软件设计文档

文档管理信息表

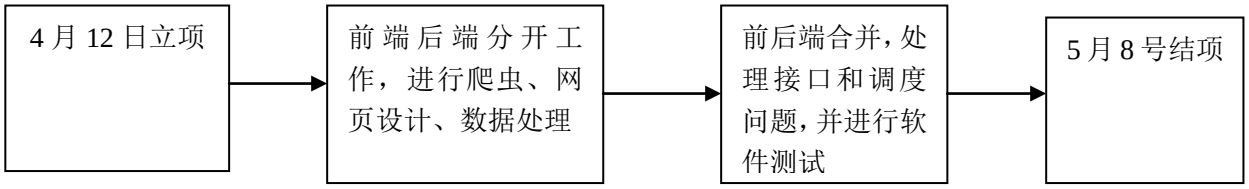
主题	汽车销售数据平台
版本	1
关键字	汽车推荐、汽车精选、情感分析
创建时间	2018.4.12
创建人	陈博楠
最新发布日期	2018.5.13

1　开发规划

1.1　开发人员

角　色	主要职责	负责模块	人员	备注
项目经理 PM	■　项目全面负责 ■　项目设计 ■　项目进度控制 ■　定义需求 ■　产品监督 ■　用户文档	■　算法模块	刘昱昊	组长
程序员 DEV	■　前端框架编写 ■　前端测试 ■　项目设计 ■　定义需求	■　前端模块	罗雨辰 张家麟 田陇宁	
程序员 DEV	■　数据爬取 ■　项目设计 ■　定义需求	■　后端模块	方潇玥 陈博楠	
数据库管理员	■　数据处理 ■　软件测试 ■　用户文档	■　数据库模块	王珊珊	
程序员 DEV	■　文本数据挖掘和情感分析	■　算法模块	刘昱昊	

1.2　开发计划



注：每周四软工下课后讨论各部分进度。

1.3 开发环境和工具

1.3.1 软件及相关环境的支持

名称	作用
相关的软件支持	
名称	作用
HBuilder	进行 HTML 的网页的布局以及样式的设计
Dreamweaver cs6	辅助进行 HTML 页面的布局设计
pycharm	应用 python 的 python web 框架 django 开发网站；利用 python3 进行汽车属性、图片等的爬取、文本评论数据的情感分析及可视化
MySQLWorkbench	提供强大的可视化设计、模型建立、以及数据库管理功能
Navicat for MySQL	连接到 MySQL，为数据库管理、开发和维护提供直观的图形界面
其他技术支持	
Django 和 xadmin	应用 django 的 xadmin 后台管理系统，兼容 Django Admin，使用 Bootstrap 作为 UI 框架，编辑页面灵活布局。

