[CS209A-24Fall] 最终项目（100 分）

背景介绍

在软件开发过程中，会出现许多问题。开发人员可能会求助于问答网站来发布问题和寻求答案。

Stack Overflow 就是这样一个面向程序员的问答网站，它隶属于 Stack Exchange 网络。Stack Overflow 为用户提供了一个提问和回答问题的平台，并通过会员制和活跃的

在 Stack Overflow 上，用户可以对问题和答案进行 "向上 "或 "向下 "投票，并以类似于维基的方式编辑问题和答案。Stack Overflow 的用户可以获得声誉分和 "徽章"；例如，一个人对一个问题或对一个问题的答案获得 "向上 "投票，可获得 10 点声誉分，并可获得

徽章来表彰他们的宝贵贡献。用户的声望越高，就能获得新的特权，如投票、评论甚至编辑他人帖子。

在这个期末项目中，我们将使用 Spring Boot 开发一个网络应用程序，用于存储、分析和可视化有关 Java 编程的 Stack Overflow 问答数据，目的是了解常见的 Java 编程问题。
与 Java 编程相关的问题、答案和解决活动。

数据收集（10 分）

在 Stack Overflow 上，与 Java 编程相关的问题通常会被标记为 java。您可以使用 java 标签来识别与 java 相关的问题。一个问题及其所有答案和评论统称为一个线程。

对于 Stack Overflow 上**与 java 相关**的**线程**，我们有兴趣回答以下问题列表。要回答这些问题，您应首先从 Stack Overflow 收集适当的数据。请查看 Stack Overflow REST API 官方文档，了解用于收集不同类型数据的 REST API。

- 您可能需要创建一个 Stack Overflow 账户，才能使用其完整的 REST API 服务。
- API 请求受速率限制。**请仔细设计和执行您的请求，否则您可能很快就会达到每日配额**。
- 与 Stack Overflow REST 服务的连接有时可能不稳定。因此，**请尽快开始数据收集！**

Stack Overflow 上有 100 多万个标记为 java 的线程。您不必全部收集。但是，您应该收集**至少 1000 个线程**的数据，以便从数据分析中获得有意义的见解。

重要：

数据收集是***离线的***，这意味着您需要先收集和持久化数据。建议使用数据库（如 PostgreSQL、MySQL 等）来存储数据。不过，也可以将数据存储在普通文件中。换句话说，当用户与**您**的应用程序交互时，**服务器应从本地数据库**（或本地文件）**中获取数据**，而不是临时向 Stack Overflow 发送 REST 请求。

因此，下面问题的数据分析应该在您收集的数据集上进行。也就是说，我们首先收集 Stack Overflow 数据的一个子集（例如，1000 个标记为 java 的线程），然后回答以下问题

利用这个子集提出以下问题。

第一部分：数据分析（70 分）

对于本部分的每个问题，您应该

- 找出回答问题所需的数据• 在后台设计和实施数据分析•

  使用适当的图表在前台可视化结果。

换句话说，当用户通过浏览器与网络应用程序交互时，可以选择感兴趣的
服务器执行相应的数据分析，并将结果返回前端，前端将结果可视化到网页上。

您的作品将通过以下方式进行评估

- •数据分析是否有意义和相关，即通过对适当数据的适当分析，是否确实能回答问题。回答一个问题可能有多种方法。要有创意！
- •可视化是否能有效传达理念，即用户能否获得想要的信息
  即刻查看可视化效果。查看数据可视化目录，寻找灵感。

1. Java 主题（10 分）

我们在本课程中涉及了各种主题，例如泛型、集合、I/O、lambda、多线程、socket 等。我们很想知道，在 Stack Overflow 上最常被问到的前 N 个（N>1，您可以根据您的数据和用户界面设计选择合适的 N，下同）主题是什么？

2. 用户参与（15 分）

声誉分数较高的用户参与度最高的 N 个主题是什么？用户参与度是指用户在主题上的任何活动（如编辑、回答、评论、向上投票、向下投票等）。

3. 常见错误（15 分）

开发人员会犯错，从而导致代码出现错误。**错误**表现为**错误**或
例外情况大致可分为以下几类：

- 致命错误：运行时无法恢复的错误，如 OutOfMemoryError。
- •异常：已检查异常和运行时异常，开发人员可通过编程处理。

Java 开发人员经常讨论的 N 大错误和异常是什么？

请注意，标签是高级信息，可能不包括低级错误或异常。因此，对于这个问题，您不能只使用标签信息。您需要进一步分析线程内容（如问题文本和答案文本），以识别与错误或异常相关的信息，可能需要使用正则表达式匹配等高级技术。

4. 答案质量（30 分）

如果一个答案被接受或有很多向上投票，我们就认为它是 "高质量 "的。了解哪些因素促成了高质量的答案很有帮助？

请调查以下因素：

- 创建问题与创建答案之间的时间间隔（例如，第一个发布的答案是否往往被接受？）
- 创建答案的用户的声誉（例如，高声誉用户创建的答案是否更容易被接受或获得更多的向上投票？）

除了这两个因素外，您还应提出另外一个可能影响答案质量的因素。

针对 3 个因素中的每个因素，使用适当的数据分析和可视化方法来证明该因素是否有助于高质量的答案。

第二部分：RESTful 服务（20 分）

您的应用程序还应提供能回答以下两个问题*的 REST 服务*，以便用户使用 RESTful API 获取他们想要的答案。

所需的 REST 服务包括

- 主题频率：用户可以查询特定主题的频率。用户还可以查询按频率排序的前 N 个话题。
- 错误频率：用户可以查询特定错误或异常的频率。用户还可以查询按频率排序的前 N 个错误或异常。

在此，您可以重复使用第一部分的数据分析。

REST 请求的响应应为 json 格式。

要求

数据分析

您应使用集合、Lambda 和流等 Java 功能自己实现数据分析。

您**不能**将数据提供给人工智能，让人工智能进行分析，并将人工智能的回复作为您的数据分析结果。
如果你这样做，本题将得 0 分。

每次客户端发送请求时，服务器都应**动态生成**数据分析结果。**不应**预先计算结果并将其存储为静态内容，然后在前端简单地显示预先计算的静态内容。**否则将扣 20 分。**

网络框架

您只能使用 Spring Boot 作为网络框架。

前端

数据可视化和交互式控件等前端功能可以任何编程语言（如 JavaScript、HTML、CSS 等）和任何第三方库或框

架实现。