Web搜索大作业：“中国裁判文书网”搜索引擎搭建

环境：ubuntu 14.04

python2.7(系统自带)

scrapy框架

mongodb

Elasticsearch 1.7.2

pyes

Django 1.9

Kibana3

本项目很大程度上借鉴了YongMan的Github项目，https://github.com/YongMan/SearchEngine

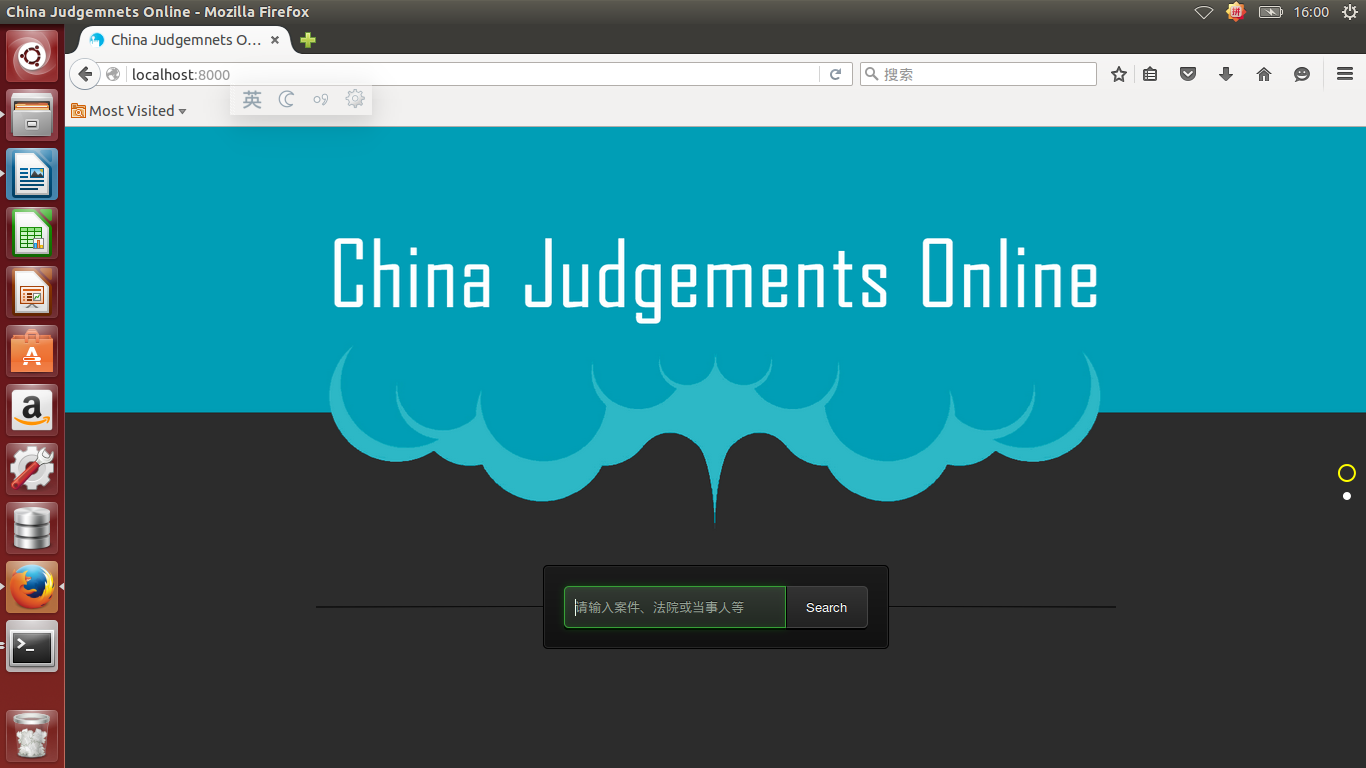
简介：基于Scrapy框架实现爬虫爬取“中国裁判文书网”内容，字段如下（课上规定拼音缩写.....）

|  |  |
| --- | --- |
| 字段 | 含义 |
| link | 详细网页url地址 |
| nr | 文书内容 |
| fy | 法院名称 |
| gbrq | 文书发布日期 |
| ID | 文书编号 |
| dq | 文书所在省份地区 |
| mc | 文书名称 |

