**实验六 文本数据可视化**

**一、本次实验的目的**

通过该实验的实践，要求学生可以理解文本数据，掌握绘图工具的使用过程，能够熟练绘制常见的适合展示文本数据的可视化方法。掌握文本数据可视化的技巧，提升对大数据分析的理解，为今后的数据驱动决策提供强有力的支持。在实际应用中，可以将所学应用于舆情分析、市场趋势研究、用户行为洞察等多个领域。

**二、实验相关知识**

文本数据通常包括电子邮件、社交媒体帖子、新闻报道、评论等各种非结构化信息。由于其复杂性和数量庞大，单纯的文字分析往往难以捕捉到关键信息。此时，可视化工具和方法就显得尤为重要。以下将详细阐述实验涉及的几个核心知识点：

1.文本预处理：这是处理文本数据的第一步，包括去除停用词（如“的”、“是”等常见词汇）、词干提取（将单词还原为其基本形式）和标点符号清理等。这些步骤有助于减少噪声，提高后续分析的效率。

2.词云：词云是最常见的文本可视化形式之一，通过调整不同词语的大小和颜色，直观展示文本中高频词汇。在实验中，使用Python的WordCloud库来创建词云图，以揭示文本的主要话题。

3.词频统计：统计每个单词在文本中的出现次数，可以发现文本的主题分布。使用如NLTK或spaCy等自然语言处理库，可以方便地进行词频统计。

4.语义分析：包括情感分析和主题建模。情感分析用于识别文本中的情绪倾向，例如积极、消极或中立；主题建模（如LDA，Latent Dirichlet Allocation）则能找出文本中隐藏的主题结构。

5.网络：当文本数据包含实体关系时，网络图是有效的表示手段。例如，人物之间的交互、关键词之间的关联等。使用Gephi或NetworkX等工具，可以构建并可视化这些关系网络。

6.时间序列分析：对于具有时间戳的文本数据，如微博或新闻，可以通过时间线或热力图来展示随时间变化的趋势和模式。

7.可视化工具和库：实验可能涉及到的Python库有matplotlib、seaborn、bokeh、plotly等，它们提供了丰富的图形选项和交互性功能。此外，Tableau、PowerBI等商业可视化工具也是处理大数据可视化的有力助手。

8.交互式可视化：为了更深入地探索和理解数据，交互式可视化允许用户自定义视图，如选择显示的维度、过滤条件等。D3.js是一个强大的JavaScript库，常用于创建此类复杂的交互式图表。文本数据可视化"这一主题，聚焦于如何通过视觉表示来揭示文本数据的内在结构和模式。

9.本次实验用到的中文分词库jieba库：

Jieba库分词的三种模式：

Jieba库是一个优秀的Python第三方中文分词库，它支持多种分词模式，并且可以自定义词典以提高分词的准确性。以下是Jieba库的一些基本介绍和使用方法：

（1） 基本介绍

功能：Jieba库提供中文文本分词功能，将中文文本切分成词语或词汇单位，便于后续的自然语言处理（NLP）和文本分析任务。

**特性**：

支持三种分词模式：精确模式、全模式和搜索引擎模式。

支持自定义词典，用户可以根据需要添加或删除特定的词汇，提高分词的准确性。

提供了关键词提取功能，有助于文本摘要和信息检索等任务。

实现了基于前缀词典和动态规划的分词算法，分词速度快且准确率高。

（2）分词模式

**精确模式**：试图将句子最精确地切开，适合文本分析。

**全模式**：把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义。

**搜索引擎模式**：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

（3）安装方法

全自动安装：在命令行下输入 pip install jieba 进行安装。

半自动安装：先下载 [jieba](http://pypi.python.org/pypi/jieba/" \t "_blank)，解压后运行 python setup.py install。

手动安装：将jieba目录放置于当前目录或者site-packages目录。

（4）使用方法

导入jieba库：

python

import jieba

简单分词：

python

text = "我喜欢Python编程"

words = jieba.cut(text)

print(" ".join(words))

