数据结构与算法大作业题目：热词统计与分析系统

题目背景

随着互联网信息的爆炸式增长，实时监控和分析网络文本中的热门词汇对于舆情监测、趋势分析、内容推荐等领域具有重要意义。本次大作业要求同学们设计并实现一个**基于滑动窗口的热词统计与分析系统**，重点考察对流式数据处理、核心数据结构设计以及性能调优的综合能力。

系统目标概述

在不断到来的文本数据流上，实时维护一个时间窗口的词频，并提供可配置的 Top-K 查询与趋势分析能力。

功能与要求

建议先把滑动窗口里的中文分词与词频统计链路打磨扎实，这部分是所有作品的共同基础；迟到/乱序处理、趋势洞察、预测等其他功能可以按兴趣和时间逐步拓展。

1. 实时数据流处理

- 支持持续接收文本数据，可使用脚本模拟或接入真实数据源。

2. 文本预处理

- 支持中文分词和基础清洗（停用词、标点、无意义符号）。允许直接调用现有工具（如 Jieba），也可简化为给定词表/字典接口。

3. 滑动窗口热词统计（核心）

- 滑动窗口语义：本题默认**基于时间的滑动窗口**，窗口大小固定，步长 10 分钟（**可自行配置**），可选增加以**基于消息数的滑动窗口**。
- 提供可配置的 **Top-K** 查询功能。
- 需在文档中说明窗口淘汰策略与时间同步方式。

4. 高级功能（可选）

1. **敏感词过滤**： 对敏感词过滤，屏蔽部分敏感词或词性的统计。
2. **迟到/乱序数据处理**：对迟到/乱序数据能正确处理和统计。
3. **趋势分析**：绘制或计算词频的增长/衰退斜率，识别新兴或降温词汇，说明检测阈值。
4. **历史查询**：对持久化的窗口快照做离线分析，可按一定时间聚合并生成报告。
5. **动态滑动窗口大小**：窗口大小可动态修改，统计动态时间大小范围内的词频。
6. **交互式可视化**：实现 CLI/GUI/Web 界面展示实时热词列表、词云或趋势曲线。

5. 性能与资源约束

- 需评估时间复杂度。
- 需估算并报告内存占用。
- （可选）给出在不同输入速率下的吞吐量与延迟。
- （可选）若使用多线程或异步处理，请描述任务划分、锁/无锁策略及其复杂度。

技术实现要求

核心数据结构

1. **实时计数器**：要求 O(1) 或接近 O(1) 的增量更新，可考虑哈希表、Trie、分段计数等方案，并在文档中进行复杂度分析。
2. **时间窗口管理器**：支持滑动更新与过期数据淘汰，可采用队列 + 哈希、时间分桶、分层计数器等策略，并在文档中进行复杂度分析。
3. **Top-K 维护结构**：考虑使用小顶堆、双堆结合、平衡树或 C++ STL set 等算法；需说明在高频更新下保持 Top-K 正确性的措施。

系统架构建议

- **数据采集层**：负责接入/生成数据流，提供统一缓冲接口。
- **预处理层**：分词、清洗、词性过滤。
- **统计核心**：实时计数器、窗口管理器、Top-K 维护结构。
- **查询与服务层**：对外提供 API/CLI/GUI，支持 Top-K、趋势、历史查询。
- **持久化与监控**：窗口快照、日志、性能指标采集。

交付物清单

1. **系统设计文档（必交）**

- 背景假设与外部依赖
- 模块/架构图与数据流
- 核心数据结构设计、复杂度分析、设计取舍
- 滑动窗口定义、实时性保证
- 性能优化与资源评估方法

2. **源代码与单元测试（必交）**
   - 完整工程（建议提供编译/运行脚本或 Makefile）
   - 代码需包含关键模块注释
   - 单元/集成测试须覆盖：分词或清洗逻辑、窗口淘汰、Top-K 正确性
3. **性能测试报告（必交）**
   - 测试环境与数据生成方式
   - 不同负载下的吞吐、延迟、内存占用等图表
   - 其它分析与改进建议
4. **演示程序或视频（选交）**
   - 展示实时热词、趋势曲线或交互查询
   - 提供操作说明或录屏链接
5. **实验日志（选交）**
   - 记录迭代过程、遇到的问题与解决方案，可作为创新性佐证

评分标准

评分项	权重	说明
核心数据结构与滑动窗口实现	40%	设计合理性、复杂度分析、正确性验证
功能完整度(高级功能)	15%	基本功能达成度及高级功能深度
系统性能与可靠性	20%	吞吐/延迟指标、资源控制、异常处理
代码质量与测试	15%	模块化、可读性、测试覆盖率
文档与展示	10%	设计文档、性能报告、演示材料质量

开发与实现建议

- 优先保证滑动窗口与 Top-K 的正确性，再逐步扩展功能。
- 充分利用现有中文分词库和停用词表，避免重复造轮子。
- 利用日志/指标组件记录系统状态，方便调试与写报告。

样例参考（可自定义输入输出格式）

输入数据流：

[00:00:00] 今天人工智能技术发展迅速
[00:01:00] 人工智能将改变未来生活方式
[ACTION] QUERY K=1
[00:02:00] 技术创新推动人工智能进步
[ACTION] QUERY K=2

输出示例：

[00:01:00]

1. 人工智能 (出现 2 次)

[00:02:00]

1. 人工智能 (出现 3 次)
2. 技术 (出现 2 次)

提交与验收

- 截止时间：第 16 周周日（2025年12月28日） 23:59，统一提交到课程平台。
- 提交内容：压缩包（命名为 `学号_姓名_hotwords.zip` ），包含源码、文档、报告、可执行程序/脚本、可选演示材料。
- 验收方式：第 17 周实验课堂，优秀报告进行现场展示。
- 引用规范：允许调用开源库，但需在文档中列出版本与许可证；不得引用未注明来源的第三方代码。