

LogoSearcher实验报告

小组成员：

杨冬 516030910538

罗雨 516030910530

刘思辰 516030910528

常峰 516030910521

实验目的，产品设计

本组大作业的设计成果为，一针对电子产品品牌的搜索引擎（LOGO searcher）。主要功能为，输入目标品牌或电子产品的相关信息——包括品牌名称，品牌LOGO（图片），产品类别，产品名称，产品功能，品牌地区等——得到相关度较高的品牌信息

和产品搜索结果。

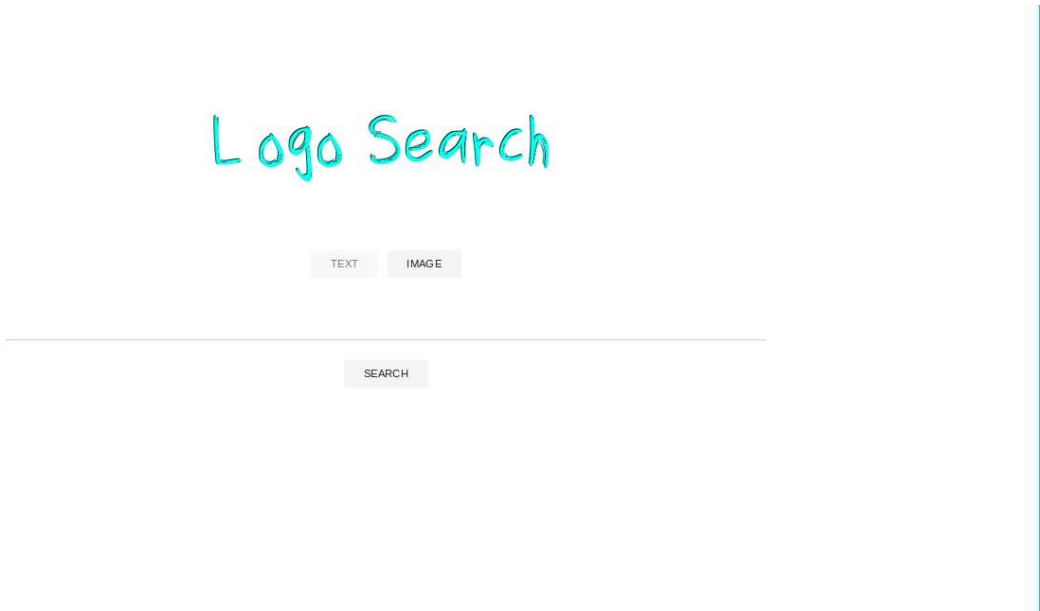
从设计成果出发，目标产品应当拥有以下模块：

1. 品牌信息爬取
2. 文字信息索引搜索
3. 图片信息搜索
4. 用户界面

实验成果，产品说明

成果展示

文字搜索界面



“三星”搜索结果

专业显示器



三星

三星集团是韩国最大的跨国企业集团，同时也是上市企业全球500强，三星集团包括众多的国际下属企业，旗下子公司有：三星电子、三星物产、三星航空、三星人寿保险等，业务涉及电子、金融、机械、化学等众多领域。三星集团成立于1938年，由李秉世创办。三星集团是家族企业，李氏家族世袭，旗下各个三星产业均为家族产业，并由家族中的其他成员管理，集团领导人已传至李氏第三代，李健熙为现任集团董事长，其子李在镕任三星电子副会长。2016年5月25日，华为公司在美国和中国提起对三星公司的知识产权诉讼，包括加州北区法院和深圳中级人民法院。7月21日，三星在北京知识产权法院起诉华为技术有限公司专利侵权，停止生产、销售mate8等机型，索赔8050万。9月6日，据道琼斯消息，在已出货或到达消费者手中的约250万部Note7手机中，有70%电池为由子公司：三星SDI（SAMSUNGSDICO.,LTD.）供应，剩下的部分则采用了中国新能源科技有限公司（ATL）的电池。10月，三星集团排2016年全球100大最有价值品牌第7名。10月，三星电子在召回GalaxyNote7手机后又宣布停止销售该手机，几乎宣布了这款旗舰手机的“死刑”，此举或令三星损失170亿美元。12月9日，三星准备在美国升级软件，永久禁用GalaxyNote7手机。2016年11月4日晚间消息，三星电子美国公司宣布，将主动召回280万台特定型号的顶部开盖洗衣机。