输出结果为：我 喜欢 Python 编程。

分词模式的使用：

精确模式：jieba.cut(text, cut\_all=False)。

全模式：jieba.cut(text, cut\_all=True)。

搜索引擎模式：jieba.cut\_for\_search(text)。

（5）自定义词典

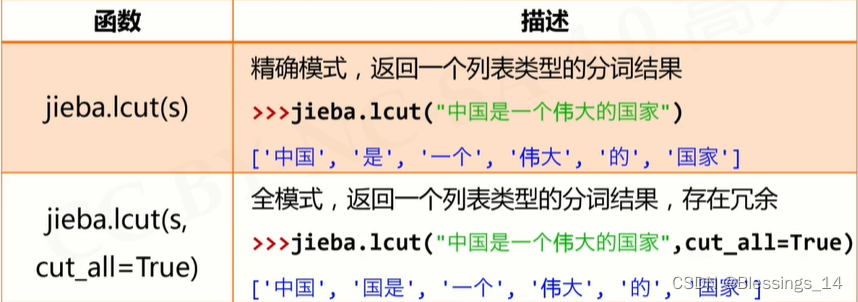
加载用户自定义词典：jieba.load\_userdict(file\_name)。

Jieba库因其高效和灵活的分词功能，在文本挖掘、搜索引擎、信息检索、文本分类、情感分析等多个领域都有广泛的应用

Jieba.cut（sentence，cut\_all=False,HMM=True）:sentence表示需要分词的句子，cut\_all表示是否采用全模式，HMM表示是否使用HMM（马尔可夫模型）,用来识别未登录词的，即那些不在词典中的词。

Jieba.cut\_search(sentence,HMM=True) sentence表示需要分词的句子, HMM表示是否使用HMM（马尔可夫模型）,该方法适合用于搜索引擎构建倒排索引的分词，粒度比较细。

Jieba.cut（）方法和Jieba.cut\_search（）方法返回的结果是可迭代对象，可使用for循环获取分词后得到的每一个词语，此外，jieba.lcut()和jieba.lcut\_for\_search()效果分别和Jieba.cut（）和Jieba.cut\_search（）效果类似，直接返回列表。



Add.word（word，freq=None，tag=None）和del\_word(word)可在程序中动态修改字典。

**三、实验内容**

本次实验主要掌握以下文本数据可视化：

1. 文本内容可视化

文本的内容可以通过关键词、短语、句子和主题进行展现。

（1）关键词可视化

标签云：

标签云是一种最常见的、最简单的关键词可视化方法。

（2）时序文本可视化

主题河流：

主题河流主要用于反映文本主题强弱变化的过程。

2.文本多特征信息可视化

关系网络图：

微博转发关系图是一种图形化展示微博用户之间信息传递和转发关系的可视化图表。在这种图表中，每个微博用户被表示为一个节点，用户之间的转发关系则用连接线连接。这样的图表有助于直观地理解微博平台上用户之间的信息传递和影响传播情况。