1.3.2 环境及相关的包

- 前端开发：

Python 版本：python 2.7

虚拟环境：整个项目在新建的虚拟环境中进行。

在开发的过程中安装和调用的包有：

Django 1.9

MySQL-python 1.2.5

Pillow 5.1.0

PyYAML 3.12

- 后端开发：

  Python 版本：python 3.6

  IDE：spyder，pycharm

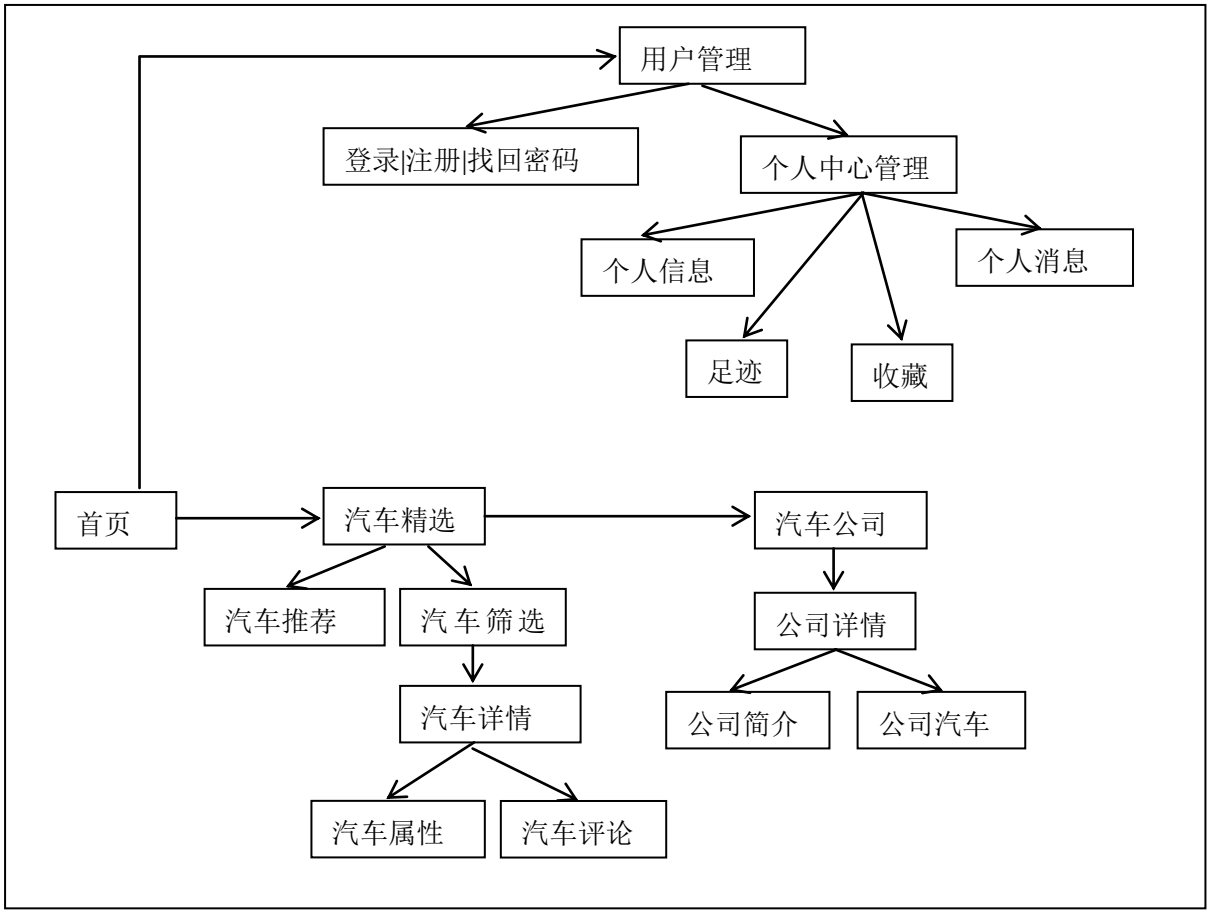
2 总体设计

2.1 基本设计描述

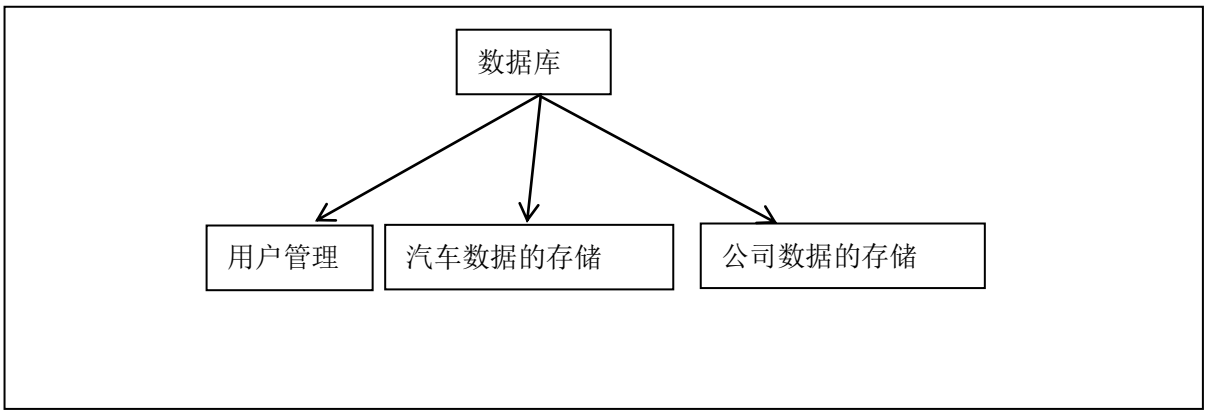
小组产品的名称是 HeyCar,其主要功能是为用户提供一个选择和了解车辆的简洁明了的平台。网站分为首页、汽车精选、汽车公司这三大模块。首页提供了用户管理包括注册和登录的个人中心的管理控制；汽车精选模块实现了依据汽车的诸多属性进行汽车车型是筛选的强大功能，并且提供深入解车型包括标准化的情感指数以及其他诸多属性的功能。同时，汽车筛选也具有相关车型推荐的功能。汽车公司模块则旨在为用户深入了解汽车公司提供技术支持。

2.1.1 系统总体逻辑结构图

网站总体结构图



数据库设计结构



2.2 主要模块及功能流程

2.2.0 模块列表综述

模块名称(英文)	功能	备注
模块 1 Module1	用户登录与密码找回	通过邮件与管理员联系
模块 2 Module2	信息向导	
模块 3 Module3	按照需求精确筛选车型以及引导用户找到符合自己实际情况的汽车，并且通过猜你喜欢模块来为用户推荐车型	
模块 4 Module4	为用户展示更加丰富的汽车公司信息	
模块 5 Module5	利用用户的浏览记录数据和文本评论数据为用户构建 TOP-K 推荐列表	
模块 6 Module6	为有公司偏好的购车人士提供公司详情及典型车系，供其选择	
模块 7 Module7	用户个人信息记录模块，包括个人信息、浏览记录和收藏列表	

2.2.1 模块 1——登录注册

图 1 登录与注册页面

- **功能详情：**
  - ✓ 注册：登录页面如上图所示，用户输入“用户名”和“密码”，点击下方登录按钮即可登录。如果没有账号可以点击下方注册，输入所需信息后，平台会向用户所填邮箱发送链接，点击链接就可以激活注册的账号。
  - ✓ 账号找回：忘记密码可以点击上图中忘记密码部分，填写必要信息，平台就会向邮箱发送验证码，输入验证码即可重置密码。

