CSV数据集数据结构分析

根据提供的列名和示例数据，该CSV数据集的数据结构分析如下（使用制表符`\t`作为分隔符）：

📊 一、核心字段结构

1. `UserID`
   - **数据类型**：字符串（如`U335175`）
   - **作用**：唯一标识用户，每条记录对应一个用户的行为数据。
2. `ClicknewsID`
   - **数据类型**：多值字符串（如`N41340 N27570 N83288...`）
   - **特点**：以空格分隔的新闻ID列表，表示用户点击的多个新闻（示例含36个ID）。
3. `dwelltime`
   - **数据类型**：多值数值（如`116 23 59...`）
   - **特点**：以空格分隔的整数列表，对应`ClicknewsID`中每个新闻的停留时间（秒/毫秒），数量需与新闻ID一致。
4. `exposure_time`
   - **数据类型**：多值时间戳（如`6/19/2019 5:10:01 AM#TAB#6/19/2019 5:11:58 AM...`）
   - **特点**：
     - 用`#TAB#`分隔的日期时间字符串（实际应为制表符，可能被转义显示）
     - 表示每个新闻的曝光时间点，数量与`ClicknewsID`匹配。
5. `pos` **与** `neg`
   - **数据类型**：多值字符串（如`pos: N55476 N103556...`, `neg: N48119 N92507...`）
   - **作用**：分别记录用户的正反馈（如点赞）和负反馈（如忽略）新闻ID列表。
6. `start` **与** `end`
   - **数据类型**：单值时间戳（如`start: 7/3/2019 6:43:49 AM`）
   - **作用**：可能表示用户会话的开始和结束时间。
7. `dwelltime_pos`
   - **数据类型**：多值数值（如`34 83 79 234 16`）
   - **作用**：对应`pos`列表中每个新闻的停留时间。

⚠️ 二、数据结构关键特点

1. **多值字段嵌套**
   - `ClicknewsID`、`dwelltime`、`exposure_time`等字段包含多个值，需解析为数组或拆分列。
   - **挑战**：不同字段的列表长度需对齐（如`ClicknewsID`与`dwelltime`均为36个值，而`dwelltime_pos`仅5个值）。

2. **时间格式混合**
    - 时间数据包含日期与精确到秒的时间（如`7/3/2019 6:43:49 AM`），需统一转换为标准时间类型。
3. **行为反馈分离**
    - 正/负反馈新闻（`pos`/`neg`）独立存储，可能用于分析用户偏好。
4. **分隔符特殊处理**
    - 列间以制表符（`\t`）分隔，字段内多值用空格或`#TAB#`分隔，需分层解析。

🔧 三、数据处理建议

1. **读取时指定分隔符**

```
import pandas as pd
df = pd.read_csv('data.csv', sep='\t')  # 显式声明制表符分隔
```