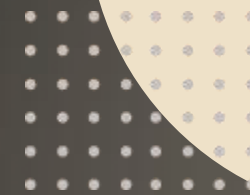


新品首发

新闻搜索——语宙

News Search——Yuzhou



01

目标网站分析

02

爬虫设计介绍

03

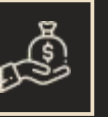
最终网站的设计介绍

04

效果展示

目录

CONTENTS



PART ONE

目标网站分析 ——网易、新浪

Target Website Analytics

01


```
var url_reg = /\//news\//article\//(\w{16}).html/;
```

https://www.163.com/news/article/I9RMI4JD000189FH.html?clickfrom=w_yw

https://www.163.com/news/article/I9RMI4JD000189FH.html?clickfrom=w_yw

在深化内化转化上下功夫——把理论学习

来源: 人民网-人民日报 北京

举报

<title>坚持不懈在深化内化转化上下功夫——把理论学习贯穿始终|中国特色社会主义|方法论|科学
_网易政务</title>

我们这么大一个党，领导着这么大一个国家，肩负着带领全国各族人民实现国家强盛、民族复兴这个艰巨任务，全党必须统一思想、统一意志、统一行动。怎么实现全党思想、意志、行动的统一？最根本的就是用党的创新理论武装全党。

拥有马克思主义科学理论指导是我们党坚定信仰信念、把握历史主动的根本所在。每逢重大历史关头，我们党都要用党的创新理论统一全党思想，每次党内集中教育也都

```
html id="ne_wrap" data-publishTime="2023-07-17 12:36:39" data-category="要闻"
class="ua-win">
<head>
<style class="ariareader ariareader--user-agent" media="screen" id="ariaf4pwhc8zp">
<!-- head -->
<title>坚持不懈在深化内化转化上下功夫——把理论学习贯穿始终|中国特色社会主义|方法论|科学
<meta name="keywords" content="马克思主义,中国特色社会主义,方法论,科学,习近平">
<meta name="description" content="坚持不懈在深化内化转化上下功夫——把理论学习贯穿始终,
<meta name="author" content="网易">
<meta name="Copyright" content="网易版权所有">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
```

```
<meta http-equiv="Cache-Control" content="no-transform">
<meta http-equiv="Cache-Control" content="no-cache">
```

html#ne_wrap.ua-win head title

样式 已计算 布局 事件侦听器 DOM 断点 属性 辅助功能

筛选器

:hov .cls +

element.style {

控制台 问题 × +

☐ 包括第三方问题 严重性: 默认级别 浏览器: 热门浏览器


☐ 按种类分组

```
var source_name = "网易";
var myEncoding = "utf-8";
var seedURL = 'https://www.163.com/';

var seedURL_format = "$('a')";
var keywords_format = "$('meta[name=\"keywords\"]').eq(0).attr(\"content\")";
var title_format = "$('title').text()";
var date_format = "$('meta[property=\"article:published_time\"]').eq(0).attr(\"content\")";
var author_format = "$('.post_author').text()";
var content_format = "$('.post_body').text()";
var desc_format = "$('meta[name=\"description\"]').eq(0).attr(\"content\")";
var source_format = "$('.post_author').text()"; // 这里和责编同一个
var url_reg = /\news\/article\/(\w{16}).html/;
```


新浪新闻



 <https://news.sina.com.cn/c/2023-07-17/doc-imzayser5094431.shtml>

URL正则表达式

```
var url_reg = /\d{4}-\d{2}-\d{2}\sdoc-(\w{15}).shtml/;
```

h1.main-title

Color
Font 38px "Hiragino" ...
Padding

h1.main-title

ACCESSIBILITY

Name 8家网约车平台被勒令下架，整顿风暴来...
Role heading
Keyboard-focusable

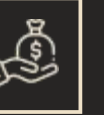


```
<!-- main content start -->
<div class="main-content w1240">
  <!-- 顶部通栏广告 start -->
  <div class="top-banner clearfix">...</div>
  <!-- 顶部通栏广告 end -->
  <!-- 面包屑 search start -->
  <!-- cID=51922, colID=51922, subCID=51922, thirdCID=third_cid, final=51922
  BoYan -->
  <div class="path-search" data-sudaclick="cnav_breadcrumbs_p">...</div>
  <!-- 面包屑 search end -->
  <h1 class="main-title">8家网约车平台被勒令下架，整顿风暴来了? </h1> == $0
  <!-- page-tools start -->
  <style type="text/css">...</style>
  <div class="ton-bar-wran" id="ton_bar_wran">...</div>
```

```
<h1 class="main-title">8家网约车平台被勒令下架，整顿风暴来了? </h1>
```

