**网络实习Lab报告**

组长： 刘子渊

组员：熊思亚东

张泽华

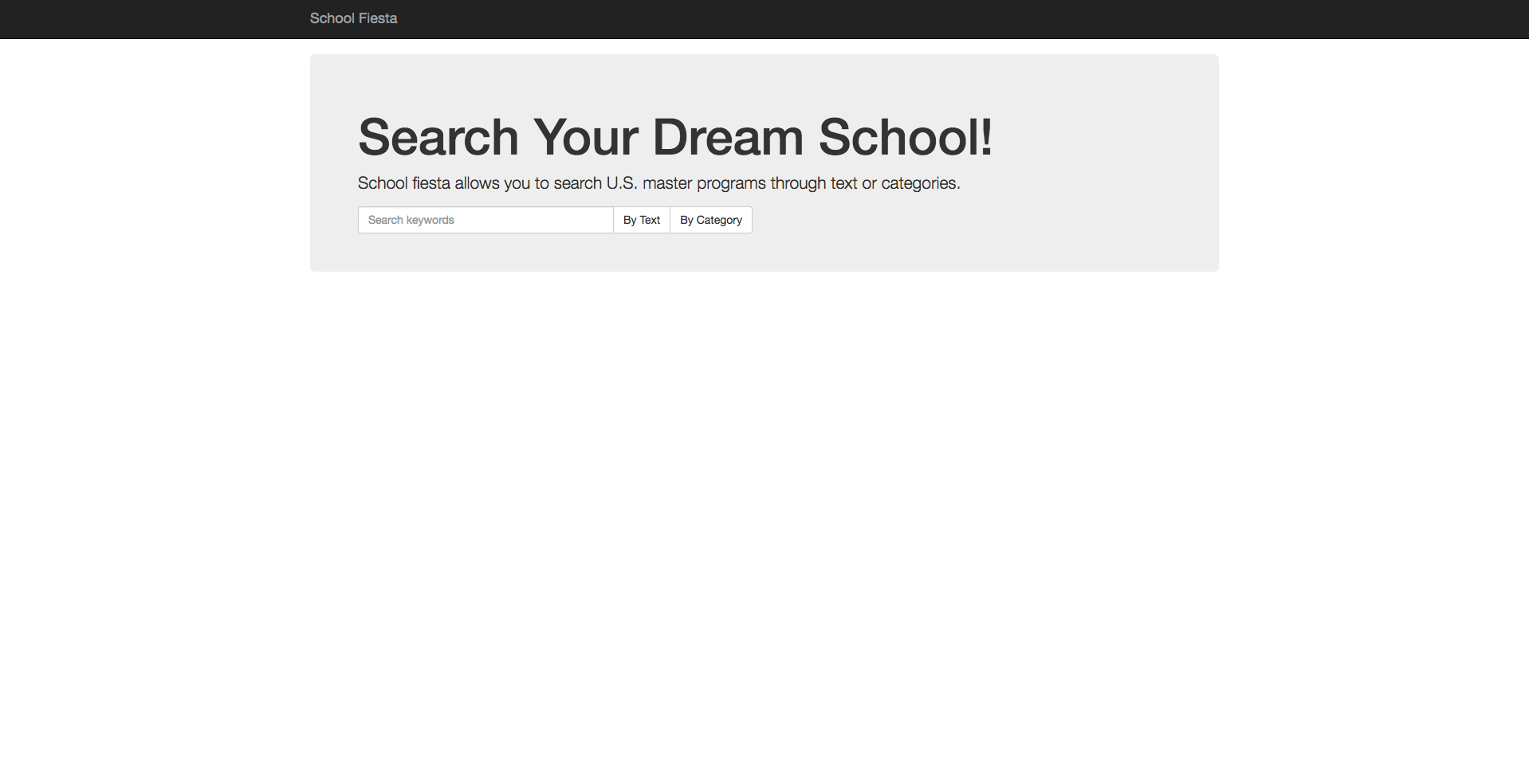
赵睿哲

1. 操作与功能

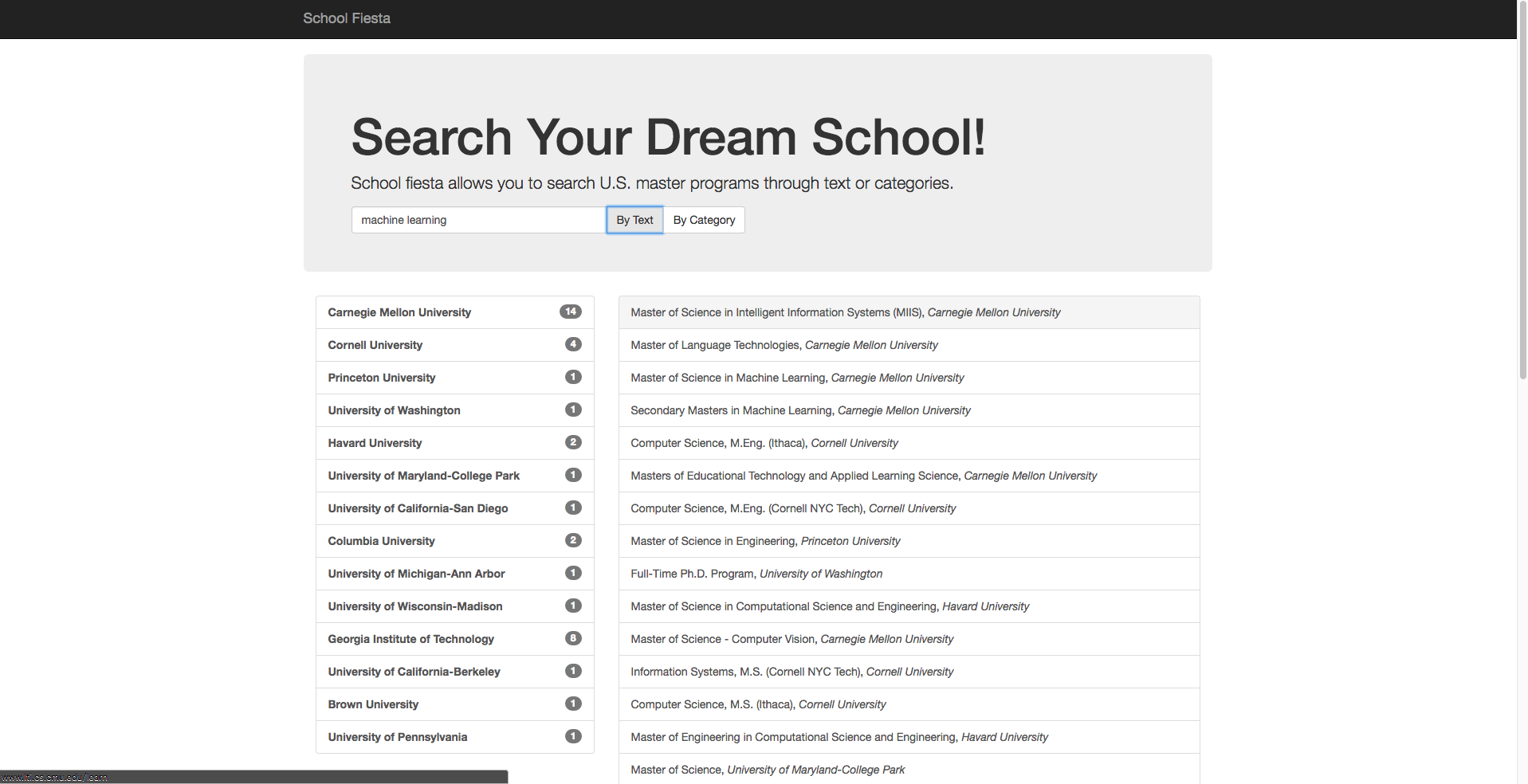
本组项目主要实现关于CS方向研究生留学项目的搜索。在NodeJS，Redis和Bootstrap框架的基础上实现了一个利用爬虫技术获取数据的留学项目搜索引擎系统。本项目中所使用而所有数据全部依靠于Python下的beautifulsoup4开源爬虫组件实现，爬取了美国排名较高院校的研究生留学项目共计59个。

