项目题目

# 项目简介

使用爬虫抓取各类新闻，在网站上展示,自动推送新闻。如果可能，希望可以出app。

# 二．项目市场分析

满足人们闲暇时间看新闻的需求

# 三．项目组成员

组长：施赛华

组员：施赛华 周伦浩 张亚明 王伟良

时间： 2015/10/10 ~ 2015/10/19

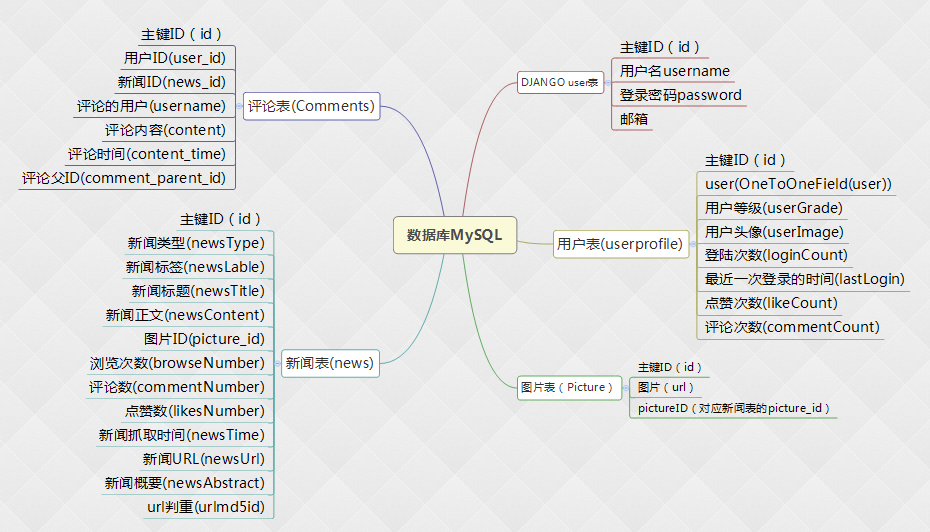
项目版本: v1.0

# 四．概要设计



# 五．数据库设计

# 六．模块详细设计



1, app目录为app应用模块

2，spider\_web目录为项目管理目录

3，static 存放静态资源文件

4，template 模板文件

5，见代码注释

# 七．总体功能设计

1. 数据库模型的设计

2. 用python 的scrapy爬虫框架抓取网络上的新闻，存入数据库

1. 在网页上显示的时候可以依照分类标签，做分类显示。用户可以浏览新闻，也可以评论新闻。

4. 用户注册，登录，修改密码（也许可以实现交友功能）

5. 根据用户的浏览记录，分类推送给用户最感兴趣的新闻

# 八．已完成功能

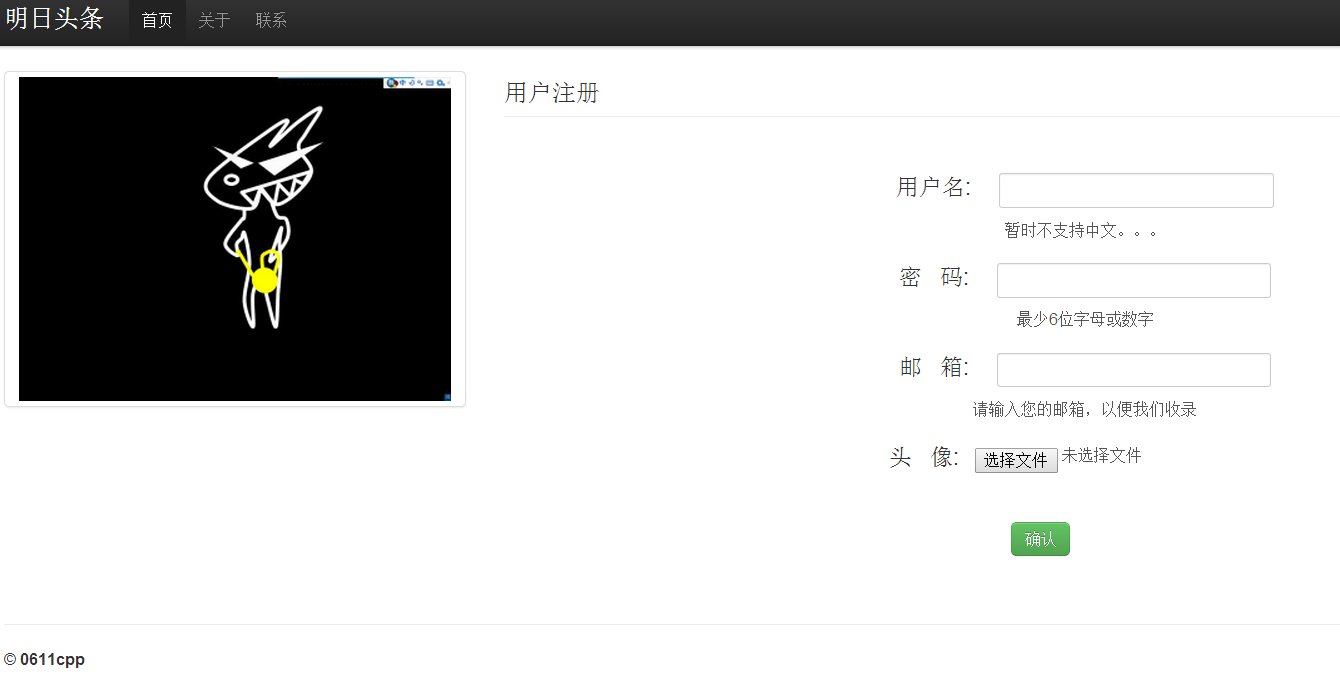
1. Crawl 将网页内的新闻内容和标题以及图片链接抓下来，存到数据库中。

2，将新闻做一下简单的分类，例如：<http://money.163.com/15/1019/10/B69J1HAG00252G50.html>

可以将链接中的关键字money作为分类词，其他新闻依此类推。

3.在网页上显示的时候可以依照分类标签，在新闻标题前显示。用户可以浏览评论新闻。

# 九．效果图



# 十．待完成功能

1. 爬虫抓取其他网站的新闻，存入数据库

2. 网页排版优化

3. 点赞功能

4. 实现新闻分类显示

5. 根据用户的浏览记录，分类推送给用户最感兴趣的新闻

# 十一.代码github地址

https://github.com/yoyomaster/spider\_web