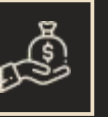
2023年07月17日 16:20 中国新闻周刊

html body..sinacMNT_logout div.main-content.w1240 h1.main-title



```
var source_name = "新浪新闻";
var myEncoding = "utf-8";
var seedURL = 'http://news.sina.com.cn';

var seedURL_format = "$('a')";
var keywords_format = "$('meta[name=\"keywords\"]').eq(0).attr(\"content\")";
var title_format = "$('.main-title').text()";
var date_format = "$('.date').text()";
var author_format = "$('.show_author').text()";
var content_format = "$('.article').text()";
var desc_format = "$('meta[name=\"description\"]').eq(0).attr(\"content\")";
var source_format = "$('meta[name=\"mediaid\"]').eq(0).attr(\"content\")";
var url_reg = /\\/(\d{4})-(\d{2})-(\d{2})\/doc-(\w{15}).shtml/;
var regExp = /((\d{4}|\d{2})(\-|\/|\.)\d{1,2}\3\d{1,2})|(\d{4}年\d{1,2}月\d{1,2}日)/
```

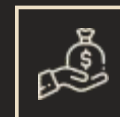


PART TWO

爬虫设计介绍 ——以网易为例

Introduction to Crawler Design

02



```
var fs = require('fs');  
var myRequest = require('request');  
var myCheerio = require('cheerio');  
var myIconv = require('iconv-lite');  
var schedule = require('node-schedule');  
var mysql = require('./mysql.js');  
require('date-utils');
```

json文件

获取网页内容

筛选和提取网页中的信息

解析编码

定时

数据库查询和写入

```
new Date(fetch.publish_date).toFormat("YYYY-MM-DD");
```



```
//防止网站屏蔽我们的爬虫
var headers = {
  'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_1)
}
```

```
//request模块异步fetch url
function request(url, callback) {
  var options = {
    url: url,
    encoding: null,
    headers: headers,
    timeout: 10000 //
  };
  myRequest(options, callback);
}
```

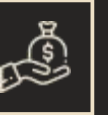
```
var rule = new schedule.RecurrenceRule();
var times = [10, 12, 14, 16, 18, 20, 22]; //每天7次自动执行
var times2 = 1; // 定义在第几分钟执行
rule.hour = times;
rule.minute = times2;

//定时执行seedget()函数
schedule.scheduleJob(rule, function() {
  seedget();
});
```

模拟浏览器

使用回调函数来异步地爬取网页的url，爬取url的同时就能对所得url进行解析

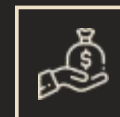
调用函数
schedule.RecurrenceRule(),
定时自动执行seedget()爬取



```
function seedget() {
  request(seedURL, function(err, res, body) { //读取种子页面
    //用iconv转换编码
    var fetch_url_Sql = 'select url from fetches where url=?';
    var fetch_url_Sql_Params = [myURL];
    mysql.query(fetch_url_Sql, fetch_url_Sql_Params, function(qerr, vals, fields) {
      if (vals.length > 0) {
        console.log('URL duplicate!')
      } else newsGet(myURL); //读取新闻页面
    });
    var href = ""; //得到具体新闻url
    href = $(e).attr("href");
    if (typeof href == "undefined") {
      return;
    }
    if (myURL == "https://jubao.163.com/") return;
    if (href.toLowerCase().indexOf('https://') >= 0) myURL = href;
    else if (href.startsWith('//')) myURL = 'http:' + href; //开头的
    else myURL = seedURL.substr(0, seedURL.lastIndexOf('/') + 1) + href; //其他
  } catch (e) { console.log('识别种子页面中的新闻链接出错: ' + e); }};

  if (!url_reg.test(myURL)) return;
```

检查读取的 url 是否在数据库中已经存储过了，避免重复插入数据

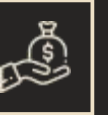


```
function newsGet(myURL) { //读取新闻页面
    request(myURL, function(err, res, body) {
        var html_news = myIconv.decode(body, myEncoding); //用iconv转换编码
        var $ = myCheerio.load(html_news, { decodeEntities: true });
        myhtml = html_news;

        console.log("转码读取成功:" + myURL);

        //动态执行format字符串, 构建json对象准备写入文件或数据库
        var fetch = {};
        fetch.title = "";
        fetch.content = "";
        fetch.publish_date = "";
        fetch.url = myURL;
        fetch.source_name = source_name;
        fetch.source_encoding = myEncoding; //编码
        fetch.crawltime = new Date();
    });
}
```

