

Homework 4

- 信息检索系统

目标：

- 1. 写一个Web爬虫，爬取网站的网页；
- 2. 解析网页内容，对内容进行结构化，并存储到文件中；
- 3. 为内容建立索引；
- 4. 通过命令行进行内容检索，并展示内容列表


- 用药助手


- 爬取丁香园用药助手(<https://drugs.dxy.cn/>)数据





药理分类


 消化系统及代谢药 >

 血液和造血系统药物 >

 心血管系统药物 >

 皮肤病用药 >

 生殖泌尿系统和性激素类药物 >

 全身用激素类制剂(不含性激素) >

口腔科用药

美克威 息洛安

抗酸药及治疗消化性溃疡和胃肠胀气用药

丽奥佳

胃肠解痉药，抗胆碱药和胃动力药

胃复安针

止吐药和止恶心药

欧贝

阿司匹林肠溶片 - 北京曙光药业有限责任公司

商品名：阿司匹林肠溶片 成分：本品主要成份为：阿司匹林
适应症：本品为非甾体抗炎药。临床可用于抗血栓，预防一过性脑缺血发作、心肌梗死、心房颤动、人工心脏瓣膜、动静脉瘘或其他手术后的血栓形成。也可用于治疗不稳定型心绞痛。如用于解热镇痛、治疗风湿症，应选用大剂量规格的阿司匹林制剂。

氯己定苯佐卡因含片 - 贵州神奇药业有限公司

商品名：氯己定苯佐卡因含片 成分：本品每片含盐酸氯己定5毫克，苯佐卡因0.5毫克。
适应症：用于口腔溃疡、扁桃体炎。

醋酸氯己定栓 - 沈阳红旗制药有限公司

商品名：醋酸氯己定栓 成分：
适应症：本品具有消炎作用。用于宫颈糜烂、化脓性阴道炎、霉菌性阴道炎、滴虫性阴道炎。

乙酰吉他霉素片 - 云南永安制药有限公司

商品名：乙酰吉他霉素片 成分：本品主要成份为：乙酰吉他霉素。
适应症：本品主要适应于革兰阳性菌所致的各种感染，特别适应于金黄色葡萄球菌、肺炎球菌及表皮葡萄球菌引起的上、下呼吸道感染及表皮软组织感染。据文献报道，本品对百日咳、猩红热、中耳炎等也有良好的疗效。

克霉唑片

请仔细阅读说明书并在医生的指导下使用

【药品名称】

通用名称: 克霉唑片

英文名称: Clotrimazole Tablets

商品名称: 克霉唑片

【成份】

本品主要成分为克霉唑。

【适应症】

预防和治疗免疫抑制病人口腔和食管念珠菌感染，但因为本品口服吸收差，治疗深部真菌感染疗效差，不良反应又多见，现已很少应用，仅作局部用药。

【用法用量】

口服，一次 0.25 ~ 1 g，一日 0.75 ~ 3 g。小儿: 按体重一日 20 ~ 60 mg/kg，分 3 次服用。

【禁忌】

肝功能不全、粒细胞减少、肾上腺皮质功能减退及对本品过敏者禁用。

【注意事项】

因吸收差且毒性大而少用于内服。出现不良反应时, 应马上停药。

【孕妇及哺乳期妇女用药】

抽取各个属性

动物实验显示，应用 100 倍于人体剂量时具胚胎毒性。孕妇应权衡利弊后决定是否应用。本品是否经乳汁分泌尚缺乏资料。但由于许多药物经乳汁分泌，哺乳期妇女应慎用。

【药理作用】

本品属吡咯类抗真菌药，对白念珠菌则可抑制其自芽孢转变为侵袭性菌丝的过程。本品具广谱抗真菌活性，对表皮癣菌、毛发癣菌、曲菌、着色真菌、隐球菌属和念珠菌属均有较好抗菌作用，对申克氏孢子丝菌、皮炎芽生... [登录](#)

【药代动力学】

本品口服后很少吸收，成人口服 3 g 后，2 小时的血药峰浓度仅 1.29 mg/L，6 小时为 0.78 mg/L。连续给药时，由于肝酶的诱导作用血药浓度反而下降。消除半衰期为 4.5 ~ 6 小时... [登录](#)

【化学成份】

【是否OTC】

甲类 OTC

• 爱课程

— <https://www.icourses.cn/cuoc/>

[首页](#)[在线开放课程](#)[视频公开课](#)[资源共享课](#)[学校云](#)[客户端](#) [登录](#) | [注册](#)[中国大学MOOC](#) | [中国职教MOOC](#) | [中国大学先修课](#) | [教师教育](#) | [考研](#) | [思政](#) | [一流大学系列课程](#) | [AI专业培养方案](#)[分类：](#) [全部](#) [哲学](#) [经济学](#) [法学](#) [教育学](#) [文学](#) [历史学](#) [理学](#) [工学](#) [农学](#) [医学](#) [管理学](#) [艺术学](#) [就业创业课](#)

