# 小荷医典知识库数据统计报告

报告日期: 2025年9月30日

分析目的: 本报告旨在对从“小荷医典”采集并清洗后的医疗知识数据进行全面的统计分析，从数据规模、内容分布和字段结构三个维度，为后续的检索增强生成（RAG）模型的开发提供数据洞察和决策支持。

## 1. 数据规模 (Scale)

数据集的总体规模反映了知识库的广度。

|  |  |
| --- | --- |
| **统计项** | **数量(条)** |
| **总词条数量** | **25076** |
| - **疾病** (Disease) | 3855 |
| - **症状** (Symptom) | 1189 |
| - **药品** (Medicine) | 17518 |
| - **检查治疗** (Examination) | 2514 |

**分析:**

* 数据集共包含 **25076**条独立的医疗词条，构成了一个内容丰富的基础知识库。
* 其中，**药品**词条数量最多，是知识库的核心组成部分，而 **症状**数量相对较少。

## 2. 数据分布 (Distribution)

### 2.1 内容丰富度 (按分区数量)

每个词条包含的分区（如“病因”、“治疗”等）数量。

|  |  |
| --- | --- |
| **统计项** | **数值** |
| 平均每个词条包含的分区数 | **4.88** |
| 最多分区数 | 15 |
| 最少分区数 | 1 |

### 2.2 内容长度 (按字符数)

每个词条所有分区内容的字符总长度。

|  |  |
| --- | --- |
| **统计项** | **数值(字符)** |
| 平均每个词条的内容长度 | **221.89** |
| 最长词条的内容长度 | 2069 |
| 最短词条的内容长度 | 3 |

### 2.3 常见分区标题

统计所有词条中最常出现的分区标题，这有助于理解数据的通用结构。

|  |  |  |
| --- | --- | --- |
| **排名** | **分区标题 (Section Title)** | **出现次数(次)** |
| 1 | 概述 | 20735 |
| 2 | 功效作用 | 16694 |
| 3 | 用法用量 | 15929 |
| 4 | 用药禁忌 | 14627 |
| 5 | 注意事项 | 13608 |
| 6 | 病因 | 4061 |
| 7 | 治疗 | 3753 |
| 8 | 诊断 | 3465 |
| 9 | 症状 | 3261 |
| 10 | 检查 | 2540 |
| 11 | 就医 | 2082 |

## 3. 字段说明 (Field Description)

最终清洗后的数据 cleaned\_medical\_data.json 是一个包含多个对象的列表，每个对象代表一个医疗词条，其结构如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名 (Field)** | **数据类型 (Type)** | **说明 (Description)** | **示例 (Example)** |
| name | String | 词条的官方名称。 | "头晕" |
| id | String | 词条在源网站的唯一标识符。 | "34828979978497" |
| url | String | 词条在源网站的URL链接。 | "https://www.xiaohe-jiankang.com/medical/disease/..." |
| type | String | 词条的类型，便于分类。 | "症状" |
| sections | Object/Dictionary | 包含该词条所有分区内容的字典。 | {"病因": "导致头晕的病因很多...", "治疗": "1. 药物治疗..."} |
| sections.key | String | 分区标题，如“概述”、“病因”、“治疗”等。 | "病因" |
| sections.value | String / Array | 对应分区标题下的详细文本内容。 | "导致头晕的病因很多..." |