爬取的信息包括：标题、内容、发表时间、URL、来源、编码和读取时间



```
var date_format = "$('meta[property=\"article:published_time\"]').eq(0).attr(\"content\")";
```

div.pub time 137.94 × 24

Color #AAAAAA

Font 14px "Microsoft YaHei"

Margin 0px 42px 0px 0px

ACCESSIBILITY

Contrast Aa 2.32

Name

Role gener

Keyboard-focusable

2023-07-10 12:46:23

谈心社

djs_normal-c9f37ccad0.js"></script>

▼ <div class="zajia_content">

<!-- 项目内容 -->

▶ <div class="banner">...</div>

▼ <div class="cc_bd w1000">

<!-- 标题 作者 时间 跟贴数 部分 -->

▼ <div class="brief">

10 12:46:23</div> == \$0
/div>

try{

temp = fetch.publish_date;

t1 = temp.split("T")[0];

t2 = temp.split("T")[1].split("+")[0];

fetch.publish_date = t1 + " " + t2;

fetch.publish_date = new Date(fetch.publish_date).toFormat("YYYY-MM-DD");

}

catch (e)

{

var pubtime_format = "\$('.pub_time').text()";

fetch.publish_date = eval(pubtime_format);

fetch.publish_date = new Date(fetch.publish_date).toFormat("YYYY-MM-DD");

}

console.log('date: ' + fetch.publish_date);

div.brief div.sub_title.clea



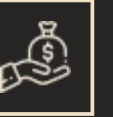
id_fetches	url	source_name	source_encoding	title	keywords	author	publish_date	crawltime	content	createtime
1382	https://news.sina.com.cn	新浪新闻	utf-8	耶伦访华将	耶伦访华将与	环球网 责任编辑: 王	2023-07-07	2023-07-10 10:00	[环球时	2023-07-10 0
1383	https://news.sina.com.cn	新浪新闻	utf-8	官方: 蚂蚁	蚂蚁集团回应	证券时报网 责任编辑: 王	2023-07-07	2023-07-10 10:00	据证监	2023-07-10 0
1384	https://www.163.com	网易	utf-8	泰国一载有	中国游客,翻车	央视新闻客户端 孙	2023-07-10	2023-07-10 10:00		2023-07-10 1
1386	https://www.163.com	网易	utf-8	习近平向全	习近平,联合国	新华社客户端 张叶	2023-07-10	2023-07-10 10:00		2023-07-10 1
1388	https://www.163.com	网易	utf-8	第1视点	习近平,总书记	新华社客户端 张叶	2023-07-10	2023-07-10 10:00		2023-07-10 1
1389	https://www.163.com	网易	utf-8	奋力谱写中	浙江,习近平,宁	新华社客户端 张叶	2023-07-10	2023-07-10 10:00		2023-07-10 1
1390	https://www.163.com	网易	utf-8	在沪上举行	上海,习近平,宁	新华社客户端 张叶	2023-07-10	2023-07-10 10:00		2023-07-10 1
1393	https://www.163.com	网易	utf-8	在沪上举行	上海,习近平,宁	新华社客户端 张叶	2023-07-10	2023-07-10 10:00		2023-07-10 1
1395	https://www.163.com	网易	utf-8	在沪上举行	上海,习近平,宁	新华社客户端 张叶	2023-07-10	2023-07-10 10:00		2023-07-10 1
1398	https://news.sina.com.cn	新浪新闻	utf-8	6月份居民消	6月份居民消费	新京报 责任编辑: 王	2023-07-10	2023-07-10 10:00	新京报	2023-07-10 1
1399	https://news.sina.com.cn	新浪新闻	utf-8	习近平向全	习近平向全球	新华网 责任编辑: 王	2023-07-10	2023-07-10 10:00	新华社	2023-07-10 1
1400	https://news.sina.com.cn	新浪新闻	utf-8	11部门联合	11部门联合部	新华每日电讯 责任	2023-07-10	2023-07-10 10:00	新华社	2023-07-10 1
1401	https://news.sina.com.cn	新浪新闻	utf-8	美国芝加哥	美国芝加哥再	海外网 责任编辑: 王	2023-07-10	2023-07-10 10:00	美国芝加哥	2023-07-10 1
1402	https://news.sina.com.cn	新浪新闻	utf-8	耶伦的“加	美财政部部长	第一财经 责任编辑: 王	2023-07-10	2023-07-10 10:00	“美国和	2023-07-10 1