微博转发关系图的作用和意义包括：

（1）信息传递和传播分析：帮助分析信息在微博平台上的传递路径和传播情况。

（2）用户关系网络分析：发现微博用户之间的关联和群体结构。

（3）平台活跃度和热点分析：识别出哪些主题或事件在微博平台上引起了热议。

（4）可视化决策支持：为决策者提供直观的信息，有助于制定相关的营销策略、舆情应对措施等。

1. **实验步骤：**
2. 文本内容可视化

（1）关键词可视化

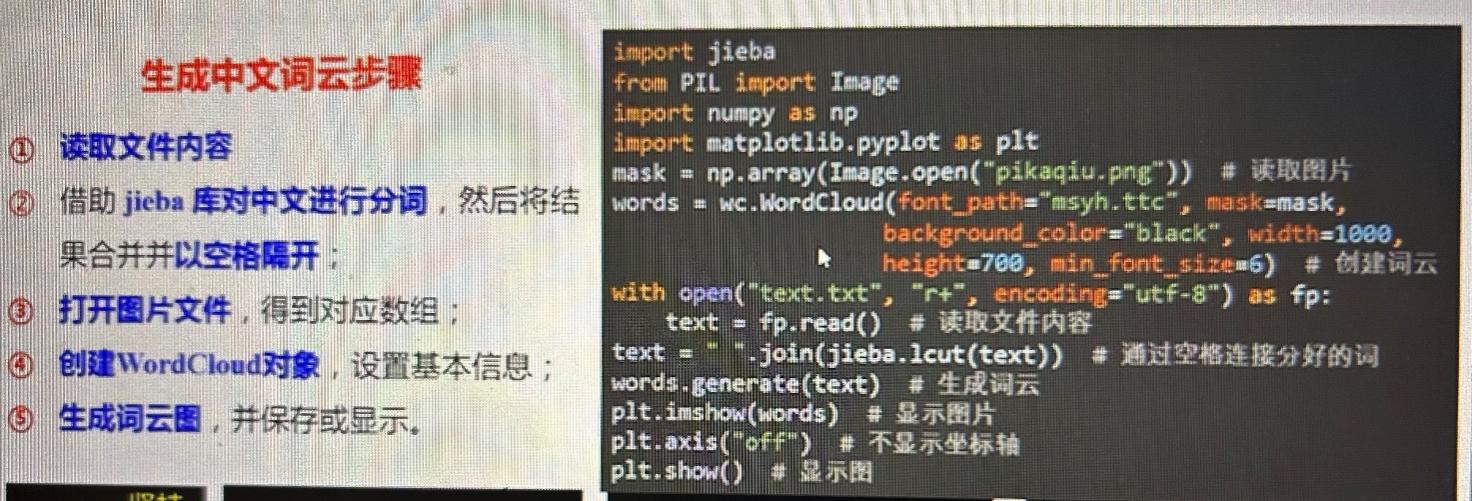
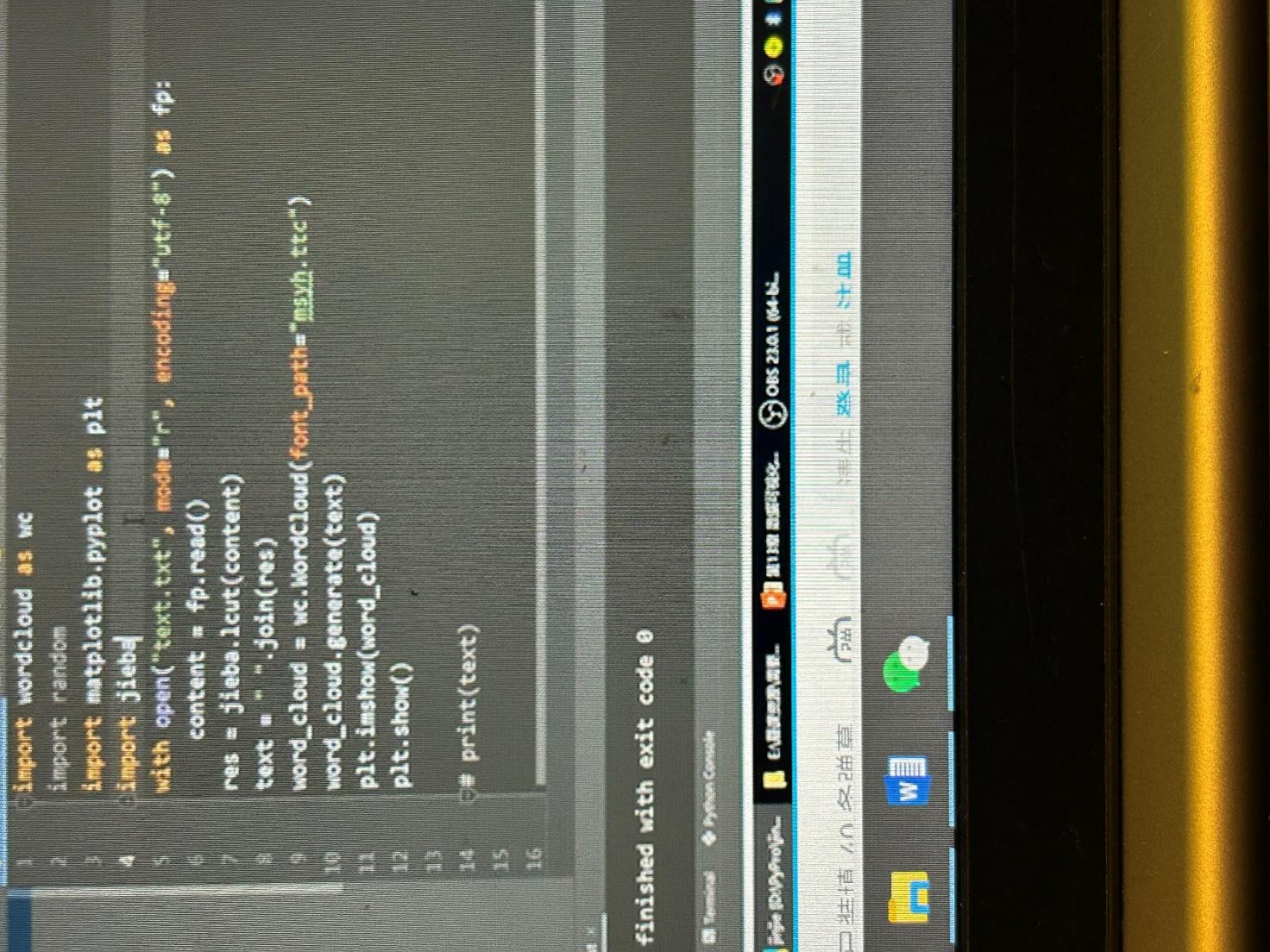
1. 标签云