共992门课

**陶瓷艺术鉴赏与制作**

视频公开课

汤书昆 | 中国科学技术大学

**大学生心理健康**

视频公开课

樊富珉 | 清华大学

**薪火传承·中国传统哲学通论**

视频公开课

宋志明 | 中国人民大学

**认识宇宙**

视频公开课

向守平 | 中国科学技术大学

**古希腊文明的兴衰**

视频公开课

赵林 | 武汉大学

**千古名月**

视频公开课

于丹 | 北京师范大学

**人工智能PK人类智能**

视频公开课

蔡自兴 | 中南大学

**六大名著导读**

视频公开课

陈洪 | 南开大学

**文化传承与建筑创新**

视频公开课

何镜堂 | 华南理工大学

**视觉文化批评**

视频公开课

冯原 | 中山大学



汤书昆

教授
中国科学技术大学

课程介绍

本课程旨在通过陶瓷这一中华文明的经典产物,让人们领会“China”的无穷魅力和伟大创新。课程以中国陶瓷的诞生、发育、演化、外传为主线,系统地讲述陶瓷艺术和陶瓷技术对中国及世界物质文化生活的重大影响,以及对人类精神文明的长期陶冶。课程从陶瓷之路、青花瓷、陶艺的起源、彩陶之美、陶俑之美、陶瓷艺术鉴赏与当代科技的交融等方面进行了专题讲授,希望能让广大学习者从陶瓷的历史背景和审美特点等方面更整体的认识中国陶瓷文化的无穷魅力!

本讲介绍

描述中国瓷的起源与材料,探讨陶与瓷从材料、工艺到审美质感的区别,讲述了中国人发明瓷器制造工艺后,拥有这项垄断技术长达千余年的辉煌历程,以及通过著名的陶瓷之路销往全世界的贸易文化。以中国最早的瓷都和秘色瓷为例,展现了中国瓷器如玉如冰的精粹意蕴;以著名的龙泉青瓷在宋元两代的外销为例,展现历史上龙泉瓷外销目的地遍布天下的史实;以禁卫军瓷和法王路易十四的传奇故事见证了当年欧洲对中国高温硬质瓷的高度迷恋。



扫码下载APP
随时随地学课程



第1/7讲 (总时长: 0时39分17秒) 19520人学习 120人评论



发表了120条评论 共12页

Homework 4

- 关键技术：
 - 爬虫
 - 信息抽取
 - 索引建立
 - 查询

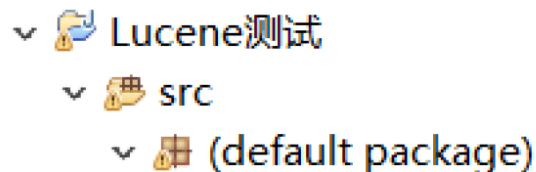
Homework 4

- Tips:
- 1. 如何在Eclipse中引入jar包

Homework 作业补充

一，在 Eclipse 引入 jar 包

- 1， 下载需要的 jar 包。在 lucene 中，需要 Lucene-core-4.10.jar 等三个 jar 包以及 IKAnalyzer2012FF.jar 分词器 jar 包。本次作业所有的 jar 包均会直接提供。
- 2， 在工程中新建文件夹，命名为 lib（右单击项目，new->folder）。将下载的 jar 包复制粘贴到此文件夹中。



Homework 4

- Tips
- 2. JAVA爬虫
 - crawler4j
 - <https://github.com/yasserg/crawler4j>

crawler4j

build passing maven-central v4.4.0 chat online

crawler4j is an open source web crawler for Java which provides a simple interface for crawling the Web. Using it, you can setup a multi-threaded web crawler in few minutes.

– JSOUP

- <https://blog.csdn.net/zbX931197485/article/details/78582407>
- jsoup 是一款 Java 的HTML 解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据，可以看作是java版的jQuery。
jsoup的主要功能如下：
从一个URL，文件或字符串中解析HTML；
使用DOM或CSS选择器来查找、取出数据；
可操作HTML元素、属性、文本；
jsoup是基于MIT协议发布的，可放心使用于商业项目。官方网站：<http://jsoup.org/>

Homework 4

- 基于jsoup: Java HTML Parser来抽取信息 (如标题等, 相同的网站同一个模板), 利用正则表达式来建立模板
 - <https://jsoup.org/>

```
File input = new File("/tmp/input.html");
Document doc = Jsoup.parse(input, "UTF-8", "http://example.com/");

Elements links = doc.select("a[href]"); // a with href
Elements pngs = doc.select("img[src$=.png]");
// img with src ending .png

Element masthead = doc.select("div.masthead").first();
// div with class=masthead

Elements resultLinks = doc.select("h3.r > a"); // direct a after h3
```

Homework 4

- Tips
- 3. 利用Lucene对文本进行索引，并进行检索

Apache Lucene™ is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

<http://lucene.apache.org/core/>

Homework 4

- 3. 利用Lucene对文本进行索引，并进行检索（输入检索词，查询得到相关的问题（或课程）列表，并显示详细信息。
 - 建索引和检索的简例

Homework 4

- 作业包括： java文件 + 文档 + 数据
- 作业打包上传到ftp homework/homework4下
- 文件： 学号_姓名_homework4.rar

Homework 4

- 代码要求：
 - 遵守编程规范，如命名、注释等规范
 - 遵守面向对象的设计原则
 - 考虑异常处理等应用

Homework 4

- 文档要求：
 - 按附件格式样例，至少包括：引用、总体设计、详细设计、测试与运行、总结
 - 包括：数据格式说明
 - 附加：程序中包含的其他特色或改进
 - 附加：数据的丰富程度