北京大学计算机科学技术研究所
Institute of Computer Science & Technology Peking University

Lemur工具简介

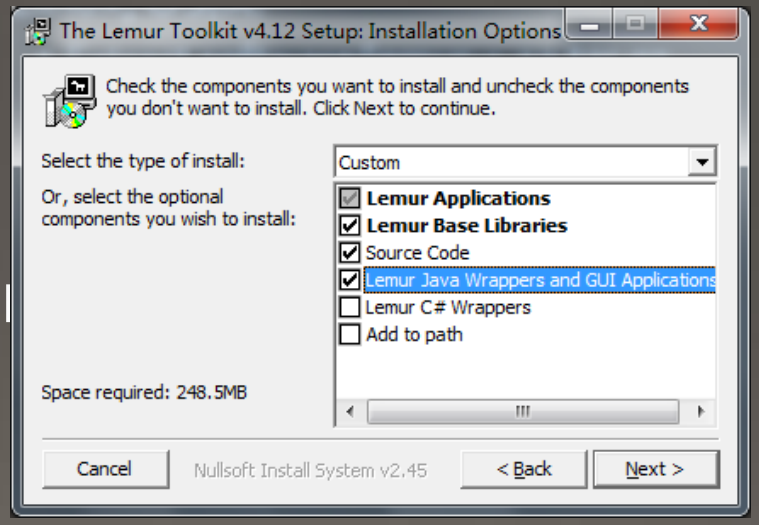
强闰伟
qiangrw@gmail.com

Lemur简介

- Lemur是一个用于辅助语言模型和信息检索研究的一个工具包。
- 信息检索广泛被认为是一种特殊的分布检索，它包含结构化的查询，跨语言的查询，统计，过滤，分类。
- Lemur系统的底层架构正是为了实现上面这些技术的，它提供了很多有用的样例应用，并且可以让你方便的定制自己的应用。
- Lemur 4.12 => Indri

安装工具

- Linux, OS/X（推荐):
  - Extract software/lemur-4.12.tar.gz
  - ./configure --prefix=/install/path
  - ./make
  - ./make install
- Windows
  - Run software/lemur-4.12-install.exe
  - Documentation in **windoc/index.html**
  - Set Environment Variable



Lemur工具简介

- 建立索引
- 检索
- 评价结果

建立索引 文档格式

- Lemur支持的文档格式
  - TREC Text （推荐使用）
  - TREC Web
  - HTML
- Indri 支持更多文档格式
  - Plain Text
  - DOC
  - PPT
  - XML
  - PDF
  - MBOX

建立索引 TREC Text

写一个简单的脚本将你的文档用TREC格式包装起来。

```
<DOC>
        <DOCNO>document_id</DOCNO>
        <TEXT>
        Index this document text.
        </TEXT>
</DOC>
```

索引类型

KeyFile

- Term Positions
- Metadata
- Offline Incremental
- InQuery Query Language

Indri

- Term Positions
- Metadata
- Fields / Annotations
- Online Incremental
- InQuery and Indri Query Languages

