提交要求：

1. 以下2道题目**任选其一**；
2. 提交word/PDF，包含源代码（带注释），执行过程截图和输出结果截图，也可以包含思考过程（可选）；
3. 大作业工程文件为非必须提交项。如果提交，请单独打包，不要和文档打包在一起。

**题目1：词频统计进阶版**

​给定一个包含多行文本的RDD，统计每个单词的出现频率，要求：

1. 过滤长度小于3的单词
2. 排除停用词（给定集合：val stopwords = Set("the", "and", "of", "to")）
3. 按词频降序排列，输出前10个结果

**示例输入：**

Spark is a fast and general-purpose cluster computing system

Spark can load data from HDFS and process it efficiently

**示例输出：**

(spark,2), (data,1), (load,1), (system,1), (fast,1), (purpose,1), (cluster,1), (computing,1), (efficiently,1), (process,1)

**题目2：销售数据聚合**

​给定订单DataFrame（schema: order\_id:String, user\_id:Int, amount:Double, date:String），完成：

1. 将date字段转换为DateType（原始格式为"yyyyMMdd"）
2. 按月份统计销售总金额
3. 输出结果按月份升序排列

**示例输入：**

ord001, 1001, 99.9, 20230115

ord002, 1002, 149.9, 20230120

ord003, 1001, 199.9, 20230201s

**示例输出：**

2023-01 - 249.8

2023-02 - 199.9