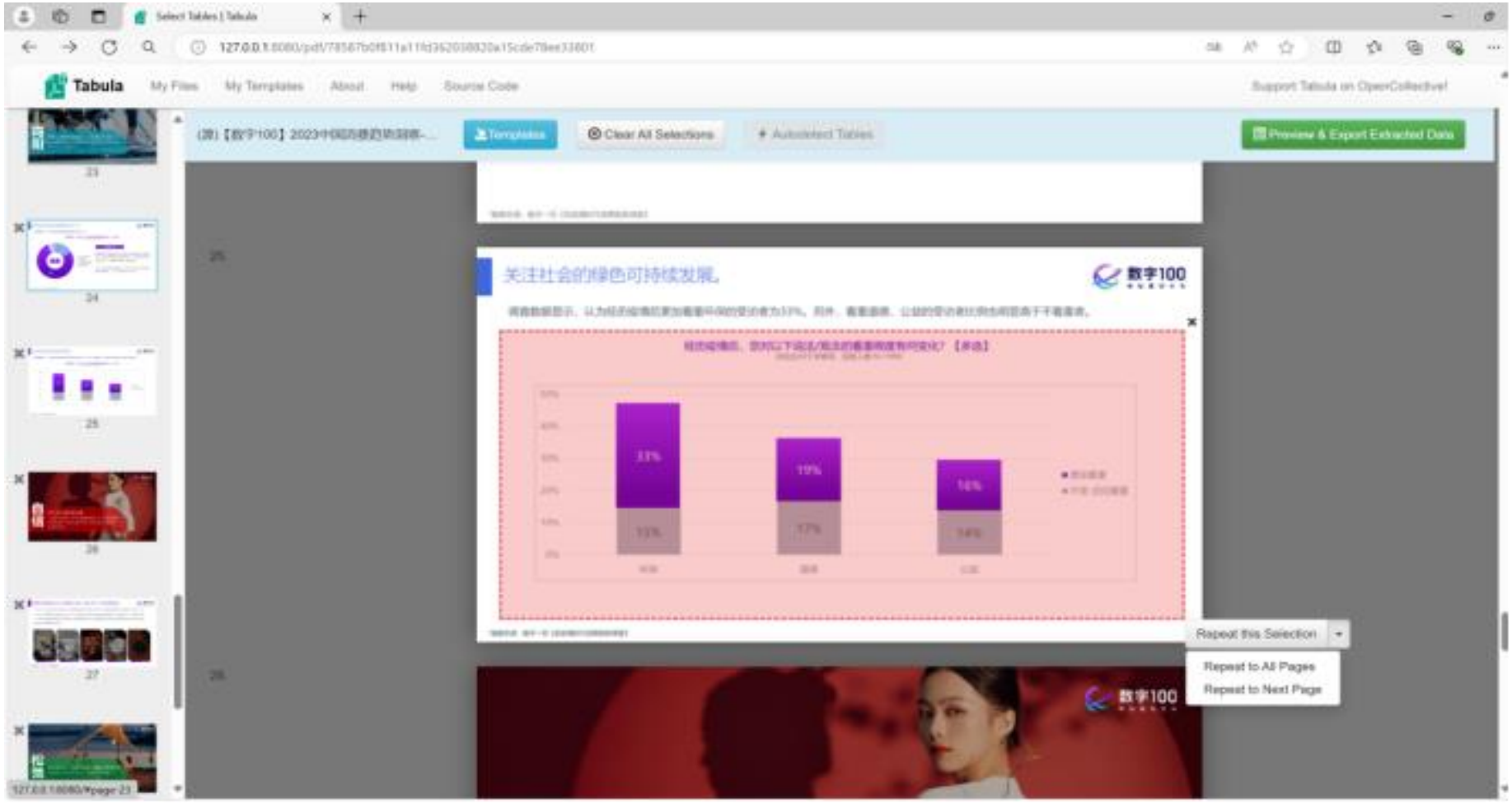


（三）数据转换

导入完成后，即可看到类似如下的界面：

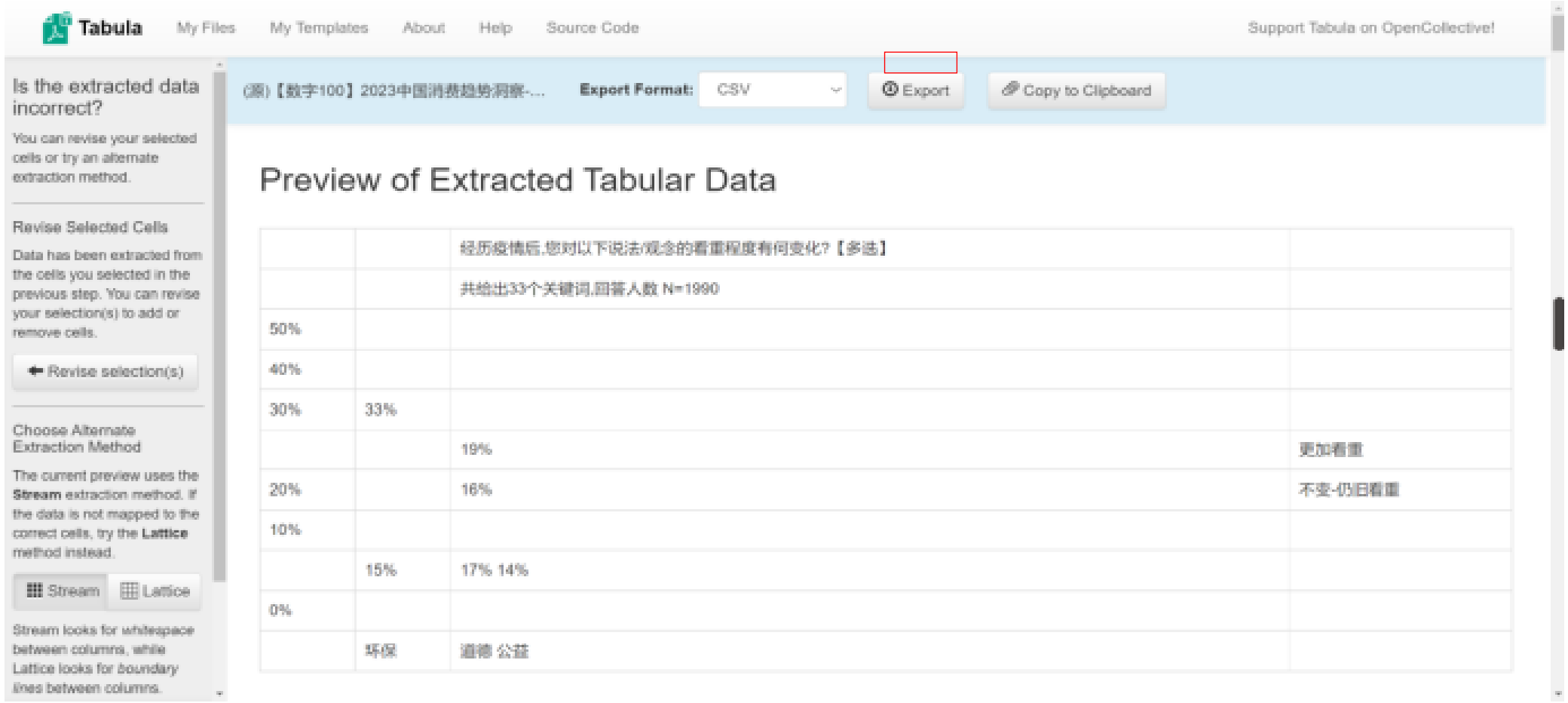
（三）数据转换

框选出包含表格的区域。将鼠标指针挪至待提取表格的左上角，按下鼠标左键不放，然后拖拽鼠标，直至将表格都框选起来。



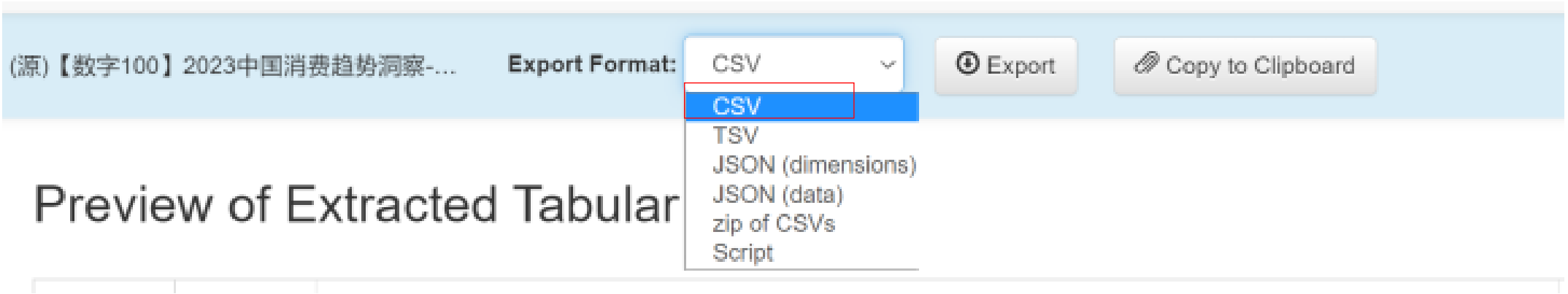
（三）数据转换

当所需表格都被框选完后，点击当前页右上方选项 Preview & Export Extracted Data，来预览和导出从被框选的表格中提取出来的数据。如下图，提取出来的数据以表格形式呈现了出来。



（三）数据转换

点击表格上方的Export选项。通过左侧的下拉框来选择导出格式。.csv文件可以用Microsoft Office 直接打开。为了便于后续操作可以用 Excel软件打开.csv文件，并导出为.xls或.xlsx文件。



