University of Glasgow | School of Computing Science

评估课程

课程名	大数据 H / 大数据 M			
课程编号	1			
最后期限	时间：	23:59	日期：	2021 年 3 月 6 日星期日
% 对最终的贡献	20% (H) / 20% (M)			
课程标记				
单人或团体✓	独奏		团体	✓
预计小时数	每组成员 8 个			
提交说明	通过 Moodle 提交一份包含您的报告的 pdf 文件，以及一份按照详细说明的 uog-bigdata.zip 文件			
请注意：此课程不能重做				

课程作业提交评估规则代码

提交正式评估的课程作业的截止日期将在课程文件中公布，迟于截止日期提交的作业将受到如下规定的处罚。

在公布的截止日期之后提交的课程作业所获得的小学成绩和中学成绩将计算如下：

(i) 对于在截止日期后不超过五个工作日提交的工作
一种。工作将按照通常的方式进行评估；
湾。如此确定的小学成绩和中学成绩将在每个工作日（或工作日的一部分）被延迟提交的工作减少两个中学成绩。

(二) 在截止日期后超过五个工作日提交的作品将被授予 H 级。

如果延迟提交有正当理由，则不会对延迟提交课程作业进行处罚。您应该通过 MyCampus 提交支持正当理由的文件。

不遵守提交说明的处罚为 2 级

您必须通过以下方式填写“自己的作品”表格 https://webapps.dcs.gla.ac.uk/ETHICS 对于所有课程
除非通过 Moodle 提交

大数据 (H/M) 评估练习任务表

概括

本练习的目的是让您熟悉使用 Apache Spark 进行大数据分析任务的设计、实现和性能测试。您将需要设计和实现一个相当复杂的 Spark 应用程序。然后，您将在一个小数据样本上本地测试此应用程序的运行，并在整个数据集上在预先准备好的 Spark 集群上远程测试该应用程序的运行。最后，您将撰写一份简短的报告，描述您的设计、设计决策，并在适当的情况下对您的设计进行评论。您将根据代码功能（是否产生预期的结果）、代码质量（是否设计良好并遵循良好的软件工程实践）和效率（它的速度有多快以及是否有效地使用资源）以及您的提交的报告。

任务描述

您将在 Apache Spark 中开发基于批处理的文本搜索和过滤管道。该管道的核心目标是接收大量文本文档和一组用户定义的查询，然后对于每个查询，按与该查询的相关性对文本文档进行排名，并过滤掉所有过于相似的文档。最终排名。每个查询的前 10 个文档应作为输出返回。每个文档和查询都应该被处理以删除停用词（具有很少区分值的词，例如“the”）并应用词干（将每个词转换为其“词干”，一个较短的版本，有助于文档和查询之间的术语不匹配）。文件应使用评分DPH排名模型 . 作为最后阶段，应分析每个查询的文档排名以删除不需要的冗余（接近重复的文档），如果发现任何一对文档的标题的文本距离（使用提供的比较函数）小于 0.5，那么您应该只保留其中最相关的内容（基于 DPH 分数）。请注意，即使在冗余过滤之后，每个查询也应该返回 10 个文档。

您将获得一个 Java 模板项目，就像已经提供的教程一样。您的角色是实现必要的 Spark 函数，以从 Dataset<NewsArticle>（输入文档）和 Dataset<Query>（要排名的查询）到 List<DocumentRanking>（每个查询的 10 个文档的排名）。除了您选择在驱动程序中执行的任何最终处理之外，您的解决方案应该只包括 spark 函数。您不应执行任何“离线”计算（例如预先构建搜索索引），即所有处理都应在 Spark 应用程序的生命周期内进行。模板项目提供以下代码的实现来帮助您：

- 加载查询集并将其转换为 Dataset<Query>。
- 加载新闻文章并将其转换为 Dataset<NewsArticle>
- 一种静态文本预处理器功能，可将一段文本转换为其标记化、停用词删除和词干形式。此函数接受一个字符串（输入文本）并输出一个 List<String>（在标记化、词干提取和停用词删除之后来自输入的剩余术语）。

- 一个静态 DPH 评分函数，根据以下信息计算 <document,term> 对的分数：
    - 术语 文档中术语的频率（计数） 文档的长度（以术
    - 语为单位）
    - 语料库中的平均文档长度（以术语为单位） 语料库
    - 中的文档总数
    - 所有文档中术语的术语频率总和
- 一个静态字符串距离函数，它接受两个字符串并计算它们之间的距离值（在 0-1 范围内）。

<document,query> 对的 DPH 分数是每个 <document,term> 对（查询中的每个术语）的 DPH 分数的平均值。在设计解决方案时，您应该主要考虑如何有效地计算使用 DPH 为每个查询对每个文档进行评分所需的统计信息。

数据集

您将用于本练习的数据集是来自《华盛顿邮报》的新闻文章语料库以及一组查询。该数据集有两个版本：本地样本；和完整的数据集。本地示例可让您使用本地 spark 部署快速迭代您的设计（如教程中所示）。这将作为模板项目的一部分提供，包含 5,000 篇新闻文章和三个查询。完整的数据集由大约 670,000 个文档和 10 个查询组成。查询包含以下信息：

```
细绳原始查询;                          // 原始查询未更改
列表<字符串>查询条款;// 查询后面的词标记化 ,停用词
                                        删除和词干
短的[]queryTermCounts;// 每个查询词出现的次数
```

