WEB编程期末项目 初步设想

10211900416 郭夏辉

主要内容：功能设想（最好有功能结构示意图）

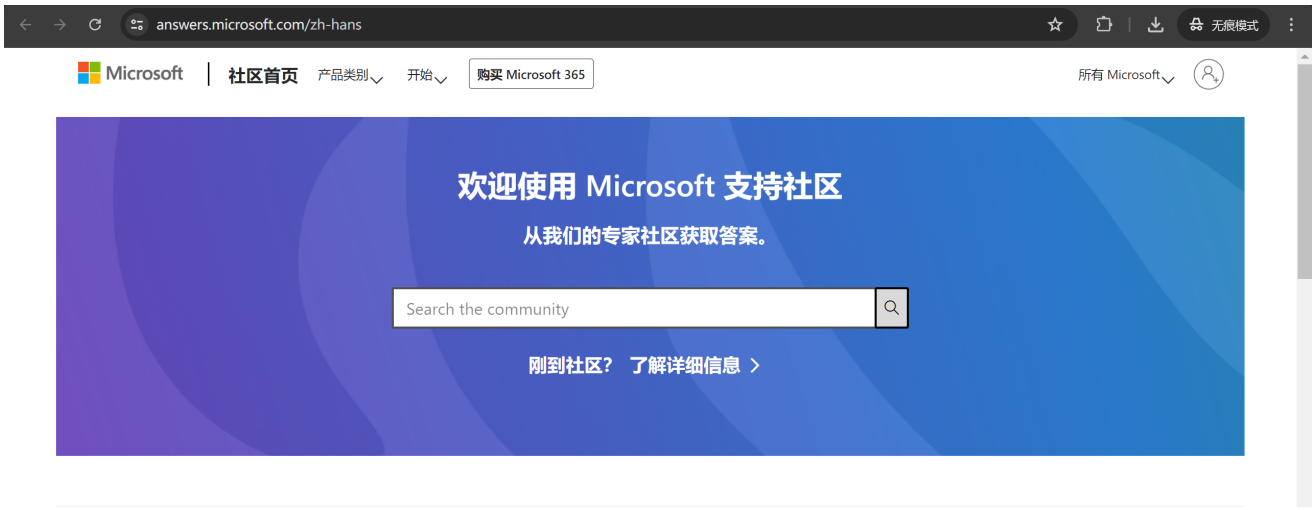
技术路线初步计划（可以包含文字，构思图，伪代码）

功能设想

老师的要求是自定主题涉及并完成一个Web网页，需要基于传统的HTML+JS+CSS技术，还要有图、文字、链接等结构，并且有显示数据的功能，至少有3个页面。

为了让项目更加有意义，我觉得数据不能来源于已有的数据集，而是网络上动态产生的数据流，这就需要利用到爬虫技术了。每天生活中都有很多新闻涌来，但是我们的精力有限，所以只能挑选感兴趣的几个领域、一些信息来阅览。我的想法是利用爬虫技术获取这样的信息中心网站的资源，将这些资源存入数据库中，然后利用一些自然语言处理技术（比如倒排索引、分词、跳表）使检索过程更加高效，提升我们生活的质量。

至于功能设想，它的样貌其实挺像这样：



简而言之就是有一个搜索栏用来确定想要检索的内容；还有一个边栏进行页面切换，这里可以有一些数据分析类的产出，比如说词云、情感分析、热点聚焦。

如果实现顺利，我的后端甚至可以引入一些AI相关的模型，并且假如用户管理这样的功能，对于每个用户个性化地在各个边栏界面中推荐与展示不同的信息。

技术路线初步计划

前端：HTML+CSS+JS 我甚至还想使用React,Vue这样的框架加速开发、提升效果

后端：应该是基于Node.js或者Express这样的框架来实现相应的功能。

数据库：大概基于SQLite或MySQL部署

构思图（其实没啥，麻雀虽小五脏俱全）：

client -> server -> crawler -> Database -> FrontEnd(for show) -> Interaction (for users)

伪代码(爬虫)：

```
var myRequest = require('request')
var myCheerio = require('cheerio')
var myURL = 'https://www.ecnu.edu.cn/info/1094/63197.htm'
function request(url, callback) {//request module fetching url
    var options = {
        url: url,  encoding: null, headers: null
    }
    myRequest(options, callback)
}
request(myURL, function (err, res, body) {
    var html = body;
    var $ = myCheerio.load(html, { decodeEntities: false });
    //console.log($.html());
    console.log("title: " + $('title').text());
    console.log("description: "
              + $('META[Name="description"]').eq(1).attr("content"));

})

var schedule = require('node-schedule')  //设置定时爬虫
var myIconv = require('iconv-lite')  //编码转换 GB2312到UTF-8
var fs = require('fs');  //保存到本地文件
var mysql = require('mysql');  //保存到mysql数据库
npm install elasticsearch  //可以用elasticsearch构建爬取数据的索引
```