项目.md 2024-11-10

[CS209A-24Fall] 最终项目 (100分)

背景

在软件开发过程中，开发者会遇到各种问题，他们可能会通过问答网站来发布问题并寻求答案。**Stack Overflow** 就是一个程序员常用的问答网站，属于 **Stack Exchange** 网络的一部分。Stack Overflow 提供一个平台，用户可以在上面提问和回答问题，并通过会员身份和积极参与来对问题和答案进行投票（点赞或点踩），以及像维基一样编辑问题和答案。用户通过参与获得声誉积分和"徽章"；例如，当一个问题或答案获得"点赞"时，用户将获得 10 个声誉积分，并且可以通过有价值的贡献获得徽章。随着声誉的增加，用户可以解锁新的特权，例如投票、评论，甚至编辑其他人的帖子。

在这个最终项目中，我们将使用 **Spring Boot** 开发一个 Web 应用程序，用来存储、分析和可视化与 **Java 编程** 相关的 Stack Overflow 问答数据，目的是理解与 Java 编程相关的常见问题、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打上 **java** 标签。你可以使用这个标签来识别与 Java 相关的问题。一个问题及其所有答案和评论统称为"线程"。

对于 Stack Overflow 上与 Java 相关的线程，我们关注以下问题。你需要先从 Stack Overflow 收集适当的数据来回答这些问题。请查阅 **Stack Overflow REST API** 的官方文档，以了解如何通过 REST API 收集不同类型的数据。你可能需要创建一个 Stack Overflow 账户，以便使用其完整的 REST API 服务。

API 请求会受到速率限制，因此请小心设计和执行你的请求，否则你可能会很快达到每日配额。连接 Stack Overflow REST 服务有时可能不稳定，因此请尽早开始数据收集工作！

Stack Overflow 上有超过 100 万个带有 **java** 标签的线程。你**不需要**收集所有这些数据，但至少应该收集 1000 个线程的数据，才能从数据分析中获得有意义的见解。

重要： 数据收集是离线进行的，这意味着你需要首先收集并持久化数据。建议使用数据库（例如 **PostgreSQL**、**MySQL** 等）存储数据，但如果将数据存储在普通文件中也是可以的。换句话说，当用户与你的应用程序交互时，服务器应该从本地数据库（或本地文件）获取数据，而不是实时向 Stack Overflow 发送 REST 请求。

因此，下面问题的数据分析应基于你收集的数据集进行。也就是说，我们首先收集一部分 Stack Overflow 数据（例如 1000 个带有 java 标签的线程），然后使用这些数据集来回答以下问题。

第一部分：数据分析 (70分)

对于本部分的每个问题，你需要：

1. 找出回答问题所需的数据
2. 设计并在后台实现数据分析
3. 使用合适的图表在前端可视化结果

换句话说，当用户从浏览器与应用程序交互时，用户可以选择感兴趣的分析内容，应用程序会向服务器发送请求，服务器执行相应的数据分析并将结果返回给前端，前端将结果可视化显示在网页上。

评估标准包括：

- 数据分析是否有意义且相关，即它能够通过适当的分析回答问题，并且分析的数据是正确的。回答问题可能有多种方式，鼓励创造性思维！

- 可视化效果是否能够有效传达信息，即用户能通过查看图表快速得到想要的信息。可以参考数据可视化的示例以获取灵感。

1. **Java 话题 (10分)**
   在课程中，我们已经涵盖了各种 Java 话题，例如泛型、集合、I/O、lambda、线程、socket 等。了解哪些话题在 Stack Overflow 上最常见是非常有趣的。请找出 **Stack Overflow** 上最常被问到的前 N 个 (N>1，可以根据你的数据和 UI 设计选择一个合适的 N) Java 编程相关话题。
2. **用户参与度 (15分)**
   哪些话题得到了来自高声誉用户最多的参与？用户参与指的是任何用户在线程中的活动，例如编辑、回答、评论、点赞、点踩等。
3. **常见错误 (15分)**
   开发者在编程时经常会犯错误，导致代码出现 bug。错误通常表现为错误或异常，可以大致分为：
   - 致命错误：如 **OutOfMemoryError**，这种错误在运行时无法恢复。
   - 异常：可以由开发者通过编程手段处理的异常，包括受检异常和运行时异常。

   哪些错误或异常在 Java 开发者的讨论中最为频繁？需要注意的是，标签信息是高层次的，可能不包括低层次的错误或异常。因此，单纯依靠标签信息不够。你需要进一步分析线程内容（例如问题文本和答案文本），识别与错误或异常相关的信息，可能需要使用正则表达式等高级技术。
4. **答案质量 (30分)**
   我们认为一个答案是"高质量"的，如果它被接受或获得了很多点赞。了解哪些因素可能影响答案质量是很有用的。请调查以下几个因素：
   - 问题创建和答案创建之间的时间间隔（例如，是否第一个发布的答案更可能被接受？）
   - 提供答案的用户的声誉（例如，声誉较高的用户发布的答案是否更可能被接受或获得更多的点赞？）

   除了这两个因素外，你还需要提出一个可能影响答案质量的额外因素。

   对每个因素，使用适当的数据分析和可视化，展示这个因素是否对高质量答案有所贡献。

第二部分：RESTful 服务 (20分)