 三星 NL22B 透明屏幕 ¥ 5300 JD TB	 三星 820DXN-2 ¥ 700000 JD TB	 三星 DB4SD ¥ 11000 JD TB
 三星 mESSA ¥ 25000 JD TB	 三星 320MX-3 ¥ 4800 JD TB	 三星 mE46A ¥ 40000 JD TB

“日本”搜索结果

Home Contact





Logo Search 日本

耳机



JVC

日本胜利公司（日本ビクター株式会社，Victor Company of Japan, Limited），公司英文简称为JVC；在中国大陆、台湾也可适用正式译名杰伟世；香港曾译作星牌，但于1990年代开始已经甚少人用这个译名。JVC是一间日本消费性与专业电子企业，总部设在日本横滨，成立于1927年。该公司最为人熟悉就是展示日本第一台电视机及发明了VHS系统。

 JVC FW8 ¥ 1099 JD TB	 JVC HA-S88BN ¥ 1199 JD TB	 JVC HA-FX850 ¥ 1688 JD TB
		

图片搜索界面

Search

Home

三星监视器

三星监控摄像机

三星手机保护套

三星家庭影院

三星液晶显示器

三星专业显示器

三星数字标牌

三星耳机

三星硬盘录像机

三星镜头

End

Information

Search

Home

JVC耳机

日立投影机

日立空调

日立洗衣机

巴络络无线路由器

松下笔记本电脑

夏普手机

夏普冰箱

夏普洗衣机

夏普投影机

End

Information

Logo Search

☐ TEXT

☐ IMAGE

NO FILE SELECTED.

成果说明

由如上展示可看出，我们的成果已基本实现设计目标，包括以品牌为基础的搜索，针对品牌logo的图片搜索，以产品类型为基础的搜索，以产品属性为特征的搜索。直接搜索“日本”，我们可以得到日本知名耳机品牌“JVC”的简介，同时下方还附有其主要热门产品的型号及其相关信息。产品仍然有许多不足，包括信息还可以进一步扩充，推荐排序不够完善等。

产品原理分块说明

信息收集（杨冬同学负责）

利用课堂上所学习的爬虫知识爬取以“中关村在线”为主的信息。

获取品牌页面

我们的产品针对的是电子电器相关的品牌信息。所以爬取策略应与之相对应，比如爬取的时候地址url符合 `re.compile('http://\w+.zol.com.cn/manu_\d+.*')` 这个页面包含了某个品牌的某个类型的产品。相关的代码可以在 `crawler_thread.py` 中获取。

获取品牌相关信息

获取品牌基本介绍

在上一步爬取的html中我们可以获取品牌名称，品牌简短介绍，以及热门产品的部分信息。但是这种url对应的页面排版不一致，需要分类讨论，获取品牌名称，logo图片地址的方法相同，但是获取热门产品及其相关信息的方法不同，所以需要分类讨论。所以在 `process_all.py` 中有两个函数 `getPartialInfoFromType1()` 以及 `getPartialInfoFromType2()` 对应爬取两种不同网页的某些信息。

获取品牌公司介绍

其中获取品牌对应的公司信息主要依靠百度百科的资源，但是品牌以及公司的名字千奇百怪，比如“苹果”就是一个多义词，我们可以通过内容中是否有“公司”、“企业”、“集团”等相关词汇来判断获取到的内容是否为所需。另外有一些品牌中英文混合，比如MSI微星，Acer宏碁，需要分隔后再爬去相关信息。相关代码可以在 `crawler_company.py` 中获取。

获取品牌相关产品信息

在上一步中获取的热门产品信息存在严重缺失，由于这些只是来源于品牌页面，而不是产品的详细页面，所以图片或者价格缺失比较常见，所以在上一步的过程中，我记录了产品对应的详细页面的url，储存在 `prod_detail_url.txt` 中，运行 `prod_img_price_thread.py` 可以获取产品的大图，以及参考价格。以便在搜索结果页面展示。最后运行 `mk_prod_info_dic.py` 与 `makeup_missing.py` 可以对原本的信息进行完善。