# 十二.项目使用说明

1, 爬虫(使用scrapy 框架)(环境建议装在python虚拟环境)

环境搭建

安装openssl

Wget http://www.openssl.org/source/openssl-1.0.0r.tar.gz

pip install lxml

pip install Scrapy（注意是大写S）

pip install scrapyd

pip install scrapy-crawlera

sudo apt-get install python-lxml

1.1 scrapy 框架使用

1.1.1创建一个新的Scrapy项目

在虚拟环境下

在抓取之前，你需要新建一个scrapy 工程。进入一个你想用来保存代码的目录，然 后执行：scrapy startproject news

这些文件主要是：

* scrapy.cfg: 项目配置文件
* news/: 项目python模块, 代码将从这里导入
* news/items.py: 项目items文件
* news/pipelines.py: 项目管道文件
* news/settings.py: 项目配置文件
* news/spiders: 放置spider的目录
* news/commands：放置命令行文件（可以同时开始多个爬虫）

1.1.2定义提取的Item

1.1.3写一个Spider用来爬行站点，并提取Items

1.1.4写一个Item Pipeline用来存储提取出的Items

其中最主要的是定义爬取规则和将数据存入数据库

2，django环境搭建

3，uwsgi web服务与nginx服务环境搭建

4， 数据库使用mysql

创建数据库 create database spider\_web character set utf8

Django: python manager.py migrate 生成数据库表

5，在爬虫项目目录下（spider\_163目录下 ）

使用命令 scrapy crawlall 命令 启动爬虫

（scrapy crawl 爬虫名字）