2.2.2 模块 2——首页信息

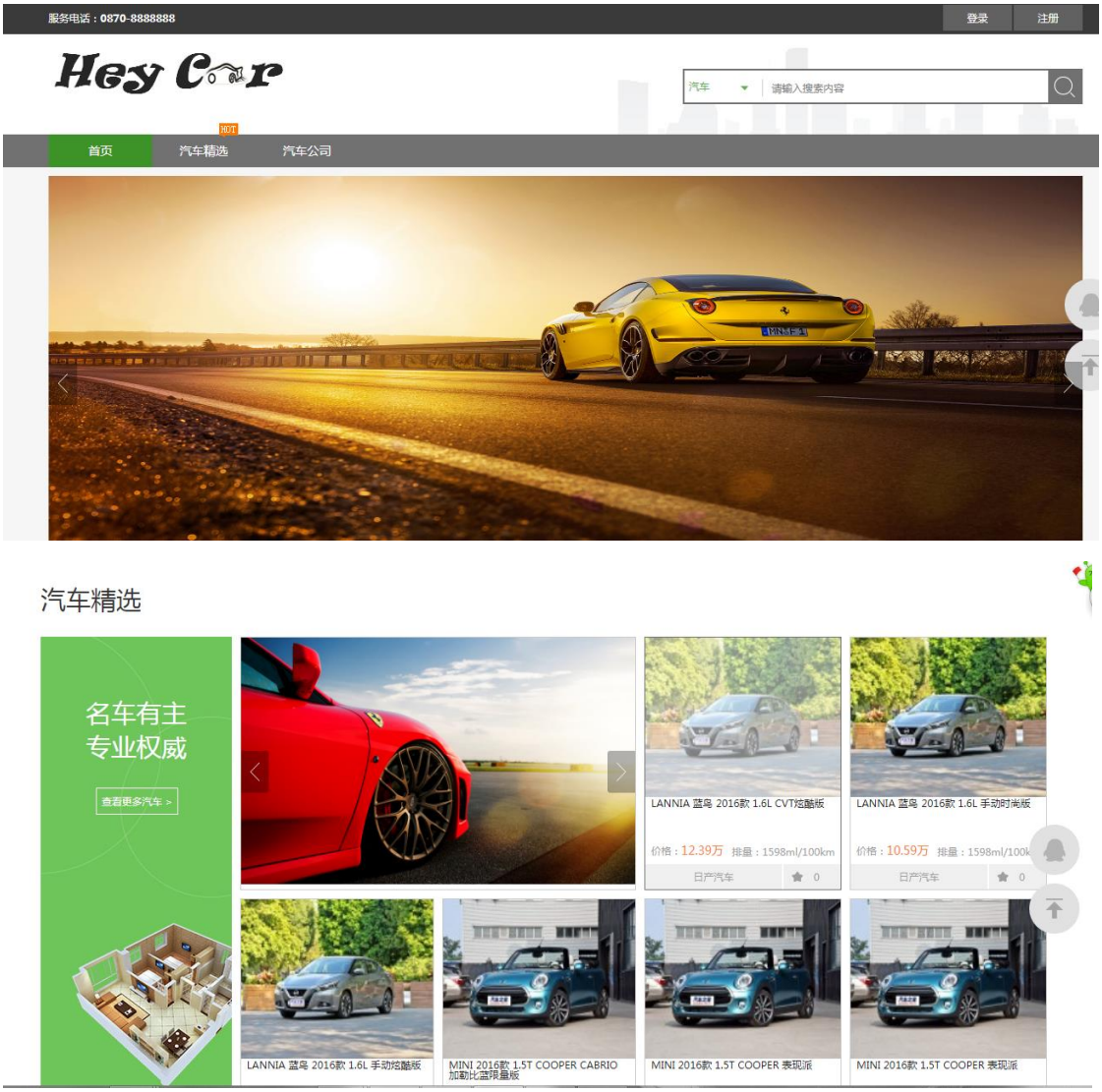


图 2 HeyCar 首页

- 功能详情：
  - ✓ 轮播图：一打开首页映入眼帘的是几张精美的轮播图，页面非常简洁美观。左上角是我们的标志和联系电话。右上角有登录注册按钮，点击按钮可以跳转到相应界面。
  - ✓ 模糊搜索：右上方的搜索框支持模糊搜索，可以先选择前边是选“汽车”还是“公司”，在“汽车”状态输入信息搜索后显示的是对应汽车，在“公司”状态操作后显示的会是对应公司。找不到的情况网页会向用户报错。
  - ✓ 信息向导：首页拖动到下方可以看到“汽车精选”和“汽车公司”的缩略部分，我们选择了销量最高的六款车以及点击率最高的六个公司放在这里。点击最左边“查看更多汽车”可以跳转到“汽车精选”页面，点击右边的汽车则可以进入这辆汽车的详情界面。

2.2.3 模块 3——汽车精选模块

图 3 汽车精选模块----汽车筛选页面

图 4：汽车精选模块----详情页面

- 功能详情：
  - ✓ 多条件筛选：如图 3 所示，筛选后会出现满足条件的车型，车型出现顺序是双重标准的，其中默认的第一标准是“销量”，第二标准如上图“最新”或者“最热门”，由客户自行选择。选择“最新”则销量相同的按更晚发布的排在前面来排序，选择“最热门”则销量相同的按点击次数多的排在前面来排序。
  - ✓ 爱车收藏：如图 4 所示，点击汽车可以进入这款汽车的详情界面，在详情界面可以收藏喜欢的汽车。同时在右侧也会列出“相关汽车推荐”，这是一个基于用户的推荐
    系统。