其他说明

1. 数据量相对于个人工作来说比较大，所以采取多线程操作。大致能减少2~3倍的时间。

- 爬取网页信息时随机获取ua，一定程度上减少被拒绝的概率。
- 如果被拒绝，无法获取页面，则将其暂时储存在 `cannot_get.txt` 然后再寻找原因，再对其中网页重新爬取。重复几次，直到其中的产品数目小于一定的值。
- 对于品牌的信息获取还不够全面，有些品牌的信息无法从百度百科获取，就只能一句话带过；相对应产品只有图片和价格，还可以获取更多的信息。

信息索引（罗雨同学负责）

索引与搜索的analyzer：使用默认的SmartChineseAnalyzer

索引与搜索的analyzer：使用默认的SmartChineseAnalyzer。在建立索引中，由于SmartChineseAnalyzer效果不佳，故使用jieba分词以达成中文分词需求。在Searchfiles中仍使用SmartChineseAnalyzer，效果符合要求。

使用SmartChineseAnalyzer

```
analyzer = SmartChineseAnalyzer(Version.LUCENE_CURRENT)
```

使用jieba中文分词

```
if len(tmp) > 4:
    goods = '\n'.join(tmp[4 :])

    for i in range(len(tmp)):
        if i > 3:
            tmp3 = tmp[i].split()
            content.extend(jieba.cut(tmp3[1]))
    content = ' '.join(content)
```

搜索结果排序优化：信息加权

尝试使用lucene一般用文档加权手段setBoost（对Field），以减少对品牌名等重要信息的搜索的干扰，但效果不佳。改用暴力加权，即在搜索文档中多次添加品牌名和产品种类。

```
for i in range(20):
    content.append(name1)
    content.append(name2)
```

亮点与不足

不足

- 加权手段。
虽然一直在尝试对Field用setBoost，但由于时间不足（事前没有规划好）一直弄不好，最后不得不换用原本的暴力加权法。如果能用好自带加权手段，结果应该会更加精准。
- 排序
由于缺少热度、销量等讯息，不能达成很好的排序效果。

图像搜索（刘思辰同学负责）

图片搜索部分主要由两部分构成，第一部分是利用我们在课程中学习的canny算法对图像进行边缘检测，第二部分是canny处理后的二值图像几何不变矩（Hu矩）的计算与匹配。

Canny算法

Canny算法在课程中已经涉及就不加赘述。

Hu矩的计算与匹配

图片的Hu矩的计算基于以下定义：

$$m_{pq} = \sum_{y=1}^N \sum_{x=1}^M x^p y^q f(x, y) \quad p, q = 0, 1, 2, \dots$$
$$\bar{x} = m_{10} / m_{00}$$
$$\mu_{pq} = \sum_{y=1}^N \sum_{x=1}^M (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad p, q = 0, 1, 2, \dots$$
$$\bar{y} = m_{01} / m_{00}$$

图片的p+q阶矩和p+q阶中心矩

其中x, y分别为像素点的纵横坐标

图片的归一化中心矩

$$\eta_{pq} = \mu_{pq} / (\mu_{00}^\rho) \quad \rho = (p+q) / 2 + 1$$

图片的Hu矩

$$M1 = \eta_{20} + \eta_{02}$$

$$M2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$M3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$M4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$M5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) \\ + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2)$$

$$M7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) \\ - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2)$$

这七个矩构成了一幅图片的特征向量，在匹配时有以下三种匹配度计算方法，经过比较，发现第三种的正确率相对更高，我也选择的这种。

1. 由于我们组做的是关于logo的搜索，而logo又是人工制造的，相对自然图片更加简单，而且边缘非常清晰，所以我选择的是较为简单并且对边缘敏感的canny算法，然后对边缘图像进行相似度计算。
2. 由于Hu矩有旋转伸缩平移不变性，在理论上应该有较好的效果。

实际上的效果不是很理想，主要有以下原因：抗噪声能力弱，非图片的边缘噪声会使得图片的矩发生变化。

3. 算法对于图形性的logo识别能力较强，偏向于文字的偏弱。
4. 计算出的7维向量有着不同的数量级，选择的匹配方法很重要，如果只是使用求距离的方法，最后相当于直接忽视其中部分数据的影响，而且网上能够找到的三个算法的准确度也不是很高，仍然有出入，这是结果不理想的主要原因，所以我们算法最主要的改进点就在找到一个合适的匹配算法。

