# 

**Spark Streaming的分析与应用**

**综合实验总结**

Version 1.0

小组成员：

陈少杰

姜鑫

蒲彦均

邹嘉欣

**版本变更记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本 | 变更时间 | 修改人 | 审核人 | 备注 |
| 1.0 | 2017/06/21 | 邹嘉欣、陈少杰  姜鑫、蒲彦均 | 邹嘉欣、陈少杰  姜鑫、蒲彦均 | 初稿 |
| 1.1 | 2017/06/21 | 邹嘉欣、陈少杰  姜鑫、蒲彦均 | 邹嘉欣、陈少杰  姜鑫、蒲彦均 | 添加概述和总结 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[一．概述 1](#_Toc485851695)

[二．各实验工作量 1](#_Toc485851696)

[2.1实验一——需求分析 1](#_Toc485851697)

[2.2实验二——需求评审 2](#_Toc485851698)

[2.3实验三——改进与扩展 2](#_Toc485851699)

[2.4实验四——测试需求 2](#_Toc485851700)

[2.5实验五——软件测试评审 3](#_Toc485851701)

[2.6实验六——项目计划与进度控制 3](#_Toc485851702)

[2.7实验七——配置管理 4](#_Toc485851703)

[2.8实验八——统计分析 4](#_Toc485851704)

[三．数据分析与说明 5](#_Toc485851705)

[3.1数据的完整性和准确性 5](#_Toc485851706)

[3.2影响因素与保证措施 6](#_Toc485851707)

[3.3偏差估计 6](#_Toc485851708)

[四．制品水平 6](#_Toc485851709)

[4.1软件需求、测试需求 6](#_Toc485851710)

[4.1.1软件需求 6](#_Toc485851711)

[4.1.2测试需求 7](#_Toc485851712)

[4.1.3评审的作用 7](#_Toc485851713)

[4.2软件改进与扩展 7](#_Toc485851714)

[4.2.1工作目标 7](#_Toc485851715)

[4.2.2软件基本功能 8](#_Toc485851716)

[4.2.3实验结果 8](#_Toc485851717)

[4.2.4测试的充分性（覆盖率） 8](#_Toc485851718)

[4.3进度计划与控制 8](#_Toc485851719)

[4.3.1实验过程及其特点 8](#_Toc485851720)

[4.3.2影响因素和监控方法 9](#_Toc485851721)

[4.3.3效果分析和说明 9](#_Toc485851722)

[4.4配置管理 9](#_Toc485851723)

[4.4.1影响因素 9](#_Toc485851724)

[4.4.2管理方法 9](#_Toc485851725)

[4.4.3管理效果的分析和说明： 10](#_Toc485851726)

[五．有效方法或实践举例 10](#_Toc485851727)

[六．结论与建议 13](#_Toc485851728)

# 一．概述

项目以Spark Streaming开源项目及相关资料为输入，分析软件设计需求，结合软件工程综合实验具体要求，基于Spark Streaming框架，以及机器学习、自然语言处理等技术，实现一个针对指定主题短文本的自动识别功能。并且根据项目内容完成了实验一至八的各项内容。其中实验一至二为软件设计的需求阶段，进行了本项目的需求分析和对其他组的需求评审。实验三完成了本项目的具体设计与改进，实验四到五进行了测试的需求分析和评审。实验六至八贯穿于整个实验过程。

# 二．各实验工作量

各实验的工作量将以表格的形式进行展现，表 1至表 8分别为实验一到实验八的工作量数据统计结果。

## 2.1实验一——需求分析

需求分析实验工作量统计结果如表 1所示。需求分析包含部分初期资料收集以及需求规格说明书编写、修改等内容。

表 1 需求分析工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 12196 |
| 用例数 | 5 |
| 制品复杂度 | 模型种类 | RUCM |
| 模型元素数量 | 5 |
| 文档修改次数 | 19 |
| 工时 | 工时（人时） | 90 |

## 2.2实验二——需求评审

需求评审实验工作量统计结果如表 2所示。需求评审主要包含需求规格说明书的评审，以及评审意见的汇总等工作。

表 2 需求评审工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 3941 |
| 制品复杂度 | 评审次数 | 6 |
| 提出问题数 | 40 |
| 收到并修改问题数 | 68 |
| 工时 | 工时（人时） | 41.5 |