（三）数据转换

3、当数据为纸质版材料时，可使用OCR文本提取工具将纸质数据转换电子文本格式。以下用易飞文字识别小程序举例，主要步骤如下：

第一步，打开微信，搜索“易飞文字识别”小程序，点击进入主页面，选择“图片转文字”选项。

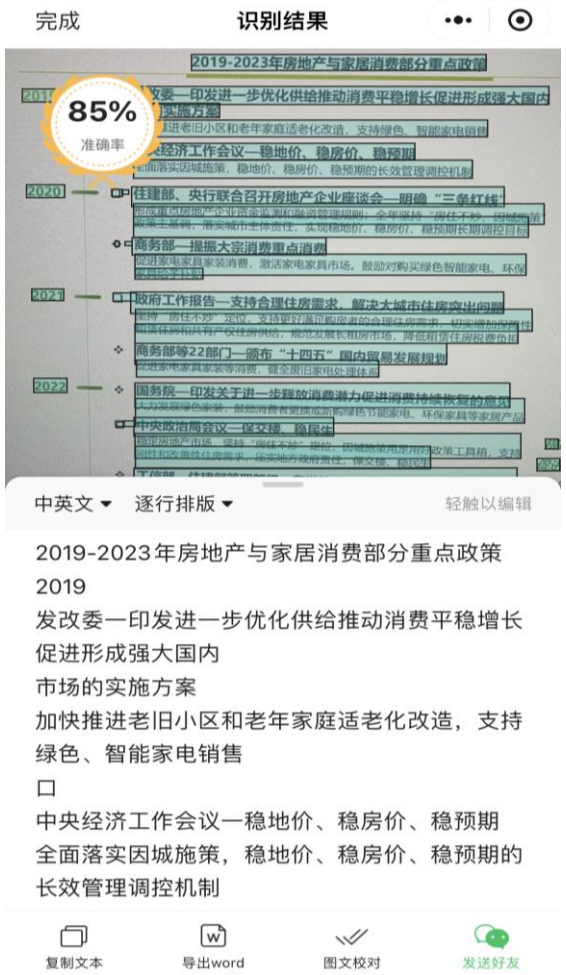
（三）数据转换

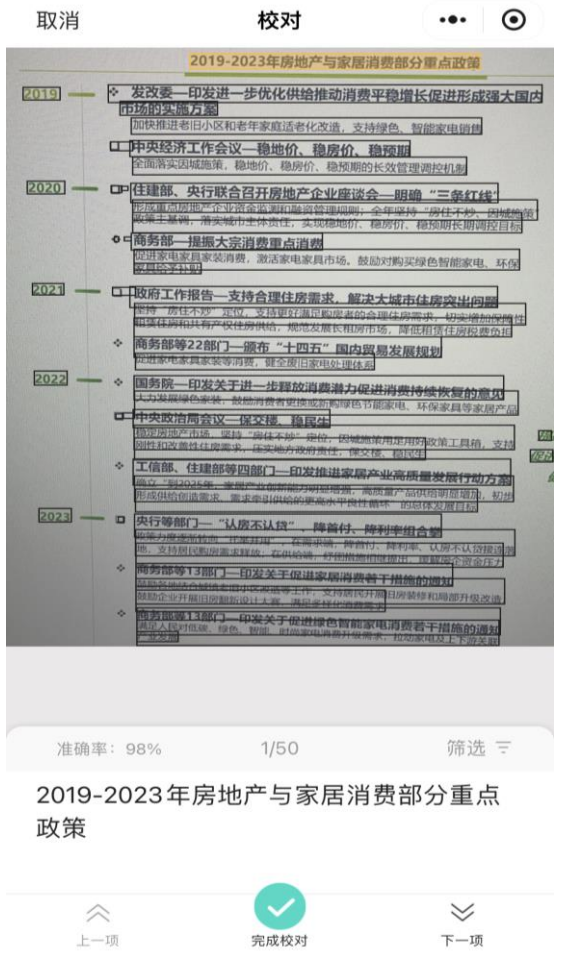
第二步，页面跳转出现“拍照” “相册选图” “微信聊天图片”三个选项，相应选择。

（三）数据转换

第三步，识别完成后，点击“图文校对”，进一步核对文本内容。

第四步，核对完成后，点击“完成校对”，导出文件。





思政融入：

数据的真实性需兼顾局部与整体，要求局部与全局都准确无误。首先，权威的数据支撑与科学的运用策略，是数据新闻效果的关键。首先，需善于借助开放的数据平台，注重自身长期积累的高质量数据库的构建。其次，面对纷繁复杂的数据，需要具备梳理其逻辑脉络，洞察其间联系的能力。另外，还需在使用数据时遵循法律法规，兼顾数据开放需求、数据安全保障与知识产权维护等因素之间的平衡点。你认为在数据采集的过程中，我们应具备哪些能力素养？

思政元素：理性思维、实证探索

谢谢观看