2.2.4 模块 4——汽车评论模块

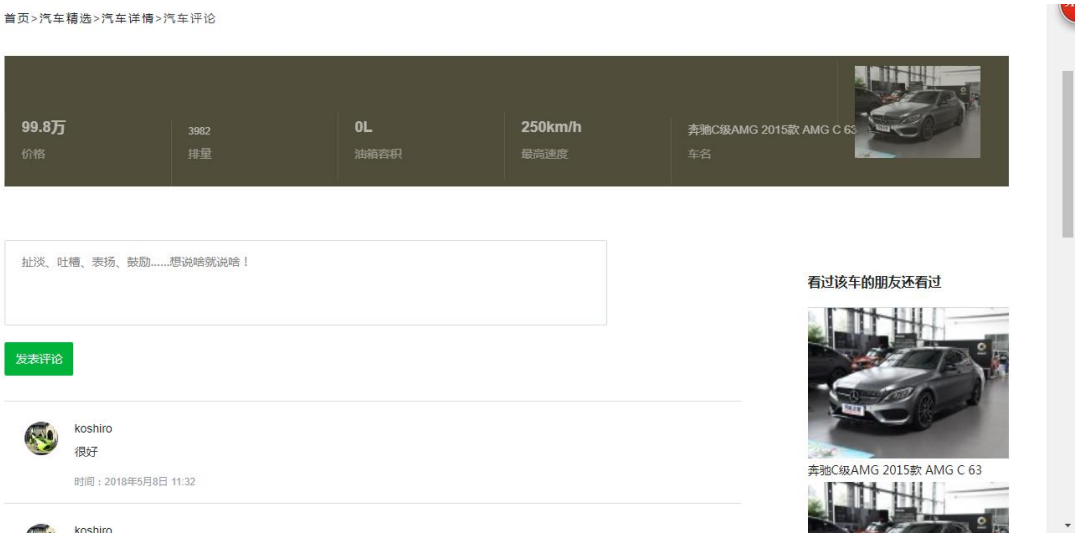


图 5：汽车评论模块

- 功能详情：
  - ✓ 查看评论：如图 4，5 所示，通过点击汽车详情页里面的“查看评论”按钮，页面跳转到评论模块。作为网站游客，可以看到所选车型的评论数据及相关汽车配置，若想对汽车评论，则需要注册登录后才能发表评论。右侧是一个基于用户的推荐模块。

2.2.5 模块 5——汽车推荐模块

如图 4 和图 5 所示，本网站利用基于用户的推荐算法构建推荐系统，利用用户的浏览记录数据和评论数据为用户构建 TOP-K 推荐列表，并将推荐信息内置于汽车详情页面和评论页面，较为符合逻辑。

