题目二 基于中文文本的分词与可视化分析

现有来自淘宝平台的手机商品评论数据集，已完成初步整理。该数据集以 CSV 格式呈现，包含 500 条真实评论数据，具体字段如下：

comment\_id：评论的唯一标识 ID。

content：用户撰写的评论内容，均为中文文本。

score：用户对商品给出的评分，取值范围为 1 至 5 星，其中 1 - 2 星代表差评，3 星为中评，4 - 5 星属于好评。

数据样例如下：

|  |  |  |
| --- | --- | --- |
| comment\_id | content | score |
| 1001 | 手机外观很漂亮，运行速度也快，就是电池续航一般 | 4 |
| 1002 | 手机拍照效果太差了，完全不符合宣传 | 2 |

作业要求

一、数据预处理

过滤掉评论内容中的特殊字符、标点符号以及类似 “嗯”“啊” 等无意义词汇。

二、分词与词频分析

运用 jieba 分词工具对评论内容进行分词操作，具体要求如下：

加载自定义词典（custom\_dict.txt），词典内包含手机相关专业术语，如 “处理器”“像素”“快充” 等，以提升分词准确性。

去除常见的停用词（stopwords.txt），如 “的”“了”“是” 等，减少无价值词汇干扰。

利用词性标注（pos\_tag）筛选并保留名词、动词、形容词，以提取关键信息。

对所有评论数据进行词频统计，找出出现频率最高的前 30 个词汇。

三、可视化分析

绘制词云图：根据所有评论的高频词生成词云图，合理设置背景颜色、字体样式以及词汇颜色，使词云图美观且能清晰展示高频词汇分布。

绘制柱状图：展示出现频率最高的前 15 个词汇的出现次数，直观呈现各词汇的提及情况。