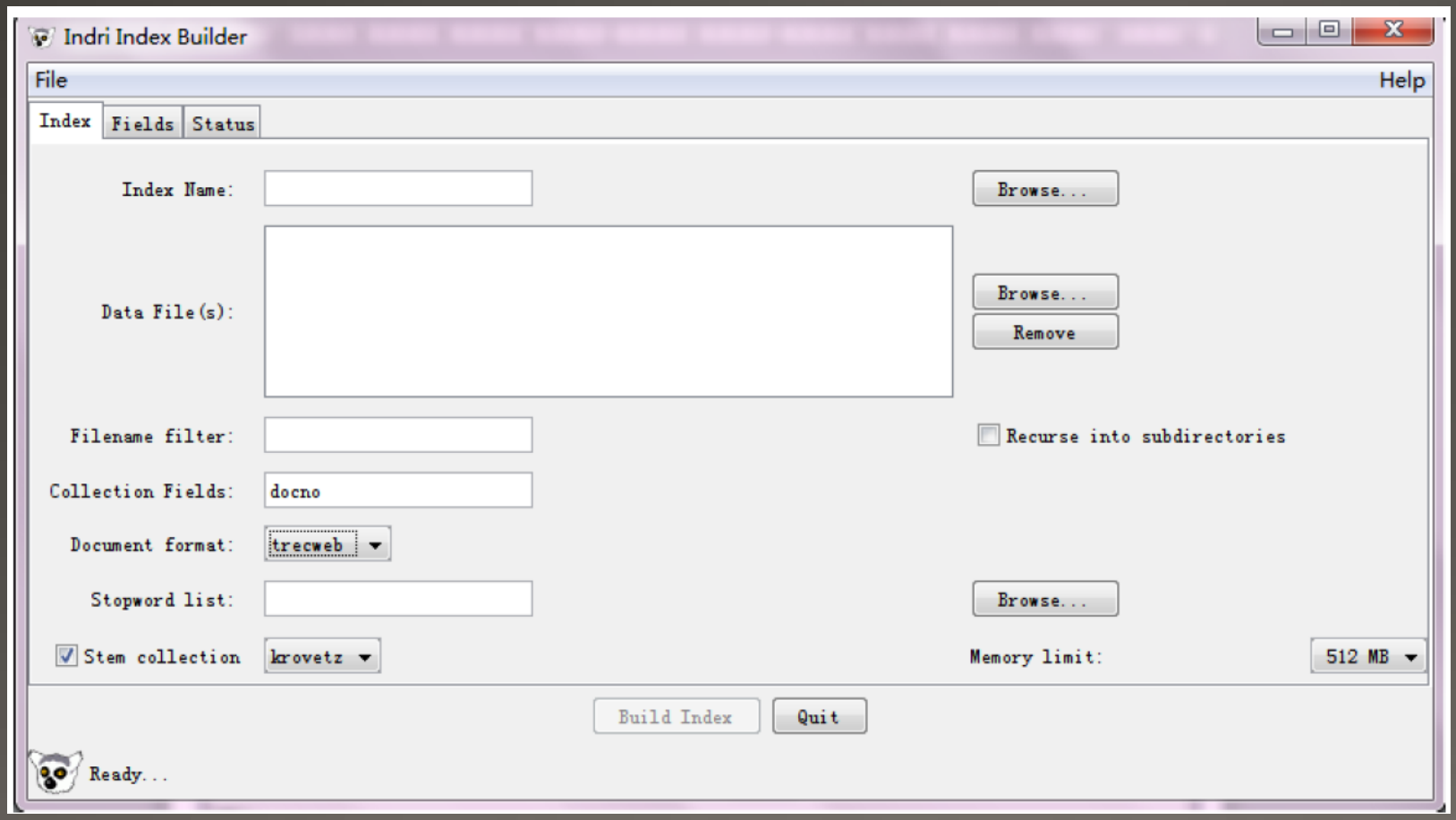
建立索引

- 建立索引命令
  - **$BuildIndex**/IndriBuildIndex <parameter_file>
- 建立索引参数
  - 数据文件的位置
  - 存放索引的位置
  - 运行命令时使用内存
  - 指定停用词，词形变换算法
  - ...

建立索引参数示例

```
<parameters>
    <dataFiles>/path/to/data</dataFiles>
    <docFormat>trec</docFormat>
    <index>/path/to/index</index>
    <indexType>key</indexType>
    <memory>1024MB</memory>
    <stemmer>Porter</stemmer>
    <countStopWords>True</countStopWords>
    <stopWords>/path/to/stopwordfile</stopWords>
</parameters>
```

Lemur Indexing UI



处理查询

- 查询文本需要做和文档模型相同的预处理操作。
- 查询文本的格式和文档文本相同，用ParseToFile命令进行预处理。
- $ParseToFile parameter_file data_file

ParseToFile

```
<parameter>
        <docFormat>trec</docFormat>
        <outputFile>\path\to\queryfile</outputFile>
        <stemmer>Porter</stemmer>
        <stopwords>\path\to\stopword</stopwords>
</parameter>
```

检索

- 检索命令
  - **$RetEval**/IndriRunQuery <parameter_file>
- 检索参数
  - 索引位置
  - 查询（集）
  - 运行命令占用内存数
  - 格式参数（核心，包括使用何种检索算法）
  - …

检索模型 RetModel

- RetModel
  - tfidf for TFIDF
  - okapi for Okapi
  - kl for Simple KL
  - cos for cosine similarity
  - indri for Indri structured query language
- See windoc/lemur-batch-retrieval.html for detials

检索参数举例

```
<parameters>
  <index>/path/to/the/index</index>
  <retModel>tfidf</retModel>
  <textQuery>/path/to/the/query</textQuery>
  <memory>1024MB</memory>
  <resultFile>/path/to/the/result</resultFile>
  <count>1000</count>
  <runID>RunTag</runID>
  <trecFormat>true</trecFormat>
</parameters>
```

