

# 信息抽取系统

- 项目概述
  - 项目要求
  - 项目架构
- 详细设计
  - 数据爬取
    - OiWiki 数据爬取
    - The Rust Programming Language 数据爬取
    - 数据存储
  - 后端设计与实现
    - 接口展示
    - 默认信息提取
    - 正则和信息点信息提取
    - 缓存优化
  - 前端设计与实现
    - 关键词、图片搜索
    - 搜索结果展示
    - 正则和指定信息搜索
    - 搜索结果反馈
      - 实体反馈
      - 热词反馈
      - 信息抽取反馈
      - 整体准确率评价反馈
  - 信息抽取服务
    - 热词和实体抽取
    - 正则和指定信息点抽取
    - 实体词检测和提取
  - 多媒体信息抽取服务
  - 抽取结果评价
    - 实体反馈
    - 热词反馈
    - 信息抽取反馈
    - 整体准确率评价反馈
- 优化与创新性
  - 后端-算法优化

- 前端

- 环境和社会可持续发展思考
- 实验总结
- 实验分工

# 项目概述

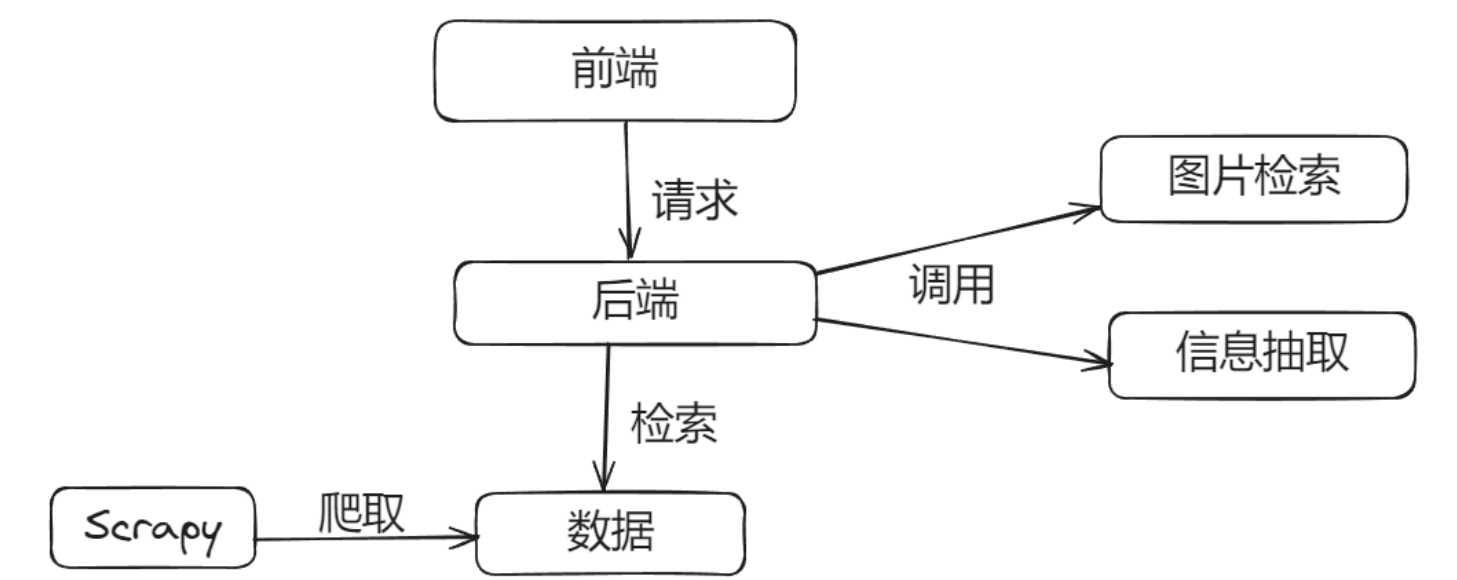
