软件开发计划书

# 概述

Scrapy是一个为了爬取网站数据，提取结构性数据而编写的应用框架。 可以应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。

所谓网络爬虫，就是一个在网上到处或定向抓取数据的程序，当然，这种说法不够专业，更专业的描述就是，抓取特定网站网页的HTML数据。抓取网页的一般方法是，定义一个入口页面，然后一般一个页面会有其他页面的URL，于是从当前页面获取到这些URL加入到爬虫的抓取队列中，然后进入到新页面后再递归的进行上述的操作，其实说来就跟深度遍历或广度遍历一样。

Scrapy 使用 Twisted这个异步网络库来处理网络通讯，架构清晰，并且包含了各种中间件接口，可以灵活的完成各种需求。

Scrapy有以下特点：

* 采取可读性更强的xpath代替正则
* 强大的统计和log系统
* 同时在不同的url上爬行
* 支持shell方式，方便独立调试
* 写middleware,方便写一些统一的过滤器
* 通过管道的方式存入数据库

# 项目内容

## 工作内容

选择Scrapy最新版本，了解Scrapy框架和用法，进行针对某网站的爬虫代码的编写，并且进行分析，具体内容如下：

* 熟悉scrapy使用方法。
* 熟悉scrapy架构
* 使用scrapy设计针对某网站的爬虫
* 分析当前爬虫的性能、功能，找出可以改进的地方。
* 阅读scrapy相关源码，针对当前爬虫做出适应性的修改
* 测试修改

## 主要参加人员

郭炜锋： SY1606324， 北航研究生计算机软件工程专业研一学生；

王益飞： SY1606116， 北航研究生计算机软件工程专业研一学生；

胡 勇： SY1606322， 北航研究生计算机软件工程专业研一学生；

武丁泽宇：SY1606416，北航研究生计算机软件工程专业研一学生。

# 项目计划

## 工作任务分工

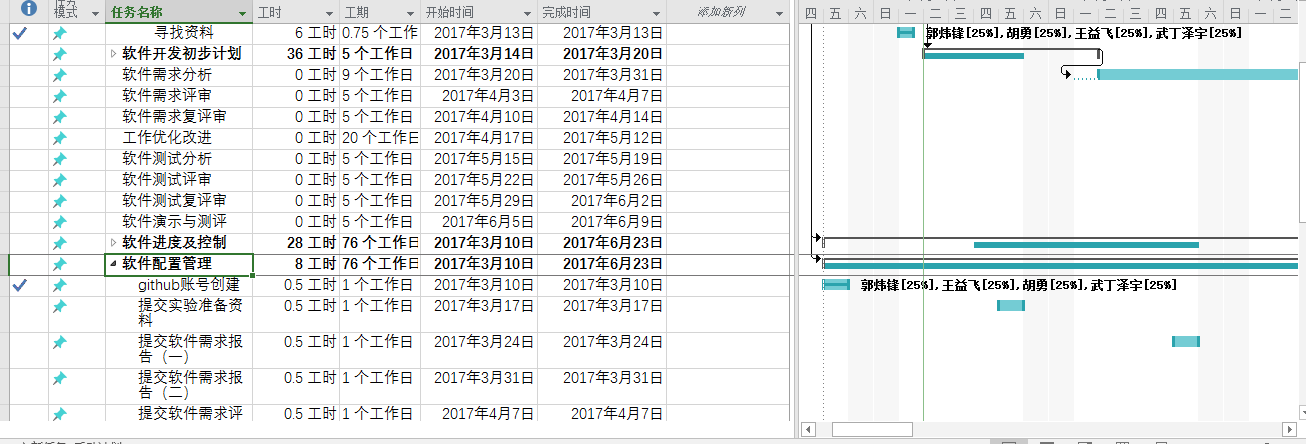
由于暂时还没有讨论出该项目的具体需求是什么，我们针对该项目要做什么具体操作，所以人员分工还需要进一步讨论中；但是其中的一个大目标已经确定了，就是针对某个网站来写爬虫代码，进而分析该项目的性能。

## 组织形式

小组会议跟分工均在微信上进行，并且随时在微信上进行问题讨论，资料分享以及各自进度汇报

## 进度安排

以下是小组进度的部分截图，进度也是第一版，由于还不了解之后的分配情况，具体分配内容跟计划还没有确定，在之后的版本中会持续更新。部分进度已经完成，也进行了追踪。



# 参考文献

Github源码参考：<https://github.com/scrapy/scrapy.git>

Scrapy官网中的pdf学习文档：<https://media.readthedocs.org/pdf/scrapy/1.3/scrapy.pdf>

《用Python写网络爬虫》 出版社: 人民邮电出版社; 第1版 (2016年3月1日)