**长春建筑学院**

**《Spark应用实训》**

**选题及任务书**

人工智能产业学院**一、课程设计选题**

**项目一：使用GitHub建立个人博客（必选）**

1．项目来源及背景

随着协同办公的迅速发展，掌握分布式的版本控制方法成为了必然要求。如何使用当今最为流行的Github进行版本控制也成为了大数据专业人才的必要条件。本项目旨在让学生学会Git的同时，掌握Github的基本操作，并可以使用简单的建站工具在Github上进行个人博客的部署，完成相应的知识输出。

2．项目需求分析

（1）个人博客采用Hexo进行相应搭建。地址：<https://hexo.io/zh-cn/>

（2）在使用Hexo前，需要安装node和npm。

（3）使用Hexo，选择相应的主题进行博客搭建。

（4）具体搭建步骤在如下示例中。

示例：<https://skykip.github.io/2022/04/10/hello-world/>

所采用主题为：Wikitten

**（5）请各位同学在搭建完成个人博客后，将之前spark课程中做的分组任务（spark全流程配置），以博文的形式发布到自己的博客中进行分享。**

其中：

1. 基础环境，JDK，Hadoop，Zookeeper作为一篇博文，

名称为：《Spark基础环境配置》

1. Spark(local)，Spark(stand-alone)作为一篇博文，

名称为：《Spark local& stand-alone配置》

1. Spark(HA)，Spark(Yarn)作为一篇博文，

名称为：《Spark HA & Yarn配置》

**各组同学需要在老师所列出的主题中进行选择（各组不能相同）：**

1. **zhl主题**：

主题示例：<https://lizehongss.github.io/>

主题github地址：<https://github.com/lizehongss/hexo-theme-zhl>



1. **MeiliDu主题**：

主题示例：<https://meilidu.github.io/>

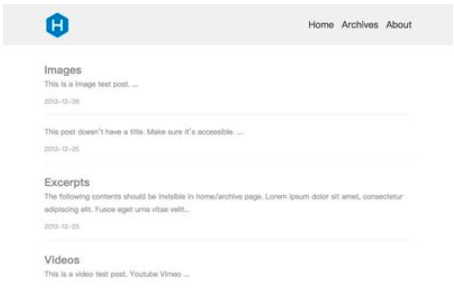
主题github地址：<https://github.com/HoverBaum/meilidu-hexo>



1. **iLiKE主题**：

主题示例：<https://caichenghan.github.io/>

主题github地址：<https://github.com/CaiChenghan/iLiKE>



1. **Volantis主题**：

主题示例：<https://volantis.js.org/>

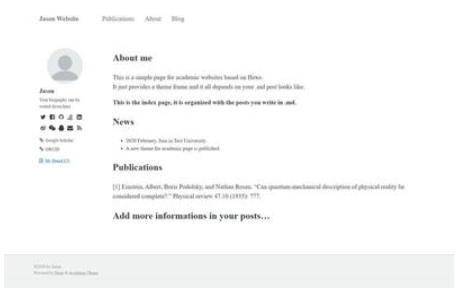
主题github地址：<https://github.com/volantis-x/hexo-theme-volantis>



1. **Academia (学术界)主题**：

主题示例：<https://phosphorw.github.io/>

主题github地址：<https://github.com/PhosphorW/hexo-theme-academia>



1. **Yilia-plus主题**：

主题示例：<https://joeybling.github.io/>

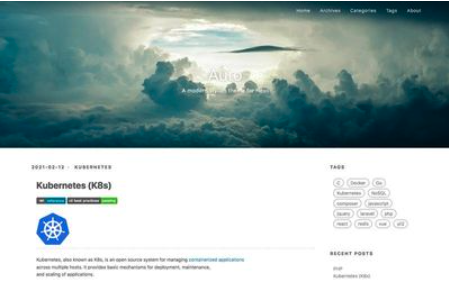
主题github地址：<https://github.com/JoeyBling/hexo-theme-yilia-plus>