## 项目要求

基本要求：自己动手设计实现一个信息抽取实验系统，中、英文皆可，可以在作业 2 信息检索系统的基础上实现，也可以单独实现。特定领域语料根据自己的兴趣选定，规模不低于 100 篇文档，进行本地存储。对自己感兴趣的特定信息点进行抽取，并将结果展示出来。其中，特定信息点的个数不低于 5 个。可以调用开源的中英文自然语言处理基本模块，如分句、分词、命名实体识别、句法分析。信息抽取算法可以根据自己的兴趣选择，至少实现正则表达式匹配算法的特定信息点抽取。最好能对抽取结果的准确率进行人工评价。界面不作强制要求，可以是命令行，也可以是可操作的界面。提交作业报告和源代码。鼓励有兴趣和有能力的同学积极尝试优化各模块算法，也可关注各类相关竞赛。

扩展要求：鼓励有兴趣和有能力的同学积极尝试多媒体信息抽取以及优化各模块算法，也可关注各类相关竞赛。自主开展相关文献调研与分析，完成算法评估、优化、论证创新点的过程。

## 项目架构

本次实验在第二次实验的基础上完成，除了数据爬取、前端展示、后端处理和图片检索服务外还而额外增加了**信息抽取服务**，整体架构图和详细内容如下：



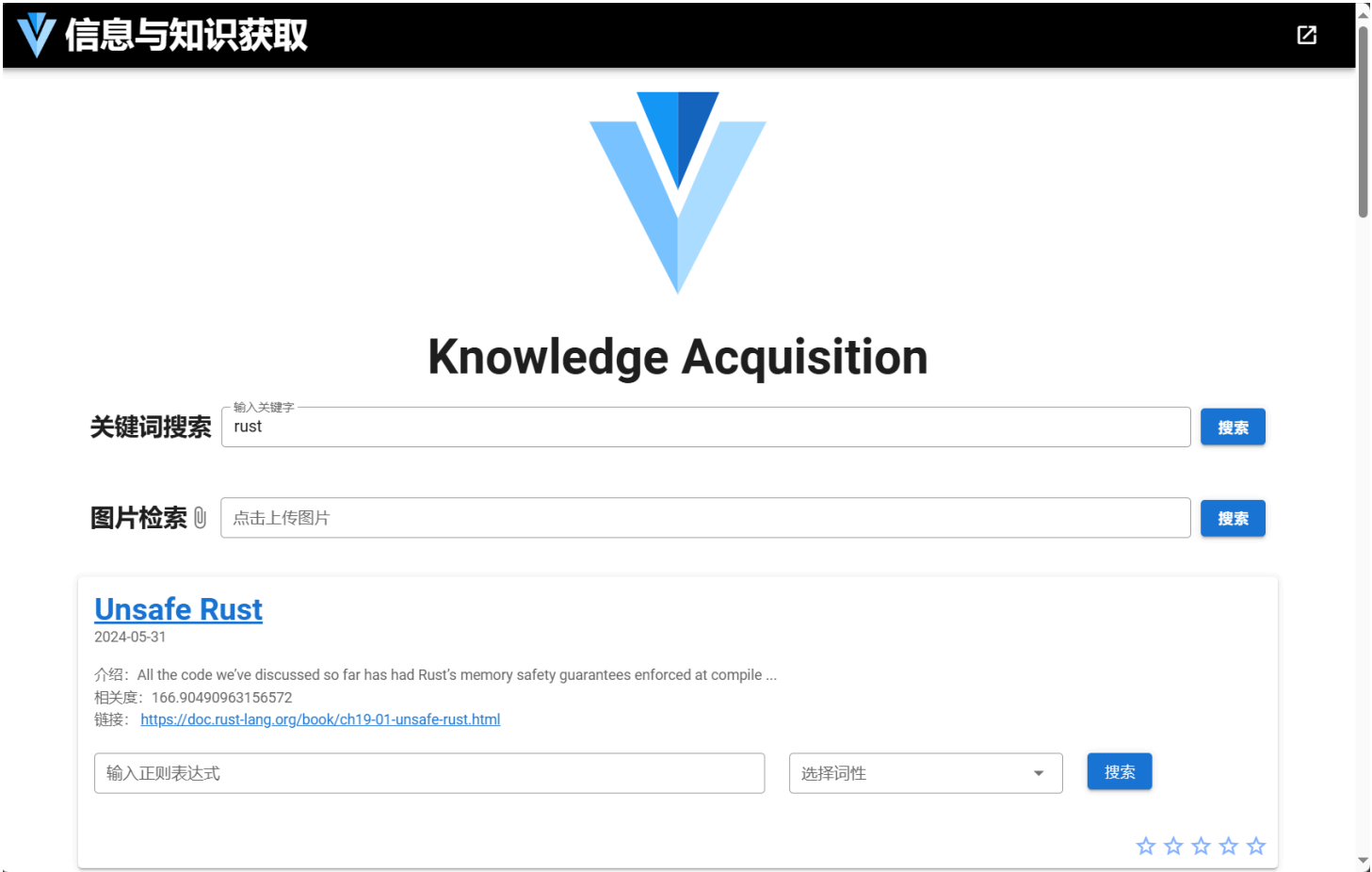
- 数据爬取使用 Scrapy 框架爬取网页文章并存储到 json 文件中，方便后续使用；
- 前端展示使用 Vue 框架实现，为用户提供了清晰直观的操作界面；
- 后端使用基于 Go 的 Gin 框架开发，接收用户请求并处理，并且对文章进行初始化处理，如：使用结巴框架进行分词、建立倒排索引、计算 TF-IDF 值等；