构建 Dataset<Query> 的预先提供的代码已经对查询执行了标记化、停用词删除和词干提取，其结果存储在 queryTerms 变量中。

一篇新闻文章包含以下信息：

```
细绳ID;// 唯一的文章标识符
细绳文章网址;//网址 指着网上的文章 细绳标题;// 文章标题 细绳作者;// 文
章作者

长发布日期;// 发布日期作为Unix时间戳 (小姐 ) 列表<内容项>内容;// 文章正文的内容
细绳类型;// 文章类型 细绳资源;// 新闻提供者
```

对于本练习，您只需要使用“id”、“title”和“contents”字段。以下是示例新闻文章的摘录。如您所见，新闻文章的“内容”是该新闻文章中的元素列表，例如踢脚线、标题元素、标题图像和各种段落。在计算文档的 DPH 时，您应该只考虑标题和 ContentItem 元素中具有非空子类型并且该子类型列为“段落”的术语。此外，如果一篇文章有 超过 5 个段落，您只需考虑前 5 个段落。提供的构建 Dataset<NewsArticle> 的代码不会应用任何文本预处理。

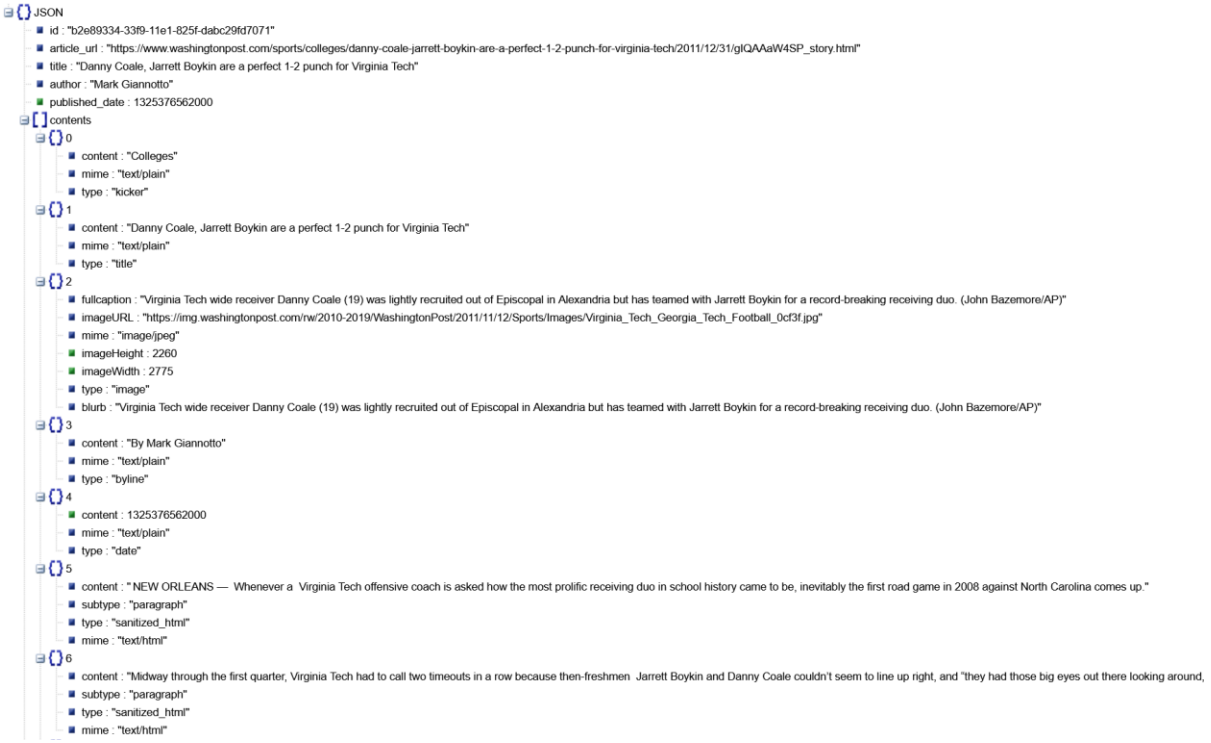

图 1：新闻文章示例

提交报告

除了您的 spark 代码库外，您还需要提交一份简短的报告，描述您的解决方案的设计逻辑。这应该是大约 2 A4 页面的总长度。您应该为您开发的每个 Spark 函数添加一个部分，其中对每个函数进行总结：

- 该功能的目标是什么
- 它如何适合您的整体管道

然后，在最后一部分，你应该包含一段讨论为什么你认为你的解决方案在项目要求的情况下是有效的，强调你为提高效率而做出的任何决定。您可能还希望强调您在实施过程中面临的任何挑战以及您如何克服这些挑战。

代码管理

对于这个项目，您将在 https://stgit.dcs.gla.ac.uk 为您的项目提供一个 git 存储库。您应该已经收到一封电子邮件，其中包含有关如何访问此存储库的详细信息。对于此项目，您需要将您每周工作的代码提交到此存储库。这个存储库有两个主要用途：

1. 我们的远程部署服务使用此存储库来获取您的源代码以进行远程测试（更多信息将在第 5-6 周进行）
2. 我们将此作为每个团队成员都在为项目做出贡献的证据来源

确保每个成员都可以访问存储库，并且他们的本地计算机配置为使用他们的学生帐户提交内容（以便我们可以追踪到他们）。

交什么

您应该通过 Moodle 提交：

- 包含您的最终报告的单个 pdf 文件。
- 从团队的 git 存储库下载的作为单个 zip 文件的代码库副本。

如何标记此练习

在 Moodle 上及时提交后，该练习将获得介于 0（未提交）和 20（各方面完美）之间的数字标记。然后数字标记将转换为带（A5、A4 等）。标记方案如下：

- 正确实施得 12 分（部分正确实施得部分分数）。
- 源代码的质量/可读性 2 分。
- 源代码文档的 2 分。
- 源代码的效率/可扩展性 4 分。