1. **Auto主题**：

主题示例：<https://autoload.github.io/>

主题github地址：<https://github.com/autoload/hexo-theme-auto>



1. **Tangyuxian主题**：

主题示例：<https://www.tangyuxian.com/>

主题github地址：<https://github.com/tangyuxian/hexo-theme-tangyuxian>



1. **white主题**：

主题示例：<https://fushaolei.github.io/>

主题github地址：<https://github.com/FuShaoLei/hexo-theme-white>



1. **Wikitten主题**：

主题示例：<https://wiki.zthxxx.me/>

主题github地址：<https://github.com/zthxxx/hexo-theme-Wikitten>



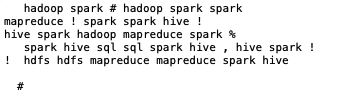
**项目二：带特殊字符的单词计数（任选）**

1．项目来源及背景

Wordcount是Spark中经典且最为基础的基本任务模板，本项目在此基础上加入了特殊字符，在RDD基础知识点的基础上，同时考察了RDD缓存与累加器的使用方法，是一个Spark数据分析的综合案例。

2．系统需求分析

（1）数据集示例：



（2）特殊字符定义如下：

abnormal\_char = [",",".","!","#","$","%"]

（3）使用RDD算子进行正常的单词计数。

**输出示例：**

正常单词计数结果: [('hadoop', 3), ('hive', 6), ('hdfs', 2), ('spark', 11), ('mapreduce', 4), ('sql', 2)]

（4）统计特殊字符出现有多少个。

**输出示例：**

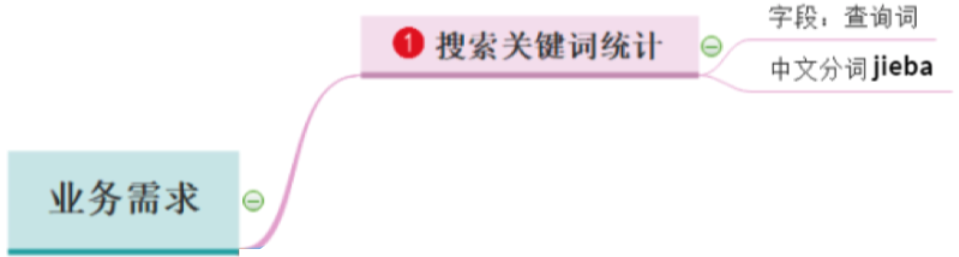
特殊字符数量: 8

**项目三：搜索引擎日志分析-搜索关键词统计（任选）**

1．项目来源及背景

搜索引擎的日志分析作为大数据分析的基本分析任务，是非常适合Spark课程的基本实训内容。其中，对日志中的关键词进行统计是日志分析首要任务，本项目旨在使用jieba库和RDD相关算子，完成用户搜索的关键词分析。

2．需求分析



根据搜索引擎日志，本项目需要完成如下的主要任务：

（1）读取日志数据，并使用jieba库进行分词处理

（2）设计RDD算子完成用户搜索的关键`词`分析

（3）主要分析热点词

（4）将所有的搜索内容取出

（5）对单词进行 分组 聚合 排序 求出前5名

**输出示例：**

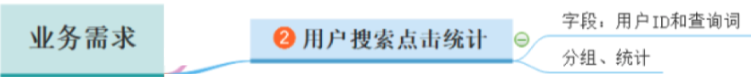
需求1结果: [('学院', 8413), ('智能', 2800), ('建筑', 2716), ('scala', 2310), ('hadoop', 2268)]

**项目四：搜索引擎日志分析-用户搜索点击统计（任选）**

1．项目来源及背景

搜索引擎的日志分析作为大数据分析的基本分析任务，是非常适合Spark课程的基本实训内容。其中，用户搜索的点击统计，是生成用户画像的基本方法，本项目旨在通过使用jieba库和RDD相关算子，完成用户搜索点击统计分析。