7.09 — 7.19 共 1640 条新闻

(原标题: 泰国一载有中国游客的巴士发生翻车事故 27人受伤)

事发现场当地时间7月9日下午,一辆载有20多名中国游客的旅游巴士在从罗勇府沙美岛前往芭堤雅的途中发生翻车事故。救援人员切割车身将被困游客救出事故发生时正在下雨,路面湿滑,事故导致车上包括司机在内27人受伤,其中有4人被困在车内,救援人员赶到后切割车身将他们救出并送往医院。





PART THREE

最终网站 的设计介绍

Introduction to the Design of the Final Website

03

Express脚手架

后端采用了Express脚手架，通过访问数据库获取数据，基本在search.html中实现，样式大多在mycss.css中

```
▼ search_site
  > bin
  > node_modules
  ▼ public
    ▼ images
      🖼 background.jpg
      🖼 cloud.png
      🖼 logo.png
      🖼 p.png
    > javascripts
  ▼ stylesheets
    # mycss.css
    # style.css
    # bootstrap-table-pagejump.css
    JS bootstrap-table-pagejump.js
    <> search.html
    <> wordcloud.html
  > routes
  > views
  JS app.js
  JS mysql.js
  {} package-lock.json
  {} package.json
```


Logo和搜索框

```
<form>
  <div class="search-box" id="box" style="display: block;">
    <input class="search-txt" type="text" name="title_text" placeholder="标题">
    <input class="search-txt" type="text" name="keywords_text" placeholder="关键词">
    <input class="search-txt" type="text" name="author_text" placeholder="作者">
      <button type="button" class="serach-btn" id="btn1">
        <i class="fa fa-search" aria-hidden="true"></i>
      </button>
    </div>
    <div class="logo" id="logo">
      
    </div>
  </form>
```

动态搜索框可同时搜索三个方面，并添加 logo

头部导航栏和两个按钮

```
<div class="header">
  <div style="width: 90%;">
    <marquee direction="left" behavior="slide" scrollamount="25">
      <h1 style="color: #F1EAE0;">家事国事天下事 事事关心</h1>
    </marquee>
  </div>
  <div class="blink" id="twinkle">
    <h1 style="color: #e84118;">你好</h1>
  </div>

  <div class="buttons">
    <button type="button" id="btn3" class="serach-btn2" style="display: none;">
      <i class="fa fa-mail-reply-all" aria-hidden="true"></i>
    </button>
    <button type="button" id="btn2" class="serach-btn2" style="display: none;">
      <i class="fa fa-line-chart" aria-hidden="true"></i>
    </button>
  </div>
</div>
```

→ 左边滑进

→ 文字闪烁

↑ 返回搜索页

↓ 图表与表格切换

四个图表及其切换按钮

```
<div id="main1" style="width: 600px; height:400px; display:none;"></div> <!--柱状图-->
<div id="main2" style="width: 600px; height:400px; display:none;"></div> <!--饼图-->
<div id="main3" style="width: 600px; height:400px; display:none;"></div> <!--面积图-->
<div id="main4" style="float: right; width: 400px; height:400px; display:none;"></div> <!--热词图-->

<div class="btns">
  <button class="button" type="button" id="btn4" style="display: none;">
    柱状图
  </button>
  <button class="button" type="button" id="btn5" style="display: none;">
    饼图
  </button>
  <button class="button" type="button" id="btn6" style="display: none;">
    面积图
  </button>
  <button class="button2" type="button" id="btn7" style="display: block;">
    近日热词
  </button>
</div>
```