## 2.3实验三——改进与扩展

改进与扩展实验工作量统计结果如表 3所示。改进与扩展主要包含环境配置、改进设计以及改进实现等内容。

表 3 改进与扩展工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 1638 |
| 代码行数 | 1500 |
| 制品复杂度 | 程序涉及功能模块 | 4 |
| 工时 | 工时（人时） | 123 |

## 2.4实验四——测试需求

需求评审实验工作量统计结果如表 4所示。测试需求主要包含测试需求的分析和《测试需求规格说明书》的编写和修改，以及实际测试及测试报告编写。

表 4 测试需求工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 11583 |
| 用例数 | 13 |
| 制品复杂度 | 模型种类 | RTCM |
| 模型元素数量 | 13 |
| 修改次数 | 7 |
| 工时 | 工时（人时） | 111 |

## 2.5实验五——软件测试评审

需求评审实验工作量统计结果如表 5所示。软件测试评审主要包含测试需求说明书的评审、问题整理等内容。

表 5 软件测试评审工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 3716 |
| 制品复杂度 | 评审次数 | 4 |
| 提出问题数 | 17 |
| 收到并修改问题数 | 52 |
| 工时 | 工时（人时） | 23 |

## 2.6实验六——项目计划与进度控制

项目计划与进度控制实验工作量统计结果如表 6所示。进度计划主要包含《项目计划书》编写和修改、小组会议、会议记录、工作日志、和MPP文件修改等内容。

表 6 项目计划与进度控制工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 15300 |
| 制品复杂度 | 更新次数 | 13 |
| 分解任务项数 | 15 |
| 工时 | 工时（人时） | 157 |

## 2.7实验七——配置管理

配置管理实验工作量统计结果如表 7所示。配置管理主要包含github的学习、配置管理文档的编写和修改。具体上传文档到github的工作分散到其他实验中了。

表 7 配置管理工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 7862 |
| 制品复杂度 | 上传次数 | 182 |
| 文档版本更新次数 | 3 |
| 工时 | 工时（人时） | 14 |

## 2.8实验八——统计分析

统计分析实验工作量统计结果如表 8所示。统计分析主要包含统计分析报告、实验总结以及综合实验报告等内容。

表 8 统计分析工作量统计

|  |  |  |
| --- | --- | --- |
| 统计项 | 统计小项 | 统计结果 |
| 制品规模 | 字数 | 8011 |
| 统计分析项数 | 15 |
| 制品复杂度 | 涉及文档数量 | 183 |
| 工时 | 工时（人时） | 62 |

# 三．数据分析与说明

## 3.1数据的完整性和准确性

第二章的数据均来自实际的统计结果。各个实验的工时，是对mpp文件中所有子任务按照实验内容来划分，进而统计得到的结果。统计中用到的mpp文件也会上传github。对于文档的字数统计均使用word软件进行。

对于需求分析，这里主要统计的内容是《需求规格说明书》。因为该文档统计的重点不在于文档中包含的字数，因而这里增加了文档修改次数作为制品复杂度的内容。

对于需求评审，主要统计检查单、评审问题清单的字数。这部分并未涉及到模型、用例、代码等。因而这部分制品复杂度的衡量主要依靠评审次数、提出问题数和修改问题数。

对于改进与扩展，主要统计《概要设计文档》字数和代码行数。制品复杂度通过设计到功能模块数量来衡量。

对于测试需求分析，主要统计《测试需求规格说明书》的字数和包含的用例数。制品复杂度除了模型种类和数量以外，还包含了修改次数。

对于软件测试评审，主要统计检查单和问题清单的字数。而该实验重在在于组件互评和修改，因而复杂度主要依靠评审次数，提出问题数、修改问题数来衡量。

对于项目计划和进度控制统计《项目计划书》、小组会议、会议记录、工作日志的字数。这里精确统计了所有文档的字数，最终得到总字数。而制品复杂度依靠更新次数，分解任务数来衡量。

对于配置管理，主要统计配置《变更与管理分析报告》字数，制品复杂度主要依靠上传次数和文档版本更新次数来衡量。

对于工作量统计，主要统计《工作量估计与统计分析》字数和分析的项数。制品复杂度主要靠其设计到的文档的数量来衡量。

## 3.2影响因素与保证措施

对于需求文档和测试需求文档的影响，其经历过多次修改。而这些在字数统计中并不能体现。因而这里通过增加文档更新次数来保证数据的有效性。对于工时统计，其最大的影响因素是工作日志中工时的估计准确度、以及任务计划与实际执行内容上的差别。所以每周初制定本周计划， mpp文件中任务分解要尽量细致。因而临时的任务变更会导致进度控制的不准确。而且，由于任务都很细致，我们无法对于每个小任务的实际工时进行精确记录。这样导致最终统计出来的工时等信息会更加不准确。因而，对于临时变更的任务，要及时在mpp文件中进行调整，以保证mpp文件中记录的工作与实际相符。在每周的工作日志提交之后，组长要对工作日志进行审核，防止出现估计偏差太大的工时。对于github上传次数，由于本组很多交流通过微信群进行，因而提交次数不能很准确的反映更新的次数。