用户界面（常峰同学负责）

文字检索

文字部分主要使用了和前面中期整合差不多的方法，构建表单然后get，在主程序中运行文字检索函数，然后返回结果给模板即可。具体代码可在 start.py 中获取。

图像检索

图像检索中使用了POST的方式，首先将上传的图片保存在一个固定的位置，然后图像搜索的函数会固定搜索某个位置的图片，然后在整个函数运行后，删除掉保存图片。以下为代码示例。

```
class img_res:
    def POST(self):
```

```

x = web.input(myfile={})
filedir = './tmp'
if 'myfile' in x:
    filepath = x.myfile.filename.replace('\\', '/')
    # filename = filepath.split('/')[-1]
    filename = 'test.jpg'
    fout = open(filedir + '/' + filename, 'wb')
    fout.write(x.myfile.file.read())
    fout.close()

f = img_run()
res=[]
for i in f:
    grand = open(unicode(('brands/'+i), "utf8"), 'r')
    grand_list = grand.readlines()
    x=[]
    for k in range(len(grand_list)-7):
        x.append(grand_list[k])
    res.append(x)
for i in range(len(res)):
    for j in range(len(res[i])):
        if j == 0 or j > 3 :
            res[i][j] = res[i][j].split('\t')

os.remove('tmp/test.jpg')
return render.img_res(res)

```

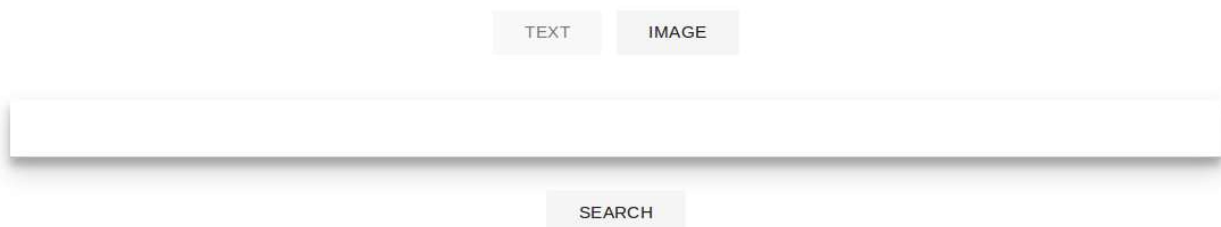
亮点和不足

亮点

我们组在界面上确实花了一些时间，甚至找出一个人专门负责界面以及最后的整合，使得整个的大作业少了一些粗糙感，整体看起来更为赏心悦目。

1. 触摸反馈

及时地反馈用户的操作，对鼠标放置的尽可能多的元素都有反馈，点击可交互文字时文字都会有轻微的变化，只是这些变化都集中在颜色，亮度，饱和度等等这些不容易产生“动”感的动画。如下图所示



2. 阴影和层级

通过不同的阴影卡片，不同的层级展示内容，而不是将内容一股脑儿全部扔出来，结果就是非常直观，简洁易懂。

Search

Home

三星监视器

三星监控摄像机

三星手机保护套

三星家庭影院

三星液晶显示器

三星专业显示器

三星数字标牌

三星耳机

三星硬盘录像机

三星镜头

End

Information

不足

整个网页的兼容性较差，如果遇到一些奇怪的内容，可能会出现排版问题。



三星 **MD32C**

¥ 7000

JD TB



三星 **DB55E**

¥ 9399

JD TB



三星 **ME95C**

¥ 859999

JD TB



三星 **DB32E**

¥ 3799

JD TB



三星 **EB40D**

¥ 2999

JD TB



三星 **PM55F**

¥ 14199

JD TB

总结

本程序主要利用上课所学习的爬虫知识获取网络中的信息，利用pylucene建立索引，利用webpy搭建网站。在图片搜索方面在上课所学习的canny检测的基础上利用了图片的其他特性来进行图片索引。有一定的效果，但是由于规模比较小，数据信息不够索引建立的还不够全面，图片搜索效果还有待提高，在以后的学习中可以再加入机器学习的方法提高其辨识能力。另一方面，组员之间相互合作，尽职尽责，每个人都功不可没。