Lemur Retrieval UI

更多资料

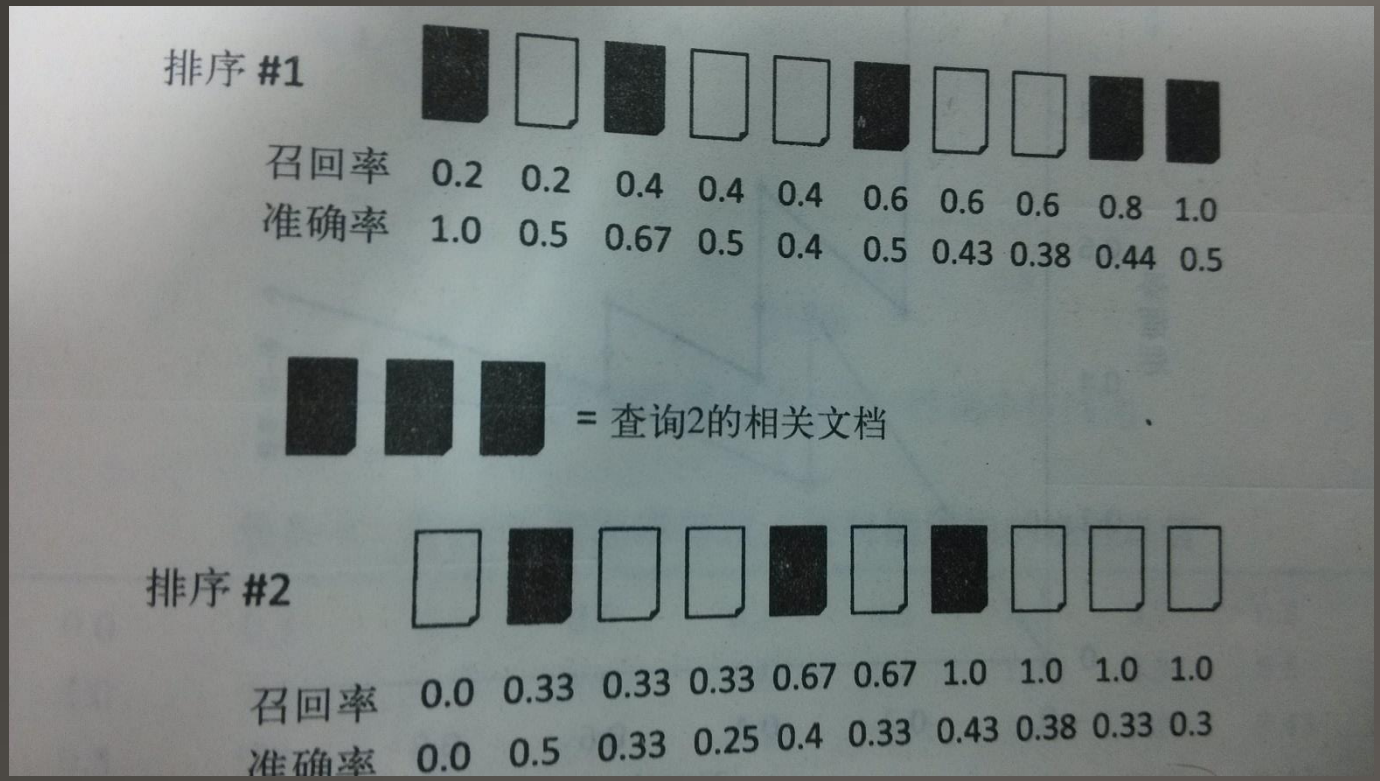
- Beginner's Guide
  - http://www.cs.cmu.edu/~lemur/LemurGuide.html
- Lemur Related Page
  - http://www.lemurproject.org/doxygen/lemur/html/pages.html
- Lemur 4.12 使用手册
  - http://lemurproject.org/lemur/
- 利用Lemur API来编写检索模型
  - http://www.cs.cmu.edu/~lemur/3.0/api.html
- 用其他开源工具:
  - https://lucene.apache.org/
  - ...

评价结果

- 使用Top 1000 文档的MAP （Mean Average Precision）

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

$$\mathrm{AveP} = \frac{\sum_{k=1}^{n} (P(k) \times \mathrm{rel}(k))}{\text{number of relevant documents}}$$



查询1 的平均准确率=(1.0+0.67+0.5+0.44+0.5)/5=0.62
查询2的评价准确率=(0.5+0.4+0.43)/3=0.44
MAP = (0.62+0.44)/2=0.53

提升检索效果

- 文本扩展（是否需要区分对待原始文本和链接信息）
- 利用微博特有的特征：主题、转发、用户提及信息
- 查询扩展（使用多少伪相关文档，如何扩展）
- 直推式学习
- ...

Waggle 竞赛提交系统

- http://webkdd.org/waggle
- 学号注册，密码将发送至PKU（的垃圾）邮箱
- 关于提交
  - 必须填写一个提交名称（Runtag）
  - 提交文件必须为*.txt后缀
  - 提交文件大小大小不能超过10MB
  - 注意提交格式（两个作业都不需要有header）

大作业打分依据

- 排名
- 代码/脚本
- 实验报告
  - 语言：准确，逻辑正确
  - 内容：充实程度
  - 组织：清楚的文章结构
- 课堂汇报

所有文件解压密码：

谢谢！