## 3.3偏差估计

对于所有的文档，其字数统计结果均来源于对每个文档的字数统计，可以保证其正确性。对于工时统计，由于本组对工作日志的审核较严格，并依据个人经验进行把控，可以保证每个子任务的实际工时与记录工时的偏差控制在30%以内。由于各任务的工时记录不会全部偏高或全部偏低，因而总工时偏差可以认为在20%以内。对于文档更新次数的统计可能比实际更新次数要少得多，可能缺少一半左右。

# 四．制品水平

## 4.1软件需求、测试需求

### 4.1.1软件需求

在软件需求规格说明书中，我们对Spark的相关核心概念进行了梳理，总结了Spark Streaming的功能与非功能需求，并在调研中确定了以应用开发为核心的实验任务，确定了要实现的功能细节并对实现过程进行了设计明确。

在开发、测试等后续工作中，软件需求规格说明书起到了极大的指导性作用，明确了工作目标，规范了名词术语等。

软件需求说明书的编写过程，也是小组成员对Spark Streaming的了解学习过程，正是在不断的了解中才确定并不断修正了我们的实验目标。

### 4.1.2测试需求

测试需求规格说明书是对测试工作的指导性文件，在编写过程中，我们曾经出过与软件需求规格说明书中的术语不一致的情况，但是经过老师的及时发现与提醒，我们也快速进行了修正。

测试需求说明书对测试工作起到了指导作用，保证了测试的充分性、规范性与合理性。在测试需求规格说明书的指导下，测试工作逐条展开，对Spark Streaming与开发的应用都做了详尽的测试。

### 4.1.3评审的作用

在文档的每一个版本更新中小组成员都会对这一版本修改的地方进行审核校对，然而随着工作的不断进行，组内的校对强度渐渐地无法得到较好的保障，组间评审的必要性就凸显了出来。组间评审为我们的各个文档都提出了规范性，完整性，一致性等方面的问题。

## 4.2软件改进与扩展

### 4.2.1工作目标

项目计划基于Spark Streaming框架，以及机器学习、自然语言处理等技术，实现了一个针对指定主题短文本的自动识别功能。本软件产品是一项独立的软件，用户根据软件提供的调用命令实现各种指定功能。

### 4.2.2软件基本功能

本软件以格式化的微博数据为输入，经过软件的一系列处理流程，输出格式化的微博以及其分类标签。输入数据包含微博的所有者，发送时间，微博正文三个字段，输出数据包含以上三个字段以及微博分类标签。

### 4.2.3实验结果

本小组对计划中的应用完成了基本功能的开发，实现了数据贯通，达到了预期的功能效果。

### 4.2.4测试的充分性（覆盖率）

在测试工作中，我们遵照测试需求规格说明书中的测试点，测试规范，对Spark Streaming和本小组的应用进行了全面的测试。而测试需求规格说明书中的测试用例又是来源于软件需求规格说明书中的需求用例。测试工作的充分性也因此得到了保障，这都是严格遵照文档书写规范的结果。

## 4.3进度计划与控制

### 4.3.1实验过程及其特点

进度计划与控制实验贯穿整个项目周期，每周都要对项目进行跟进。且每周的跟进结果的准确性，会很大程度的影响最终进度计划与控制结果的准确性。

进度计划与控制实验使用的最主要的工具是MS Project。由于进度计划与控制不是一人可以完成的，每次都需要组内进行讨论，最终由组长进行会议记录并更新MS Project生成的mpp文件。在每周结束时，组员各自写本周工作日志，日志中记录本周工作内容及各项工作的工时。组长根据工作日志用MS Project进行记录。最后利用MS Project生成的统计结果进行分析，进而完成进度计划与控制实验。