- 图片检索服务则使用 Python 进行开发，使用 Flask 框架给后端提供调用接口，使用 Tensorflow 框架对图片进行识别并提取关键词用于检索。
- 信息抽取服务也是使用 Python 进行开发，针对中英文两种语言分别使用 jieba 和 spaCy 两种框架进行处理和抽取。

项目运行截图如下：

前端界面展示：

整体内容：



结果列表

## Unsafe Rust

2024-05-31

介绍: All the code we've discussed so far has had Rust's memory safety guarantees enforced at compile ...

相关度: 166.90490963156572

链接: <https://doc.rust-lang.org/book/ch19-01-unsafe-rust.html>




☆☆☆☆☆

## Appendix D - Useful Development Tools

2024-05-31

介绍: In this appendix, we talk about some useful development tools that the Rust project provides. We'l...

相关度: 0.4401068682116109

链接: <https://doc.rust-lang.org/book/appendix-04-useful-development-tools.html>




☆☆☆☆☆

## RefCell<T> and the Interior Mutability Pattern

2024-05-31

介绍: is a design pattern in Rust that allows you to mutate data even when there are immutable references...

相关度: 0.07846236503761751

链接: <https://doc.rust-lang.org/book/ch15-05-interior-mutability.html>

详细信息:

## Unsafe Rust

2024-05-31

介绍: All the code we've discussed so far has had Rust's memory safety guarantees enforced at compile ...

相关度: 166.90490963156572%

链接: <https://doc.rust-lang.org/book/ch19-01-unsafe-rust.html>

### Unsafe Rust

All the code we've discussed so far has had Rust's memory safety guarantees enforced at compile time. However, Rust has a second language hidden inside it that doesn't enforce these memory safety guarantees: it's called *unsafe Rust* and works just like regular Rust, but gives us extra superpowers.

Unsafe Rust exists because, by nature, static analysis is conservative. When the compiler tries to determine whether or not code upholds the guarantees, it's better for it to reject some valid programs than to accept some invalid programs. Although the code *might* be okay, if the Rust compiler doesn't have enough information to be confident, it will reject the code. In these cases, you can use unsafe code to tell the compiler, "Trust me, I know what I'm doing." Be warned, however, that you use unsafe Rust at your own risk: if you use unsafe code incorrectly, problems can occur due to memory unsafety, such as null pointer dereferencing.

Another reason Rust has an unsafe alter ego is that the underlying computer hardware is inherently unsafe. If Rust didn't let you do unsafe operations, you couldn't do certain tasks. Rust needs to allow you to do low-level systems programming, such as directly interacting with the operating system or even writing your own operating system. Working with low-level systems programming is one of the goals of the language. Let's explore what we can do with unsafe Rust and how to do it.