2.2.6 模块 6——汽车公司模块

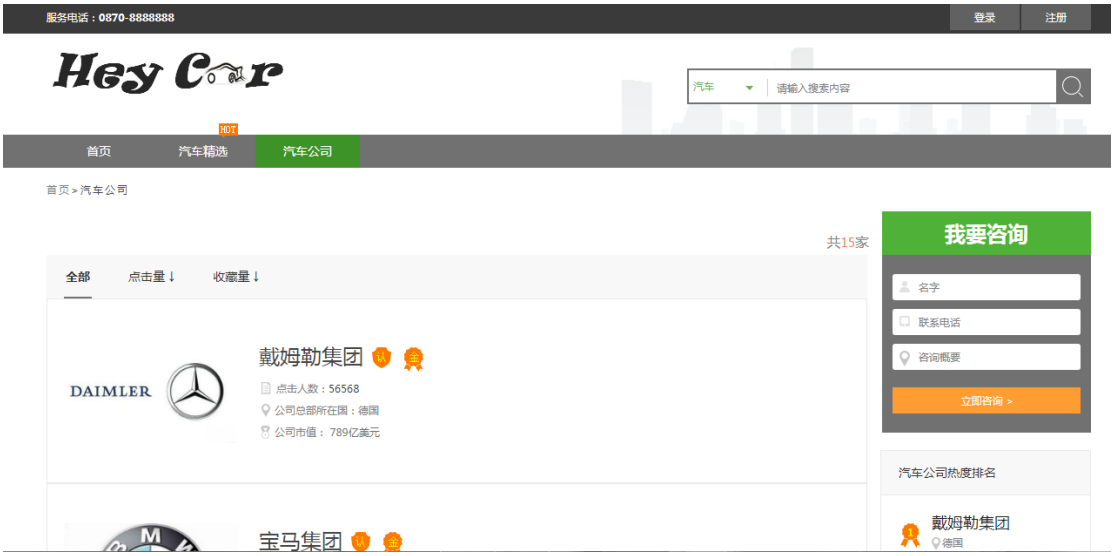


图 6：汽车公司列表页

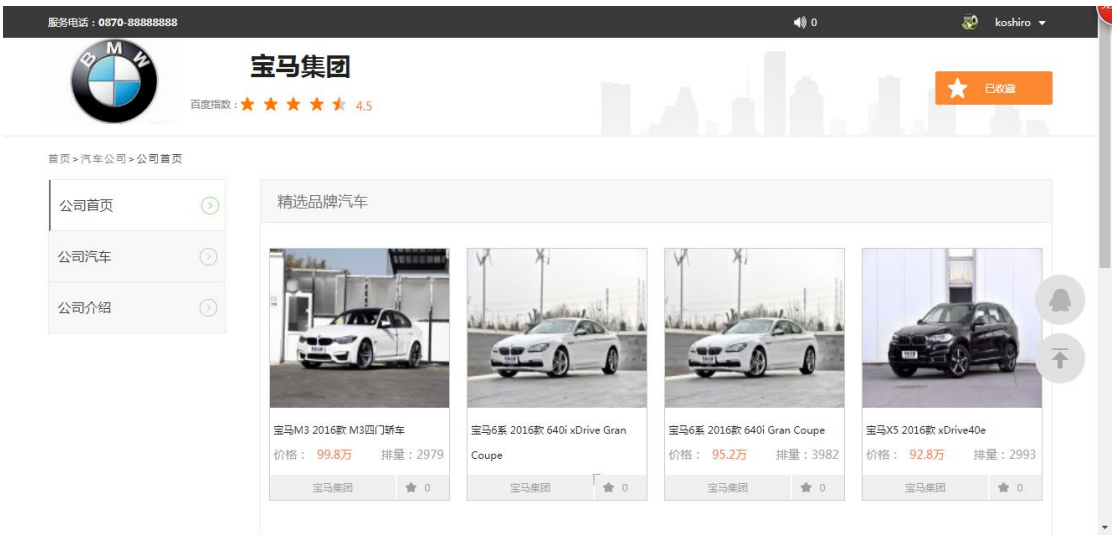


图 7：汽车公司详情页----公司首页

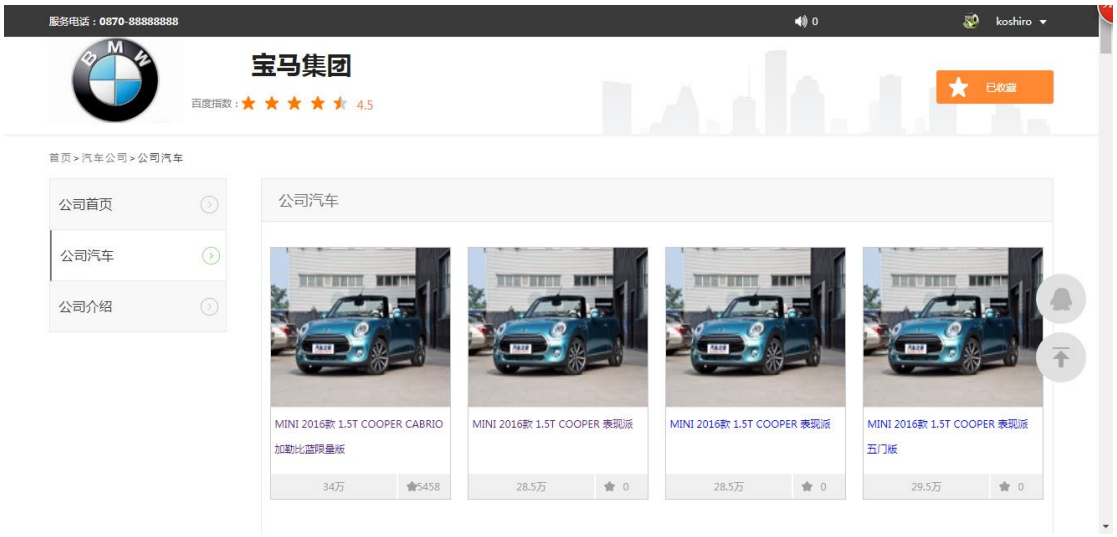


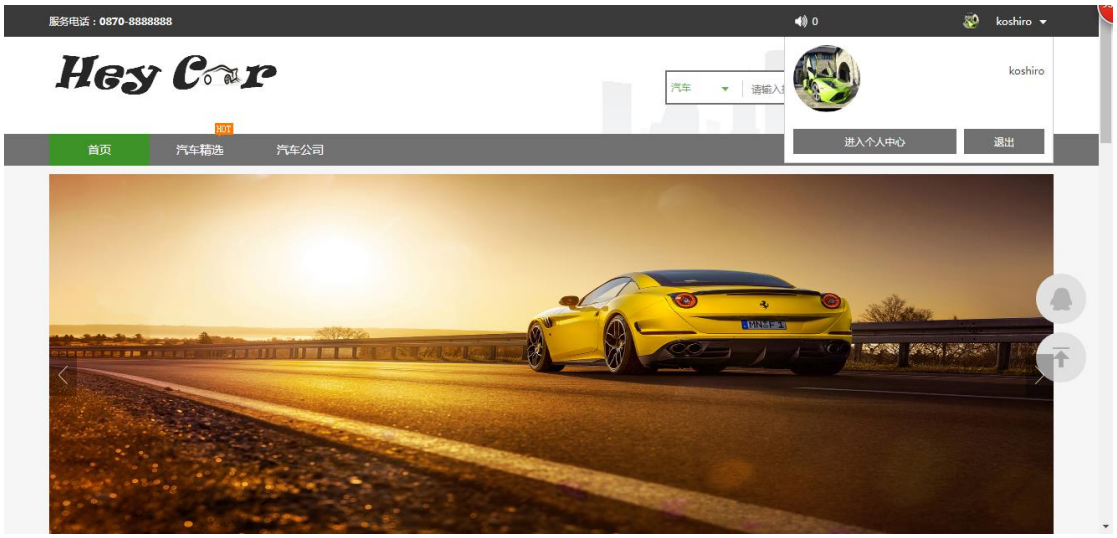
图 8：汽车公司详情页----公司代表型汽车

图 9：汽车公司详情页----公司介绍

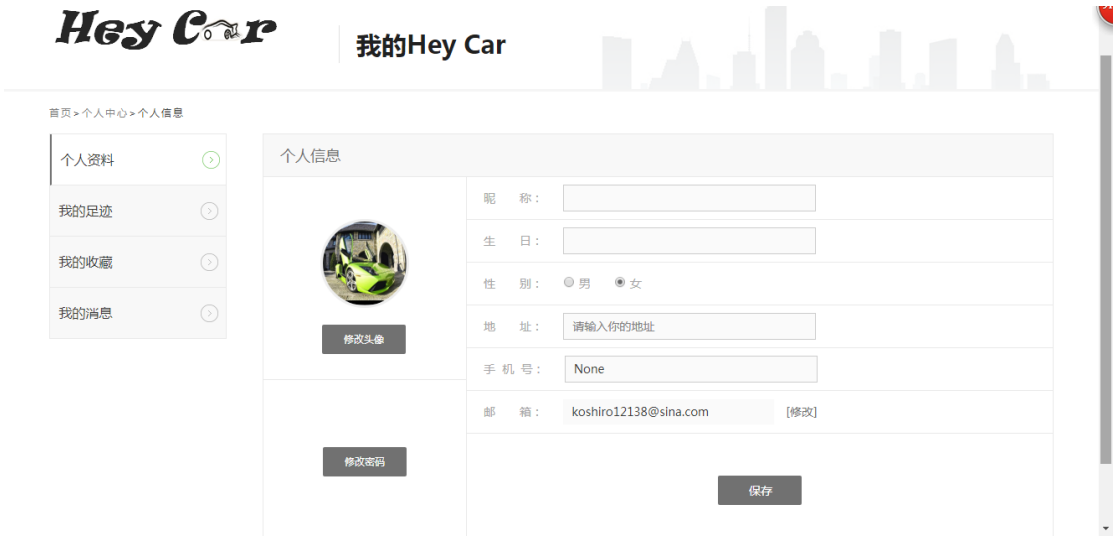
- 功能详情：
  - ✓ 汽车公司简介：如图 6 所示，点击首页上方“汽车公司”按钮，或者点击首页下方汽车公司缩略介绍部分的按钮，或者在首页右上方搜索框“公司”状态下搜索，我们都可以进入“汽车公司”页面。汽车公司页面介绍了十几家常见的汽车公司，点击对应图片可以进入详情页面。汽车公司有对应的汽车品牌，可以点击汽车品牌 logo，了解汽车品牌的更多知识和详情。
  - ✓ 公司首页：各公司的主页可以看到这家公司的精选品牌汽车，还可以在登录状态下对这家公司进行收藏。
  - ✓ 公司代表性汽车：可以在这个界面中看到该公司的所有车型代表。
  - ✓ 公司简介：在这个页面可以看到对当前公司的简要介绍。