步骤分为两步：

第一：统计出文本中词语的出现频率，提取中出现频率较高的关键词。

第二：按照一定的顺序和规律将这些关键词展示出来。

示例代码如下：

利用mask生成词云图的形状

一般方法：

1. 时序文本可视化
2. 主题河流

包括两个属性：

颜色，表示主题的类型，一个主题用一个单一颜色的涌流表示。

宽度，表示主题的数量（或强度），涌流的状态随着主题的变化，可能扩展、收缩或者保持不变。

示例代码：

import pyecharts.options as opts

from pyecharts import ThemeRiver

x\_data = ["DQ", "TY", "SS", "QG", "SY", "DD"]

y\_data = [

["2015/11/08", 10, "DQ"],

["2015/11/09", 15, "DQ"],

["2015/11/10", 35, "DQ"],

["2015/11/11", 38, "DQ"],

["2015/11/12", 22, "DQ"],

["2015/11/13", 16, "DQ"],

["2015/11/14", 7, "DQ"],

["2015/11/15", 2, "DQ"],

["2015/11/16", 17, "DQ"],

["2015/11/17", 33, "DQ"],

["2015/11/18", 40, "DQ"],

["2015/11/19", 32, "DQ"],

["2015/11/20", 26, "DQ"],

["2015/11/21", 35, "DQ"],

["2015/11/22", 40, "DQ"],

["2015/11/23", 32, "DQ"],

["2015/11/24", 26, "DQ"],

["2015/11/25", 22, "DQ"],

["2015/11/26", 16, "DQ"],

["2015/11/27", 22, "DQ"],

["2015/11/28", 10, "DQ"],

["2015/11/08", 35, "TY"],

["2015/11/09", 36, "TY"],

["2015/11/10", 37, "TY"],

["2015/11/11", 22, "TY"],

["2015/11/12", 24, "TY"],

["2015/11/13", 26, "TY"],

["2015/11/14", 34, "TY"],

["2015/11/15", 21, "TY"],

["2015/11/16", 18, "TY"],

["2015/11/17", 45, "TY"],

["2015/11/18", 32, "TY"],

["2015/11/19", 35, "TY"],

["2015/11/20", 30, "TY"],

["2015/11/21", 28, "TY"],

["2015/11/22", 27, "TY"],

["2015/11/23", 26, "TY"],

["2015/11/24", 15, "TY"],

["2015/11/25", 30, "TY"],

["2015/11/26", 35, "TY"],

["2015/11/27", 42, "TY"],

["2015/11/28", 42, "TY"],

["2015/11/08", 21, "SS"],

["2015/11/09", 25, "SS"],

["2015/11/10", 27, "SS"],

["2015/11/11", 23, "SS"],

["2015/11/12", 24, "SS"],

["2015/11/13", 21, "SS"],

["2015/11/14", 35, "SS"],

["2015/11/15", 39, "SS"],

["2015/11/16", 40, "SS"],

["2015/11/17", 36, "SS"],

["2015/11/18", 33, "SS"],

["2015/11/19", 43, "SS"],

["2015/11/20", 40, "SS"],

["2015/11/21", 34, "SS"],

["2015/11/22", 28, "SS"],

["2015/11/23", 26, "SS"],

["2015/11/24", 37, "SS"],

["2015/11/25", 41, "SS"],

["2015/11/26", 46, "SS"],

["2015/11/27", 47, "SS"],

["2015/11/28", 41, "SS"],

["2015/11/08", 10, "QG"],

["2015/11/09", 15, "QG"],

["2015/11/10", 35, "QG"],

["2015/11/11", 38, "QG"],

["2015/11/12", 22, "QG"],

["2015/11/13", 16, "QG"],

["2015/11/14", 7, "QG"],

["2015/11/15", 2, "QG"],

["2015/11/16", 17, "QG"],

["2015/11/17", 33, "QG"],

["2015/11/18", 40, "QG"],

["2015/11/19", 32, "QG"],

["2015/11/20", 26, "QG"],

["2015/11/21", 35, "QG"],

["2015/11/22", 40, "QG"],

["2015/11/23", 32, "QG"],

["2015/11/24", 26, "QG"],

["2015/11/25", 22, "QG"],

["2015/11/26", 16, "QG"],

["2015/11/27", 22, "QG"],

["2015/11/28", 10, "QG"],

["2015/11/08", 10, "SY"],

["2015/11/09", 15, "SY"],

["2015/11/10", 35, "SY"],

["2015/11/11", 38, "SY"],

["2015/11/12", 22, "SY"],

["2015/11/13", 16, "SY"],

["2015/11/14", 7, "SY"],

["2015/11/15", 2, "SY"],

["2015/11/16", 17, "SY"],

["2015/11/17", 33, "SY"],

["2015/11/18", 40, "SY"],

["2015/11/19", 32, "SY"],

["2015/11/20", 26, "SY"],

["2015/11/21", 35, "SY"],

["2015/11/22", 4, "SY"],

["2015/11/23", 32, "SY"],

["2015/11/24", 26, "SY"],

["2015/11/25", 22, "SY"],

["2015/11/26", 16, "SY"],

["2015/11/27", 22, "SY"],

["2015/11/28", 10, "SY"],

["2015/11/08", 10, "DD"],

["2015/11/09", 15, "DD"],

["2015/11/10", 35, "DD"],

["2015/11/11", 38, "DD"],

["2015/11/12", 22, "DD"],

["2015/11/13", 16, "DD"],

["2015/11/14", 7, "DD"],

["2015/11/15", 2, "DD"],

["2015/11/16", 17, "DD"],

["2015/11/17", 33, "DD"],

["2015/11/18", 4, "DD"],

["2015/11/19", 32, "DD"],

["2015/11/20", 26, "DD"],

["2015/11/21", 35, "DD"],

["2015/11/22", 40, "DD"],

["2015/11/23", 32, "DD"],

["2015/11/24", 26, "DD"],

["2015/11/25", 22, "DD"],

["2015/11/26", 16, "DD"],

["2015/11/27", 22, "DD"],

["2015/11/28", 10, "DD"],

]

(

ThemeRiver()

.add(

series\_name=x\_data,

data=y\_data,

singleaxis\_opts=opts.SingleAxisOpts(

pos\_top="50", pos\_bottom="50", type\_="time"

),

)

.set\_global\_opts(

tooltip\_opts=opts.TooltipOpts(trigger="axis", axis\_pointer\_type="line")

)

.render("theme\_river.html")

)

3.微博转发关系图

微博转发关系图是一种图形化展示微博用户之间信息传递和转发关系的可视化图表。在这种图表中，每个微博用户被表示为一个节点，用户之间的转发关系则用连接线连接。这样的图表有助于直观地理解微博平台上用户之间的信息传递和影响传播情况。