### Unsafe Superpowers

To switch to unsafe Rust, use the `unsafe` keyword and then start a new block that holds the unsafe code. You can take five actions in unsafe Rust that you can't in safe Rust, which we call *unsafe superpowers*. Those superpowers include the ability to:

- Dereference a raw pointer
- Call an unsafe function or method
- Access or modify a mutable static variable
- Implement an unsafe trait
- Access fields of `union`s

It's important to understand that `unsafe` doesn't turn off the borrow checker or disable any other of Rust's safety checks: if you use a reference in unsafe code, it will still be checked.

The `unsafe` keyword only gives you access to these five features that are then not checked by the compiler for memory safety. You'll still get some degree of safety inside of an unsafe block.

In addition, `unsafe` does not mean the code inside the block is necessarily dangerous, so that it will definitely have memory safety problems; the intent is that as the programmer you'll

反馈：

信息与知识获取

Accessing Fields of a Union

The final action that works only with `unsafe` is accessing fields of a `union`. A `union` is similar to a `struct`, but only one declared field is used in a particular instance at one time. Unions are primarily used to interface with unions in C code. Accessing union fields is unsafe because Rust can't guarantee the type of the data currently being stored in the union instance. You can learn more about unions in [the Rust Reference](#).

When to Use Unsafe Code

Using `unsafe` to take one of the five actions (superpowers) just discussed isn't wrong or even frowned upon. But it is trickier to get `unsafe` code correct because the compiler can't help uphold memory safety. When you have a reason to use `unsafe` code, you can do so, and having the explicit `unsafe` annotation makes it easier to track down the source of problems when they occur.

关键字: All the code we've discussed so far has had Rust's...

语言: 英文

实体	频率	反馈
ABI	3	☆☆☆☆
Listing	3	☆☆☆☆
five	5	☆☆☆☆
one	4	☆☆☆☆
two	5	☆☆☆☆
热词	频率	反馈
Rust	36	☆☆☆☆
code	41	☆☆☆☆
raw	27	☆☆☆☆
unsafe	45	☆☆☆☆
,	95	☆☆☆☆

输入正则表达式

选择词性

搜索

☆☆☆☆

图片搜索：

信息与知识获取

Knowledge Acquisition

关键词搜索

mask comic\_book book\_jacket

搜索

图片检索

photo.png

搜索

Introduction

2024-05-31

介绍: Welcome to , an introductory book about Rust. The Rust programming language helps you write faster...

相关度: 3.8353177405018393

链接: <https://doc.rust-lang.org/book/ch00-00-introduction.html>

输入正则表达式

选择词性

搜索

☆☆☆☆

Installation

2024-05-31

介绍: The first step is to install Rust. We'll download Rust through , a command line tool for managing ...