2.2.7 模块 7——用户信息模块

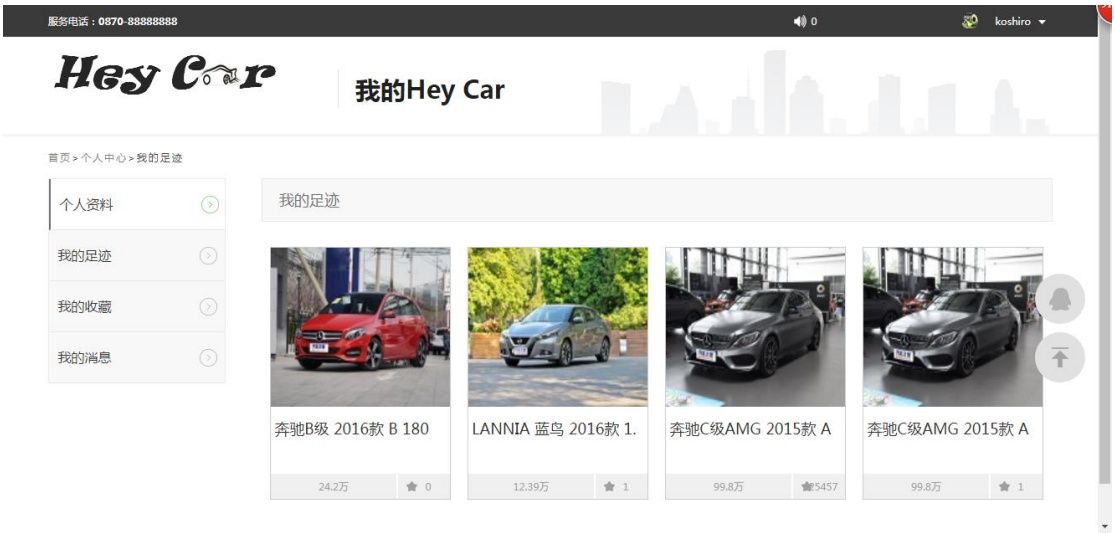
- 进入个人中心



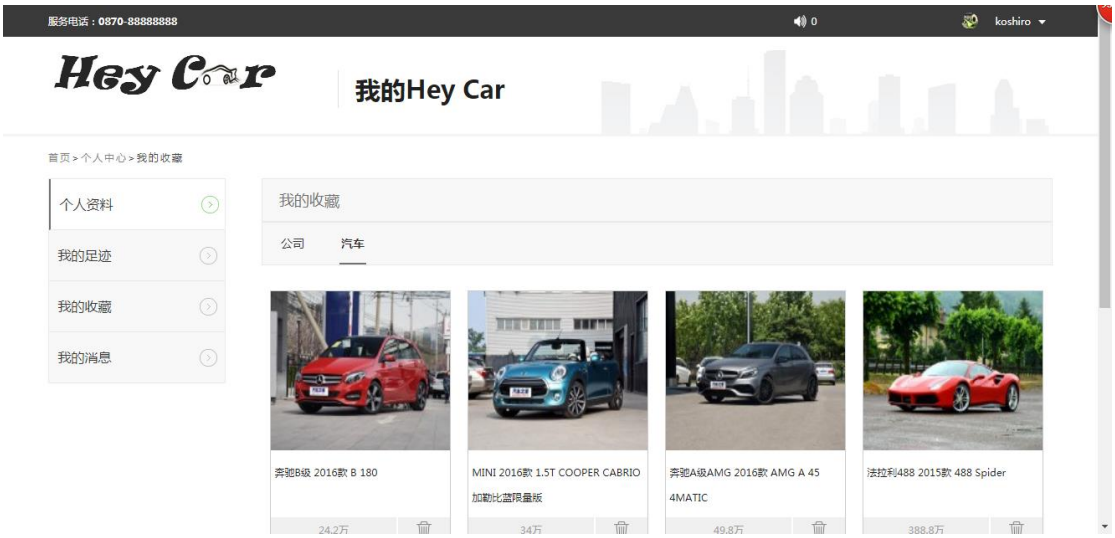
- 个人信息维护



- 浏览足迹



- 我的收藏



3 数据结构

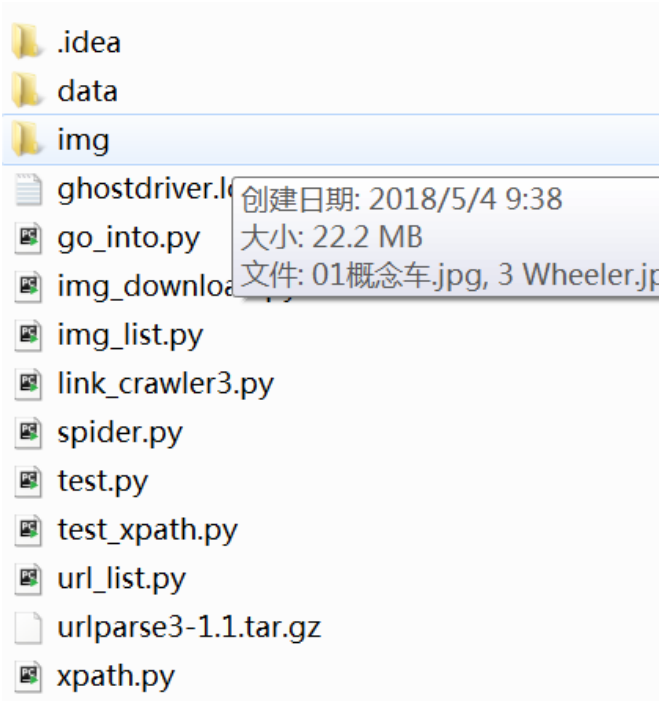
数据库表列表

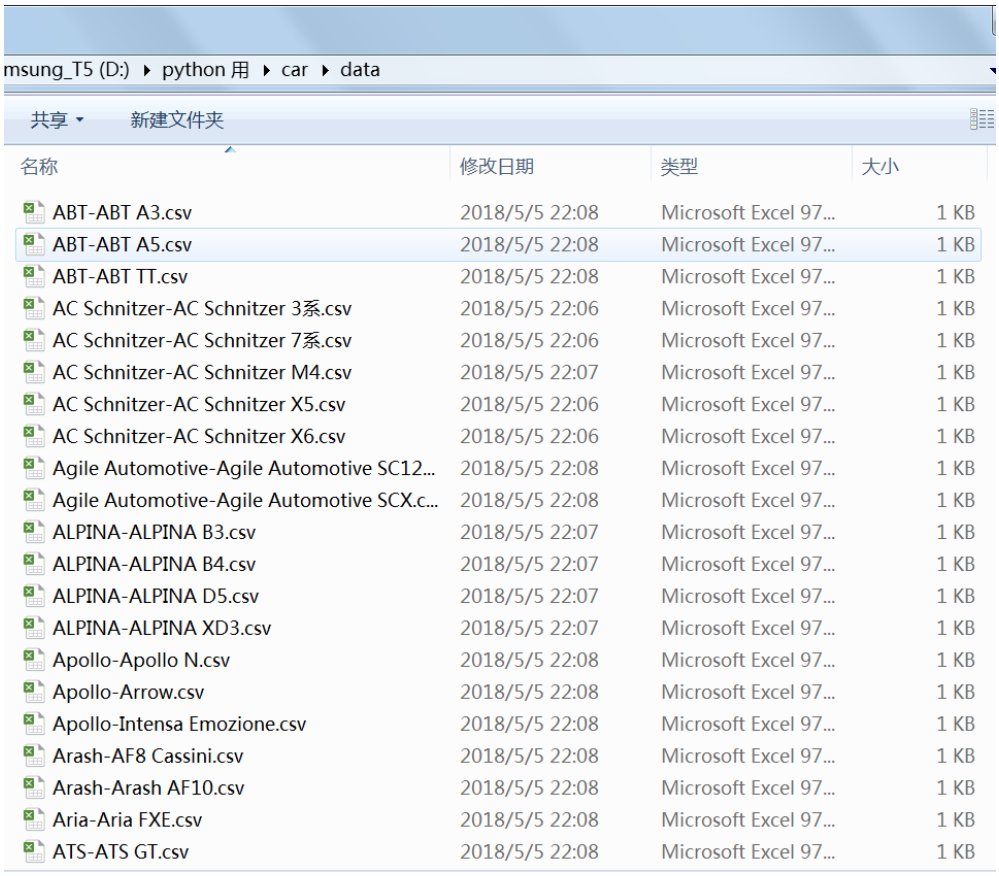
表名	作用
User	进行用户的信息记录以及用户的管理
Comment	以往用户评论
emotion_avg	每种车型的情感的平均值
Company	保存众多的公司信息，涉及的属性有：公司名称、公司总部所在的国家、公司的营业额、公司的市值、公司的联系方式、公司的图片。
Sentiment	每种评论对应的情感指数
Car	保存众多的汽车信息，其数据主要来自汽车之家、太平洋汽车网等多个汽车网站。
Operation	保存用户操作信息，诸如验证码、用户浏览记录和收藏记录以及用户与汽车和汽车公司之间的关联信息。