2．系统需求分析



根据搜索引擎日志，本项目需要完成如下的主要任务：

（1）读取日志数据，并使用jieba库进行分词处理

（2）设计RDD算子完成用户和关键词组合分析

（3）对用户的搜索内容进行分词, 分词后和用户ID再次组合

（4）对单词进行 分组 聚合 排序 求出前5名

**输出示例：**

需求2结果:

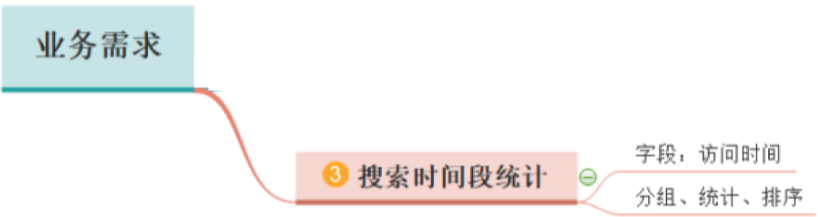
[('6185822016522959\_scala', 2016), ('41641664258866384\_智能', 1372), ('41641664258866384\_学院', 1372), ('44801909258572364\_hadoop', 1260), ('7044693659960919\_数据', 1120)]

**项目五：搜索引擎日志分析-搜索时间段统计（任选）**

1.项目来源及背景

搜索引擎的日志分析作为大数据分析的基本分析任务，是非常适合Spark课程的基本实训内容。其中，搜索时间段统计同样也是日志分析的基本任务，本项目旨在通过使用jieba库和RDD相关算子，完成热门搜索时间段分析。

2.项目需求



根据搜索引擎日志，本项目需要完成如下的主要任务：

（1）读取日志数据，并使用jieba库进行分词处理

（2）设计RDD算子完成热门搜索时间段分析

（3）取出日志中的所有的时间（0点至23点）

（4）对时间进行处理, 只保留小时精度即可

（4）对得到的分析结果进行分组 聚合 排序

**输出示例：**

需求3结果: [('20', 3479), ('23', 3087), ('21', 2989), ('22', 2499), ('01', 1365), ('10', 973), ('11', 875), ('05', 798), ('02', 756), ('19', 735), ('12', 644), ('14', 637), ('00', 504), ('16', 497), ('08', 476), ('04', 476), ('03', 385), ('09', 371), ('15', 350), ('06', 294), ('13', 217), ('18', 112), ('17', 77), ('07', 70)]

**二、设计要求**

每个小组必须选择项目一，同时在项目二、三、四、五中任选一题，相同题目者设计报告不能雷同；

2.课程设计报告要做到层次清晰，论述清楚，图表正确，书写工整；

3.成绩根据平时、设计报告和答辩给定。

**三．设计任务**

1.使用Hexo + Github，选择主题，搭建个人博客。

2.将spark相应的配置内容作为博文在个人博客上进行发布。

3.使用pycharm远程连接远程服务器的spark环境进行项目开发。

4.根据需求，设计Spark RDD的构建方式

5.使用RDD的缓存和检查点机制

6.使用RDD的共享变量（广播变量与累加器）

7.请同学们在设计时注意存储备份，以免机器故障、或其他误操作而丢失程序。

8.在硬盘上建立自己的目录，将程序文件存储到该目录下，即有利于调试文件，又保证了文件的安全性。

**四．设计参考资料**

《Spark快速大数据分析》[美] 卡劳（Holden Karau），[美]肯维尼斯科（Andy Konwinski），[美]温德尔（Patrick Wendell），[加]扎哈里亚（Matei Zaharia）著，王道远 译、人民邮电出版社、2015年

<https://yelog.org/2017/03/23/3-hexo-instruction/>

<https://yelog.org/2016/10/02/git-command/>

<https://wiki.zthxxx.me/wiki/index/>

https://ohlia.github.io/Wiki-site/wiki/Hexo/build-blog-by-hexo/