项目设计文档——胡洋

项目负责板块：

1.前期与小组成员搭建hadoop和spark 集群

2.django 与MongoDb，spark的处理并展示数据

3.优化搜索引擎

前期搭建hadoop和spark平台：

hadoop安装与配置：

首先下载java安装并配置环境变量

并且相互设置免密登录

1.下载hadoop

2.解压安装，配置变量环境

3.配置core-site和 hdfs-site（确认好和

namenode ）

4.格式化namenode 启动并确认集群是否搭建好

spark按照配置：

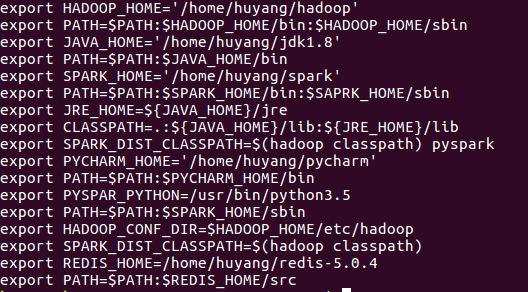
1.下载spark

2.解压安装，配置变量环境

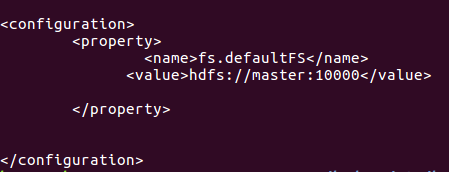
3.配置slave和master

总结：整体完成下来，大致的效果如下：

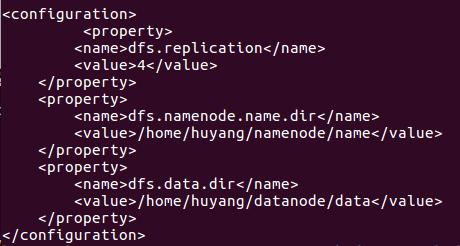
配置环境变量：



hadoop core-site.xml设置如下：



hadoop hdfs-site.xml设置如下：



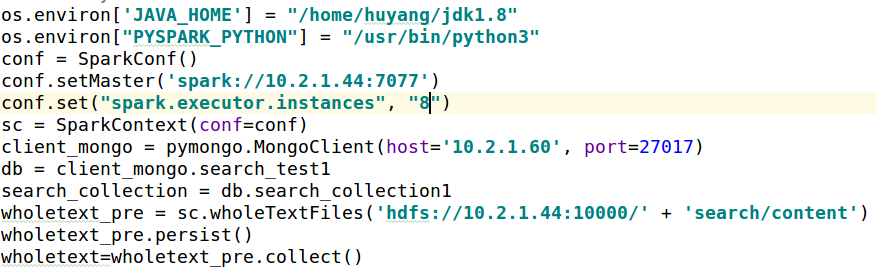
spark主要配置一下slave和master，以及其变量环境

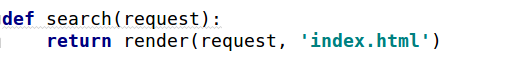
Django 和spark，MongoDb的数据提取：

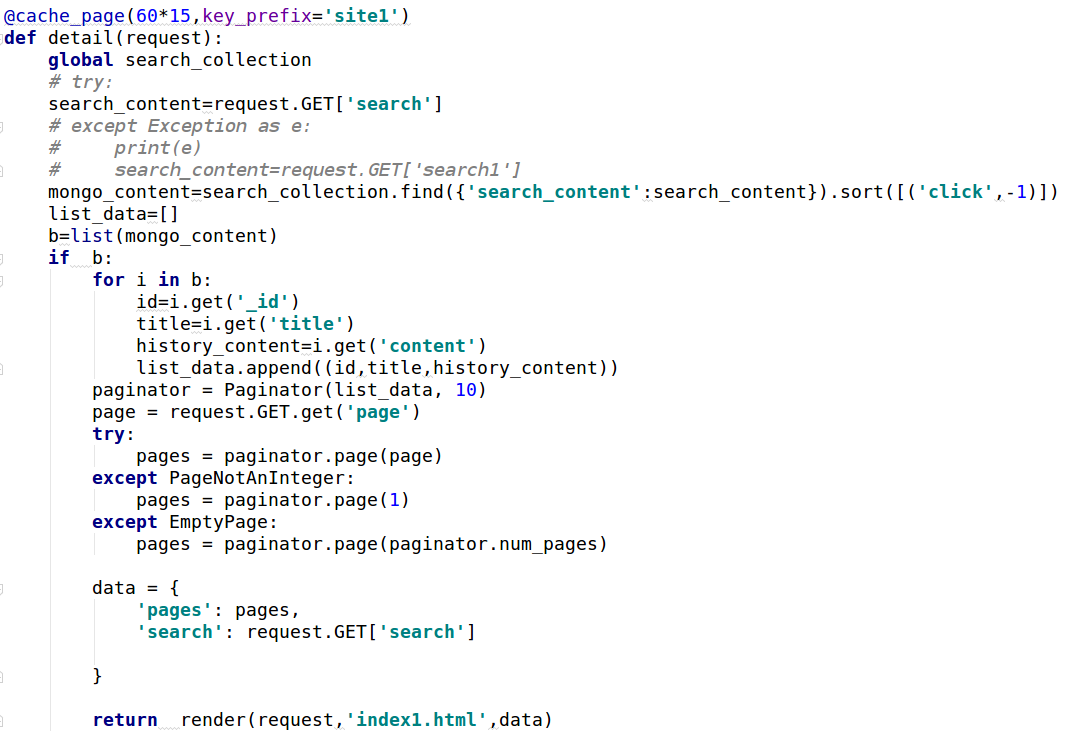
在这个过程中，主要运用到三个库做数据提取、计算、展示，分别是：spark 对hadoop中的数据读取，这里主要是应用到

wholeTextfiles的方法对hadoop整个文件夹内的所有文件进行读取，读取格式，以及相应的信息和存储方式和组员协商好，可以很方便的提取 Mongodb的数据读取主要是采用的是pymongo这个库来实现的 Django主要用于前端数据展示和页面缓存，当然，还有一个就是实现根据点击排序，这个主要是在MOngo中实现的，通过一个链接跳转，然后记录好数据中的click，用自增1的方式记录并跟新数据库，大致的代码如下：

各种参数设置：

首页的函数视图：

查询页面函数视图：



跳转页面的函数视图：

