

EE208 Final Project Report
Integrated Search Engine for Electronic Commodities
电子商品集成搜索引擎

董世文 方少恒 杨弘博 周李韬
GROUP 14
SJTU F1803016

2020 年 1 月 10 日

目录

前言	i
I Crawler	1
1 京东爬虫	3
1.1 网页 URL 爬取	3
1.2 商品信息提取	3
1.3 商品评论标签提取	3
1.4 商品评分计算	3
2 苏宁爬虫	5
2.1 网页 URL 爬取	5
2.2 商品信息爬取	5
2.3 商品评论标签爬取	5
2.4 商品评分计算	5
II Index & Search	7
3 构建索引	9
4 图片匹配	11
4.1 LOGO 匹配	11
4.2 LSH 匹配	11
5 商品检索与排序	13
5.1 按相关性检索	13
5.2 按属性检索	13

III Web Front-end	15
6 Web 框架	17
6.1 web.py 配置	17
6.2 Bootstrap 框架	17
7 网页布局	19
7.1 搜索主页	19
7.2 商品信息陈列	19
7.3 商品过滤功能	19
Appendix	21
总结	23

前言

Part I

Crawler

Chapter 1

京东爬虫

1.1 网页 URL 爬取

1.2 商品信息提取

1.3 商品评论标签提取

1.4 商品评分计算

Chapter 2

苏宁爬虫

2.1 网页 URL 爬取

2.2 商品信息爬取

2.3 商品评论标签爬取

2.4 商品评分计算

Part II

Index & Search

Chapter 3

构建索引

Chapter 4

图片匹配

4.1 LOGO 匹配

4.2 LSH 匹配

Chapter 5

商品检索与排序

5.1 按相关性检索

5.2 按属性检索

Part III

Web Front-end

Chapter 6

Web 框架

6.1 web.py 配置

6.2 Bootstrap 框架

Chapter 7

网页布局

7.1 搜索主页

7.2 商品信息陈列

7.3 商品过滤功能

Appendix

写作分配

Acknowledgements

总结