**关键要素和解释包括**：

节点(Nodes)：每个微博用户都被表示为图中的一个节点。节点通常用不同的形状、颜色或图标来区分用户的不同属性或类别。

连接(Links)：连接线表示用户之间的转发关系。如果用户A转发了用户B的微博，那么就会有一条连接线从节点A指向节点B。这样的连接线构成了整个图的基础结构。

节点类别(Categories)：节点类别可以用来对微博用户进行分类，例如按照用户的地理位置、兴趣爱好、认证身份等。这有助于在图中更清晰地展示不同类别用户之间的关系。

斥力(Repulsion)：斥力是一种设置，用于控制节点之间的排斥程度。在布局过程中，斥力会使得节点更均匀地分布在图上，避免节点之间过于拥挤。

曲线(Line Curve)：连接线的曲率可以通过设置曲线参数，如 curve=0.2，使得连接线呈现曲线形状。这样的设计可以使图更具美感，同时保持了节点之间的关系。

实验步骤：

第一：导入数据

第二：定义节点

第三：绘制图表

部分json数据：

    [{

            "name": "",

            "symbolSize": 5,

            "draggable": "False",

            "value": 1,

            "category": "",

            "label": {

                "normal": {

                    "show": "True"

                }

            }

        },

        {

            "name": "Camel3942",

            "symbolSize": 5,

            "draggable": "False",

            "value": 1,

            "category": "Camel3942",

            "label": {

                "normal": {

                    "show": "True"

                }

            }

        },

  {

            "name": "血红暴鲤魚",

            "symbolSize": 5,

            "draggable": "False",

            "value": 0,

            "category": "战争史研究WHS"

        },

        {

            "name": "傲血困意",

            "symbolSize": 5,

            "draggable": "False",

            "value": 0,

            "category": "简木生--包丰瀛"

        },

        {

            "name": "只道是寻常草履虫",

            "symbolSize": 5,

            "draggable": "False",

            "value": 0,

            "category": "新浪体育"

        },

示例代码：

from pyecharts import Graph

# 定义节点

nodes = [

{"name": "", "symbolSize": 5, "draggable": "False", "value": 1, "category": "", "label": {"normal": {"show": "True"}}},

{"name": "Camel3942", "symbolSize": 5, "draggable": "False", "value": 1, "category": "Camel3942", "label": {"normal": {"show": "True"}}}

]

# 创建图表

graph = Graph("网络关系图", width=1200, height=600)

graph.add("", nodes, links=[], categories=[], label\_pos="right", is\_legend\_show=False)

graph.render("network\_relationship.html")

**第一个节点：**

"name": ""：节点的名称为空字符串，这可能意味着这个节点没有特定的名称或者是一个占位符。

"symbolSize": 5：节点的大小被设置为5，这可能影响节点在图表中的显示大小。

"draggable": "False"：节点不可拖动，这意味着用户不能在图表中通过拖动来移动这个节点。

"value": 1：节点的值被设置为1，这个值可能用于影响节点的视觉表示，如大小、颜色等。

"category": ""：节点的类别为空字符串，这可能意味着这个节点没有被分配到特定的类别。

"label"：节点标签的设置。

"normal"：正常状态下的标签样式。

"show": "True"：在正常状态下显示节点的标签。

**第二个节点：**

"name": "Camel3942"：节点的名称为"Camel3942"，这可能是一个用户ID、名称或其他标识符。

"symbolSize": 5：同样，节点的大小被设置为5。

"draggable": "False"：节点不可拖动。

"value": 1：节点的值同样被设置为1。

"category": "Camel3942"：节点被分配到了名为"Camel3942"的类别，这可能用于在图表中对节点进行分组或分类。

"label"：节点标签的设置与第一个节点相同，正常状态下显示标签。

使用场景：

这些节点可以用于 pyecharts 的 Graph 类来创建图表。例如，你可以将这些节点添加到图表中，以展示社交网络、组织结构或其他类型的网络关系。每个节点的属性，如名称、大小、是否可拖动、值和类别，都可以影响其在图表中的显示和行为。

**四、实验任务**

根据实验内容和实验步骤的所学知识，完成下面的任务：

1. 根据“post\_data”中views列的数据量将category列生成词云图。

2.利用jieba库对“text”进行分词，并生成词云图。

3.利用text1中的数据，完成主题河流的可视化。

4.利用“weibo-2”json数据完成微博转发关系图的绘制。