### 4.3.2影响因素和监控方法

对于进度计划与控制实验，其最大的影响因素是工作日志中工时的估计准确度、以及任务计划与实际执行内容上的差别。

由于本实验要求每周初制定本周计划，得到的mpp文件中任务要尽量细致。因而临时的任务变更会导致进度控制的不准确。而且，由于任务都很细致，我们无法对于每个小任务的实际工时进行精确记录。这样导致最终统计出来的工时等信息会更加不准确。

因而，对于临时变更的任务，要及时在mpp文件中进行调整，以保证mpp文件中记录的工作与实际相符。在每周的工作日志提交之后，组长要对工作日志进行审核，防止出现估计偏差太大的工时。

### 4.3.3效果分析和说明

本组对mpp文件的安排按照周来进行，每个大任务都对应着一周的工作内容。因而可以很容易看出来每周的计划工时与实际工时等内容，也能与每周的工作日志很好的对应起来。由于本组对每周工作日志的要求不够严格，未能每周按时提交，因而对工时估计结果会增大偏差程度。但由于本组对工作日志的审核较严格，并依据个人经验进行把控，可以保证每个子任务的实际工时与记录工时的偏差控制在30%以内。由于各任务的工时记录不会全部偏高或全部偏低，因而总工时偏差可以认为在20%以内。

## 4.4配置管理

### 4.4.1影响因素

主要影响因素有人员分工，任务时间安排等

### 4.4.2管理方法

通过github进行配置管理。负责人为我们小组的四位成员，每周各自的任务完成之后会直接上传到github中。有时并不能保证每日更新，但每当项目有新的进展，比如更新了某个文档，就会将该文档上传到github中，并进行简要的更新说明。

关于汇总合并，一般是由某个成员先将自己完成的部分上传至github,再由其他成员在这个基础上继续修改。每次修改都会在文档的版本变更记录中进行说明。

### 4.4.3管理效果的分析和说明：

通过github进行配置管理，小组成员之间可以有效的协同工作。同时可以避免版本混乱。此外，github提供了历史的记录，追踪和分析功能。

# 五．有效方法或实践举例

|  |  |
| --- | --- |
| 软件需求分析 | |
| 目标 | 确定软件的需求，并撰写《软件需求规格说明书》 |
| 问题特点说明 | 此时处于项目前期，项目成员之间需要磨合，对项目也是刚刚接触的阶段，了解不深。此外，由于组员之间对选定的项目spark streaming的了解程度参差不齐，如何确定分工也是需要解决的问题。 |
| 实践要点 | 首先确定每位成员对我们组的项目spark streaming的了解程度：项目成员姜鑫和蒲彦均对项目有一些了解，而陈少杰和邹嘉欣对spark streaming基本没有了解，所以让前两位侧重于代码，后两位侧重文档。当然所有成员都需要加强对项目的学习和了解。此外，我们会共同商讨当前需要解决的问题，并把任务细分到个人。 |
| 结果 | 成功撰写出《软件需求规格说明书》 |
| 注意事项 | 在商讨过程中，如果分配给自己的任务有异议或者疑虑，一定要及时提出，大家共同沟通。 |

|  |  |
| --- | --- |
| 软件需求评审 | |
| 目标 | 对D组和E组的软件需求评审 |
| 问题特点说明 | 本组同学对其他两组的项目了解比较少，通过看他们的文档很难找出内容上的问题，大部分只是语言文字排版方面的问题。 |
| 实践要点 | 本组成员需要对D组和E组项目进行初步的了解，并且在每周的汇报时着重关注这两组的讲解。此外，利用好微信和课程中心这些交流平台，遇到问题及时请教其他同学。 |
| 结果 | 完成对D组、E组的软件需求评审，编写相应的问题清单 |
| 注意事项 | 应该提前了解一下其他两组的项目。 |

|  |  |
| --- | --- |
| 改进和扩展 | |
| 目标 | 利用spark streaming 平台对微博数据进行分类 |
| 问题特点说明 | 需要确定整个设计实现的流程，以及文本分类的模型训练工具。每个设计实现步骤采用的方案也需要进一步确定。此外，可能存在一些诸如分布式模型调用等方面的工程问题 |
| 实践要点 | 小组成员共同讨论设计实现的基本方案，确定了详细实现方案。该方案主要包括四部分：模型训练、数据的产生、文本预处理以及文本的分类。 |
| 结果 | 撰写《概要设计文档》，并确定项目实现方案 |
| 注意事项 | 在进行编码实现之前需要确定整体方案。 |