你的应用程序还需要提供一个 REST 服务，回答以下两个问题，用户可以通过 RESTful API 来获取他们想要的答案。需要的 REST 服务包括：

- **话题频率**：用户可以查询特定话题的频率。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。

在这里，你可以重用第一部分的数据分析。

REST 请求的响应应为 JSON 格式。

要求

1. **数据分析**
   你应该使用 Java 的功能，如 **Collections**、**Lambda** 和 **Stream** 来实现数据分析。
   不能将数据交给 AI 来处理、让 AI 完成分析或使用 AI 提供的分析结果。这样会导致该问题得 0 分。
   数据分析结果应该每次客户端发送请求时由服务器动态生成。**不应该**预先计算结果并将其存储为静态内容，只是将静态内容展示在前端。如果这样做，将扣 20 分。
2. **Web 框架**
   你应该仅使用 **Spring Boot** 作为 Web 框架。

3. **前端**
   前端功能，例如数据可视化和交互式控制，可以使用任何编程语言（如 JavaScript、HTML、CSS 等）和任何第三方库或框架来实现。

以下是两个文件的翻译结果：

评分标准.md 2024-11-10

[CS209A-24Fall] 最终项目评分标准 (100分)

数据收集 (10分)

- 使用 Stack Overflow API 收集至少 1000 个带有 Java 标签的线程。 (5分)
- 将数据存储在数据库（例如 PostgreSQL、MySQL）或本地文件中，以确保后续数据分析可以从本地源访问数据。 (5分)

Java 话题分析 (10分)

- 分析收集到的数据中的常见 Java 话题。 (5分)
- 数据可视化。 (5分)

用户参与分析 (15分)

- 计算用户参与度。 (5分)
- 分析高声誉用户参与较多的话题。 (5分)
- 数据可视化。 (5分)

常见错误分析 (15分)

- 正确提取错误和异常信息。 (5分)
- 分析常被讨论的错误和异常。 (5分)
- 数据可视化。 (5分)

答案质量分析 (30分)

分析答案是否为“高质量”，基于以下因素：

1. 问题和答案创建之间的时间间隔。 (5分)
2. 提供答案的用户声誉。 (5分)
3. 另一个自定义的有意义因素。 (5分)

每个因素对应的可视化（每个因素 5 分，共 15 分）。

RESTful 服务 (20分)

- **话题频率查询服务 (10分)：**
  - 用户可以查询特定话题的频率。 (5分)
  - 用户可以查询按频率排序的前 N 个话题。 (5分)
- **错误/异常频率查询服务 (10分)：**

- 用户可以查询特定错误或异常的频率。 (5分)
- 用户可以查询按频率排序的前 N 个错误或异常。 (5分)

进一步说明

可视化

- 好的可视化应以清晰简洁的方式准确有效地传达信息。
- 选择合适的可视化方法，确保用户可以正确理解图表并了解数据的含义。同时，需关注美观性和用户交互性。
- 可视化部分的评分不仅限于功能性实现，还会综合考虑以上所有方面。

第16周展示候选项目 具有以下特点的项目可能被选中进行第16周的展示：

- 高质量、深入的数据分析，具有意义和有趣的见解。
- 有效传达分析结果的优秀可视化；出色的 UI/UX 设计。
- 能让项目脱颖而出的其他特点。

项目评分.md 2024-11-10

[CS209A-24Fall] 最终项目 (100分)

背景

在软件开发过程中，开发者会遇到各种问题，可能需要通过问答网站来发布问题并寻找答案。**Stack Overflow** 是一个专门为程序员设计的问答网站，属于 **Stack Exchange** 网络的一部分。
用户通过提问、回答、投票、编辑等方式参与平台活动，获得声誉积分和徽章，并解锁新特权。

在本项目中，我们将使用 **Spring Boot** 开发一个 Web 应用程序，存储、分析并可视化与 Java 编程相关的 Stack Overflow 问答数据，以了解 Java 编程相关的常见问题、答案和解决活动。

数据收集 (10分)

- 使用 **java** 标签定位 Stack Overflow 上与 Java 编程相关的线程。
- 收集至少 1000 个线程，并将数据存储在数据库（如 PostgreSQL、MySQL）或本地文件中，确保分析基于本地数据源进行。

第一部分：数据分析 (70分)

1. **Java 话题 (10分)**
   - 找出最常被问到的前 N 个 Java 话题，并使用图表展示。
2. **用户参与度 (15分)**
   - 计算用户在线程中的各种活动，并分析高声誉用户参与较多的话题。
3. **常见错误 (15分)**
   - 分析线程内容，找出常被讨论的错误和异常（如 **OutOfMemoryError**、运行时异常等）。
4. **答案质量 (30分)**
   - 调查影响“高质量”答案的因素，例如响应速度、用户声誉，以及另一个自定义因素。

第二部分：RESTful 服务 (20分)

1. **话题频率查询服务**
    - 用户可查询某话题的频率，或按频率排序的前 N 个话题。
2. **错误频率查询服务**
    - 用户可查询某错误的频率，或按频率排序的前 N 个错误。