本搜索引擎系统实现了两种搜索方式，一为朴素的传统搜索引擎文本匹配搜索方法。而是基于机器学习及文本分类技术的新型搜索引擎匹配方式。

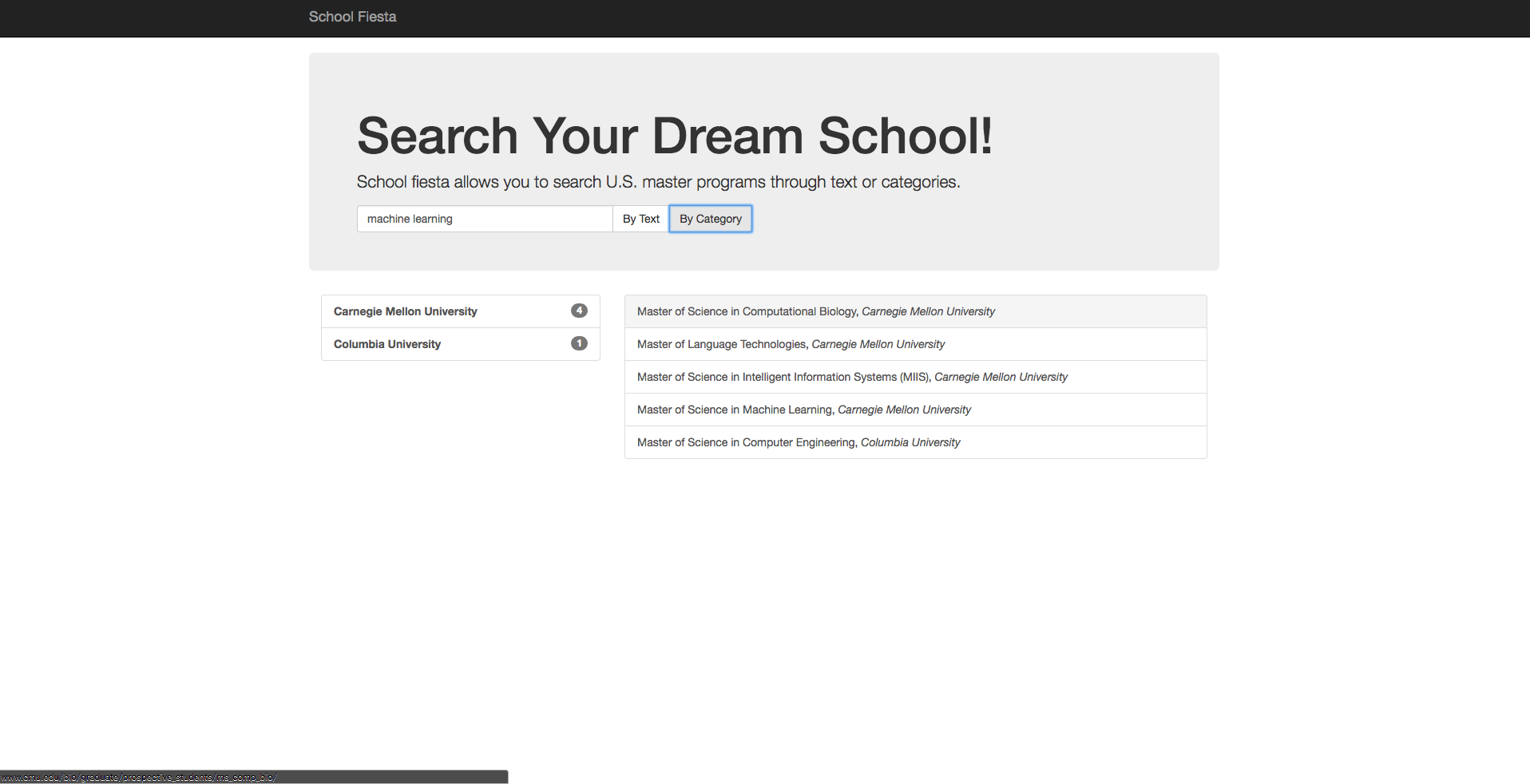
主要使用流程如下，下图所示图片全部源自项目运行时截图。



在进入系统尚未进行任何操作时网站界面如上图所示，主要组件为用于键入搜索关键词的文本输入框和选择搜索类型的两个按钮，分别代表本组项目中实现的两种搜索方式。



再输入搜索关键词之后，在本页面内会刷新出搜索得到的结果并根据院校进行了分类。可以看到左边一列展示了返回结果中所属学校的列表并表明了每个院校所包含的项目数量，便于用户进行筛选和检索。右方列表当中展示的内容则是项目的名称信息，点击后会直接跳转至对应项目主页。



点击”By Category”按钮后会使用基于机器学习的匹配技术进行检索，返回结果更加精确，展示形式上和上文所述并无不同。

1. 数据集

本项目中所使用的数据集共包括来及不同学校的共59个研究生留学项目。获取信息使用的爬虫模块使用Python下beautifulsoup4开源爬虫组件实现。由于各个学校的网页路由规则并不相同，因此无法使用传统的统一规则爬虫进行实现。于是在本次项目的实现过程当中人为加入了网页列表进行启发式搜索，从预先设定的主页列表中选取起点，进行递归的链接检索。最终获取到的数据内容以Json形式进行存储以便后续使用。

数据格式如下：

{

schools:[

{

progs:[

{

text:”xxx”,

[url:”xxx](url:)”,

dept:”xxx”,

name:”xxx”

},…

]

},…

]

}

1. 能否动态部署
2. 修改意见反馈

老师对于数据集的要求将根据实际需求很具学校排名列表继续扩大搜索范围进行对数据集的补充。此外还可以考虑将其他小组的不同类型数据进行一定程度上的整合最终获得较好的实现效果。

1. 组员分工：

刘子渊：机器学习算法实现

熊思亚东：爬虫部分实现

张泽华：项目前端实现

赵睿哲：项目架构设计与实现

1. 亮点

爬虫部分：由于各个学校的网页路由规则并不相同，因此无法使用传统的统一规则爬虫进行实现。于是在本次项目的实现过程当中人为加入了网页列表进行启发式搜索，从预先设定的主页列表中选取起点，进行递归的链接检索。

搜索引擎部分：搜索引擎实现部分传统匹配方式运用了大量的自然语言处理方法对文本内容进行预处理。而运用机器学习算法的文档评分模块使用Softmax和基于特征向量提取算法实现了对目标文本的分类和内容分析，具有一定的技术和实现难度，代码量不少。

服务端、前端实现：整个项目架构使用NodeJS的Express配合angular.js框架实现，