|  |  |
| --- | --- |
| 测试需求分析 | |
| 目标 | 确定测试需求，撰写《测试需求规格说明书》 |
| 问题特点说明 | 需要确定项目的测试需求，设计需求用例。由于之前项目组的四位成员都没有软件工程方面的经验，所以对如何设计测试用例比较困惑。 |
| 实践要点 | 大家一起商讨之后确定了分工，并且在GitHub上找到去年的软件工程试验小组的测试需求文档中的测试用例作为参考，对于老师的课件也进行了仔细的研究。 |
| 结果 | 完成《测试需求规格说明书》的撰写 |
| 注意事项 | 《测试需求规格说明书》的测试项应该与《软件需求规格说明书》的需求项对应 |

|  |  |
| --- | --- |
| 测试需求评审 | |
| 目标 | 对D组、E组的测试需求进行评审 |
| 问题特点说明 | 需要对两组的测试策略进行了解，才能更好地对D、E两组的测试用例做出评审。 |
| 实践要点 | 我们一起开会确定了测试需求评审的分工，姜鑫和蒲彦均主要侧重于对D组进行评审，邹嘉欣和陈少杰则侧重于对E组进行评审。在评审过程中，如果遇到问题通过微信或者课程中心的讨论区积极与D、E组成员交流。 |
| 结果 | 完成对D组、E组的测试需求的评审， |
| 注意事项 | 评审过程中遇到问题可积极与对方组的成员沟通。 |

|  |  |
| --- | --- |
| 进度计划与控制 | |
| 目标 | 利用Microsoft project对项目进行计划的制定与进度的控制 |
| 问题特点说明 | 由于我们组的四位成员都没用接触过这个软件，所以最初在使用上有很多困惑。 |
| 实践要点 | 经过开会讨论之后，决定让组长陈少杰来对该软件进行深入的学习，并由他负责后续与项目进度计划与控制相关的内容。 |
| 结果 | 顺利完成项目进度计划与控制方面的实验内容。 |
| 注意事项 | 遇到不熟悉的软件，应该积极学习如何使用。 |

|  |  |
| --- | --- |
| 配置管理 | |
| 目标 | 利用Github进行配置管理 |
| 问题特点说明 | 在项目前期阶段，项目组成员对Github如何使用不太熟悉；对于文档的不断更新，怎么利用Github来控制版本迭代。 |
| 实践要点 | 项目前期阶段，因为蒲彦均对Github的基本操作较为熟悉，在他的帮助之下，其他三位成员也学习了Github的基本操作，并且可以完成文档的上传与同步等。每周各自的任务完成之后会直接上传到github中。有时并不能保证每日更新，但每当项目有新的进展，比如更新了某个文档，就会将该文档上传到github中，并进行简要的更新说明。一般是由某个成员先将自己完成的部分上传github,再由其他成员在这个基础上继续修改。每次修改都会在文档的版本变更记录中进行说明。 |
| 结果 | 顺利完成配置管理方面的实验内容。 |
| 注意事项 | 在上传新文档之前，应该先同步Github上的内容。 |

# 六．结论与建议

在小组的全体成员经过15周的共同工作与努力，我们完成项目预期的全部工作，其中完成了在实验一种提出的5个功能性需求和5个非功能需求。并且达成了计划的工作目标­基于Spark Streaming框架，以及机器学习、自然语言处理等技术，实现一个针对指定主题短文本的自动识别功能。在需求评审阶段，制定了对应的需求规格说明书检查单并对目标组按照检查单的格式进行了检查。在概要设计阶段，设计了概要设计及详细实现方案，并针对项目的实际情况提出了技术难点来重点突破。在完成设计和实现之后，针对本组的项目制定了相应的测试需求规格说明书，同样和需求规格说明书中的功能需求和非功能需求以及设计实现相对应。有三大测试需求。之后根据测试需求形成了本组的测试报告及对互评组的测试报告。实验六七八贯穿于整个实验，分别是实验的任务计划，我们在项目进行期间不定期的根据项目实际需求修改和添加项目计划，合理的项目计划对我们项目的有序顺利完成提供了很大的帮助，配置管理实验让我们小组的协助更方便和科学，每一个人修改的版本都可以保存和回滚，并且我们可以分析我们整个项目期间的修改情况。实验八更科学的统计了我们每一个人在这个项目中的贡献程度以及让我们更科学的进行任务分工。总的来看，我们成功的完成了软件工程试验的课程要求，经历每一个实验阶段都能让我们学习到新的软件工程知识及发现软件工程实践中的一些实际问题。感谢两位老师一学期的努力，让我们有收获，有进步。