使用了热词榜、柱状图、饼图和面积图来可视化结果

两个表格

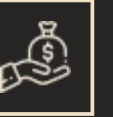
```
<div class="table table-hover" style="table-layout: fixed; word-break: break-all;" id="table1">
|   <table width="100%" id="record2" border="1" align="center" style="background-color: #F1EAE0;"></table>
</div>

<div class="table2" style="table-layout: fixed; width: 50%; height: auto;" id="table2">
|   <table width="100%" id="record1" border="0px" align="center" style="background-color: #F1EAE0;"></table>
</div>
```

table1是搜索结果，用 bootstrap-table美化、分页

table2是按日期统计搜索数量的展示，增添鼠标悬浮效果

css样式及script部分详见代码

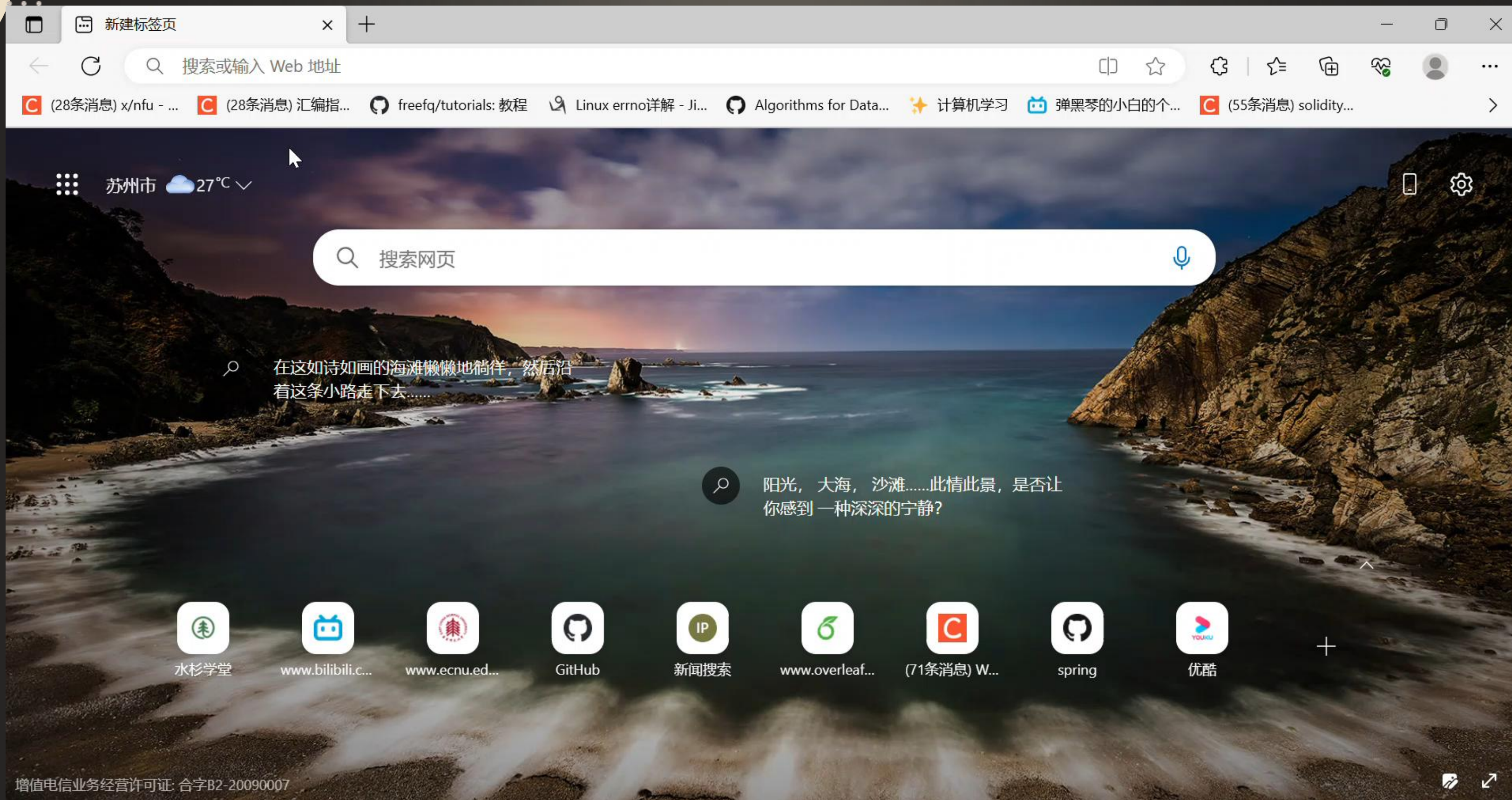


PART FOUR

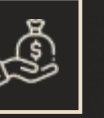
效果展示

Effect Display

04



亮点与展望



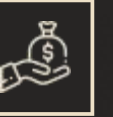
动态界面和搜索框

复合和二次搜索

多种图表

美化词云

增加或搜索



谢谢观看

Thanks for Watching