4 爬虫设计

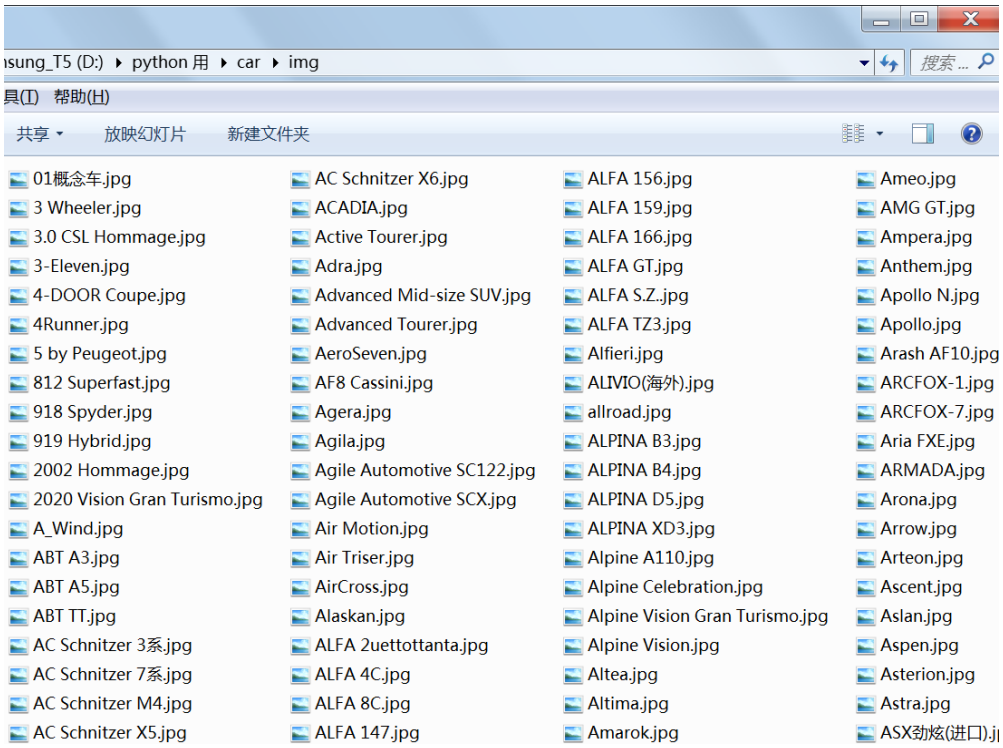
本次爬虫主要有两个部分，由四个分开的程序完成。

- 爬取所有汽车的详细文本信息：通过运行 url_list.py，我们可以获取汽车之家主页面的所有车名和链接列表，并将这些信息写入文件 url_list.csv，存储到 data 文件夹汇总；然后再运行 spider.py 依次读取 url_list 中的链接，进入每辆车的详情页面爬取具体信息，爬取到的信息以及第一步爬取到的车名被存入一个以车名命名的 csv 文件（例如：ABT-ABT A3.csv）。
- 爬取所有汽车的图片链接并批量下载：通过运行 img_list.py 爬取主页面汽车名和图片链接，并将信息写入 csv，生成 img 文件夹下的 img_list.csv 文件；然后再运行 img_download.py 依次读取 img_list 中的图片链接，将下载下来的图片以对应的车名存入 img 文件夹，从而完成批量下载。





Data 文件夹的部分内容



Img 文件夹的部分内容

本次总共调用了以下包用来完成爬虫：

import requests

用来根据目标网址打开网页，为 BeautifulSoup 的操作做准备

from bs4 import BeautifulSoup

用来完成各种网页链接、图片链接的定位以及少数文字标题信息的定位，爬取了汽车名、详细汽车名以及参考价格。

from selenium import webdriver

用来模拟登入网页并解析地址，用于下一步 xpath 定位

from lxml import etree

使用 xpath 对所需要爬取的具体信息进行定位，爬取油耗、上市时间、最大功率等二十余条详细信息。

import pandas as pd

将爬取下来的数据存储为 csv 文件输出

import urllib. requests

用来进行图片的批量下载

import string

辅助爬取下来的信息的处理

5 文本挖掘与情感分析

- 5.1 开发环境

  Python3.6 + spyder

- 5.2 理论依据

首先利用 pandas 构建数据框，把各个车型的评论集读入到一个数据框中。之后根据逐一修改过的停用词词典，采用 jieba 包中的 TF-IDF 算法对原评论集进行停用词的去除与分词，得到分词后的评论集；之后利用 nltk 包中的词频统计功能对分词后的数据集进行词频统计，最后根据清华大学的汉语情感词典计算各个车型的情感值，计算公式如下：

$$\frac{1}{n}\sum_{i=0}^{n} x_i k_i$$

其中 x_i 为第 i 个词汇的词频，k_i 为该词汇的情感分。得出每个车型的情感分之后，再对情感分进行标准化，第 i 种车型的标准化公式为：

$$\frac{x_i - min}{max - min}$$

这样所有车型的情感得分均为 0-1 之间的数字。最后进行频率统计，得到两个情感指数分界点，用于划分“口碑较差”，“口碑一般”和“口碑较好”。