爬取内容存储于Mongodb数据库，并在ElasticSearch中建立索引，最后基于Django进行Web

开发， 实现效果如下：

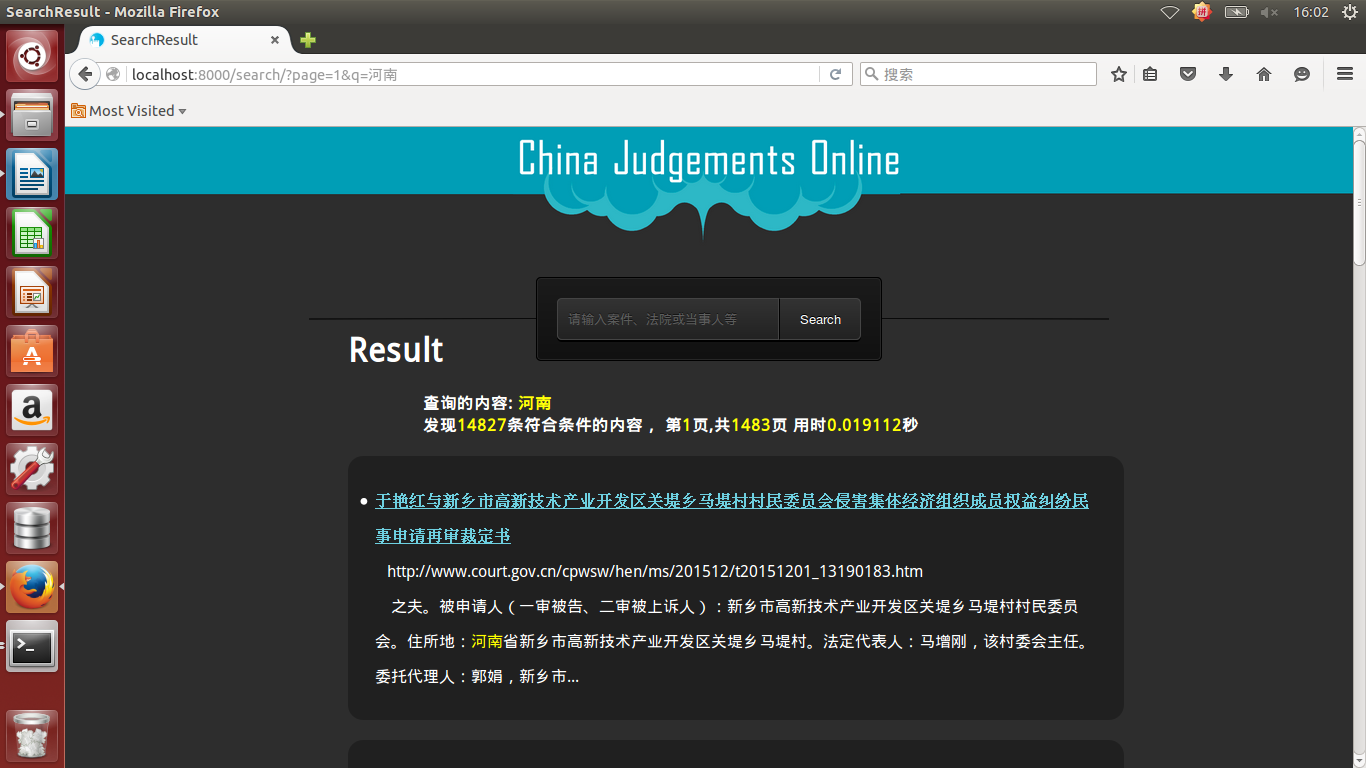
(1) 首界面——搜索界面



（2）Kibana界面调用



（3）结果显示界面



（4）点击链接进入“中国法律文书网”详情页



实现步骤：

一.数据爬取

（1）根据此网页介绍安装Scrapy框架，http://cuiqingcai.com/912.html

（2）下载安装Mongodb，<http://www.yiibai.com/mongodb/mongodb_environment.html>，

如果是新手，可以下载Robomongo进行图形化界面操作

（3）根据scrapy框架中pipelines.py文件所写，在Mongodb中建立LAW数据库和WS集合

（4）为防止重复抓取，项目中使用pybloomfiltermmap,请自行安装

（5）SearchIndex.py文件中实现连接ES，建立索引，设置mapping及文档插入功能

（6）tmp文件夹下运行 scrapy crawl myspider，即可看到程序运行效果（但由于项目会在数据爬取

时建立索引，所以需要开启ElasticSearch服务）

二.配置ElasticSearch

ElasticSearch基础知识及原理介绍，参见链接 <http://es.xiaoleilu.com/>

Elasticsearch安装参见链接， https://segmentfault.com/a/1190000000703282

（1）启动ES服务

在ElasticSearch官网下载所需版本安装包，解压后进入目录，使用./bin/elasticsearch命令开启服

务,在浏览器中访问9200端口，若返回200状态码则表示服务正常启动。

（2）IK分词插件

根据此GitHub项目，<https://github.com/medcl/elasticsearch-analysis-ik>，安装IK分词。直接

下载的项目代码中已经完成安装，可以跳过此步骤。

（3）ES集群配置

通过配置ElasticSearch的config文件夹下的elasticsearch.yml文件可以形成分布式集群，配置

时，协同工作的各主机需要有相同的cluster.name，及不同的node.name。

可以通过安装head插件，来检测集群工作状态，安装方式

<http://blog.csdn.net/july_2/article/details/24481935>

启动ElasticSearch服务后并访问 http://localhost:9200/\_plugin/head/ 即可看到博文中效果

三.搜索引擎实现

（1）安装pyes——使用python操作ES的第三方库

注意版本问题的函数变更,参考官方文档 http://pythonhosted.org/pyes/

search.py文件实现了对ES的连接及关键字查找功能

（2）安装Django框架

推荐慕课网Django基础教程，<http://www.imooc.com/learn/458>

（3）设计网站效果时，html文件中引用的js，css及image文件路径需要特殊处理

参考博文 <http://www.douban.com/note/504517892/>

项目中

本项目中则全部放在law应用文件夹下的static下的对应文件夹中，html文件的路径则写为如下格

式，/static/img/title.jpg（引用title.jpg文件）

（4）对于Kibana的介绍详见 http://kibana.logstash.es/content/kibana/index.html

项目中首个网页效果是通过对Kibana分享的url引用实现的。