相关度: 0.20090786228222082

链接: <https://doc.rust-lang.org/book/ch01-01-installation.html>

输入正则表达式

选择词性

搜索

☆☆☆☆

后端日志展示：

```
[GIN] 2024/06/08 - 21:45:55 | 200 | 606µs | 10.29.94.205 | GET | "/api/v1/search?q=test&page=1&limit=10"
[GIN] 2024/06/08 - 21:46:40 | 200 | 1.0925ms | 10.29.94.205 | GET | "/api/v1/search?q=test&page=1&limit=10"
time="2024-06-08T21:50:30+08:00" level=info msg="queryWords: [1]"
time="2024-06-08T21:50:30+08:00" level=info msg="queryVector:map[1:-0.9999983468443772]"
time="2024-06-08T21:50:30+08:00" level=info msg="202 results"
time="2024-06-08T21:50:30+08:00" level=debug msg=">>> scoreMap"
time="2024-06-08T21:50:30+08:00" level=debug msg="83:Doc:{83 <h2 id=\"storing-keys-with-associated-values-in-hash-maps\"><a class=\"header\" href=\"#storing-keys-with-associated-values-in-hash-maps\">Storing Keys with Associated Values in Hash Maps</a></h2> The last of our common collections is the . The type \nstores a mapping of keys of type to values of... https://doc.rust-lang.org/book/ch08-03-hash-maps.html 2024-05-31}Score:0.0001931839082104849"
time="2024-06-08T21:50:30+08:00" level=debug msg="188:Doc:{188 <h1>杨氏矩阵</h1> 杨氏矩阵引入 (Young tableau), 又名杨表, 是一种常用于表示论和舒伯特演算中\Xe7... https://oi-wiki.org/math/young-tableau/ 2024-05-31}Score:3.9090714015602266e-05"
time="2024-06-08T21:50:30+08:00" level=debug msg="198:Doc:{198 <h1>公平组合游戏</h1> 公平组合游戏经典的公平组合游戏有很多, 包括取数游戏, 31 点, 以及 Nim \Xe6\xB8... https://oi-wiki.org/math/game-theory/impartial-game/ 2024-05-31}Score:0.0019040737123493896"
time="2024-06-08T21:50:30+08:00" level=debug msg="48:Doc:{48 <h2 id=\"using-threads-to-run-code-simultaneously\"><a class=\"header\" href=\"#using-threads-to-run-code-simultaneously\">Using Threads to Run Code Simultaneously</a></h2> In most current operating systems, an executed program's code is run in a\n, and the operating syst... https://doc.rust-lang.org/book/ch16-01-threads.html 2024-05-31}Score:0.0004254136865664769"
time="2024-06-08T21:50:30+08:00" level=debug msg="388:Doc:{388 <h1>bitset</h1> bitset介绍 是标准库中的一个存储 的大小不可变容器。严格来讲, 它并不属\Xe4... https://oi-wiki.org/lang/csl/bitset/ 2024-05-31}Score:5.162520772654723e-05"
time="2024-06-08T21:50:30+08:00" level=debug msg="422:Doc:{422 <h1>出题</h1> 出题出题前的准备具备一定的水平一方面, 一个人自己出题, 很难出出难度\Xe5... https://oi-wiki.org/contest/problemsetting/ 2024-05-31}Score:8.88146634694935e-06"
time="2024-06-08T21:50:30+08:00" level=debug msg="55:Doc:{55 <h2 id=\"cargo-workspaces\"><a class=\"header\" href=\"#cargo-workspaces\">Cargo Workspaces</a></h2> In Chapter 12, we built a package that included a binary crate and a library\ncrate. As your project ... https://doc.rust-lang.org/book/ch14-03-cargo-workspaces.html 2024-05-31}Score:6.963541137311594e-06"
time="2024-06-08T21:50:30+08:00" level=debug msg="119:Doc:{119 <h1>矩阵树定理</h1> 矩阵树定理Kirchhoff 矩阵树定理 (简称矩阵树定理) 解决了一张图的生成树个数... https://oi-wiki.org/graph/matrix-tree/ 2024-05-31}Score:0.010477520794562186"
time="2024-06-08T21:50:30+08:00" level=debug msg="291:Doc:{291 <h1>Lyndon 分解</h1> Lyndon 分解定义首先我们介绍 Lyndon 分解的概念。Lyndon 串: 对于字符串, 如\Xe6\x9e... https://oi-wiki.org/string/lyndon/ 2024-05-31}Score:6.930629463264774e-06"
time="2024-06-08T21:50:30+08:00" level=debug msg="256:Doc:{256 <h1>洲阁筛</h1> 洲阁筛前置知识定义洲阁筛是一种能在亚线性时间复杂度内求出大多数积性\Xe5... https://oi-wiki.org/math/number-theory/zhou/ 2024-05-31}Score:0.0005353866616791869"
time="2024-06-08T21:50:30+08:00" level=debug msg="177:Doc:{177 <h1>划分树</h1> 划分树引入划分树是一种来解决区间第 大的一种数据结构, 其常数、理解\Xe9\x9a... https://oi-wiki.org/ds/dividing/ 2024-05-31}Score:6.444740970351023e-06"
time="2024-06-08T21:50:30+08:00" level=debug msg="69:Doc:{69 <h2 id=\"refactoring-to-improve-modularity-and-error-handling\"><a class=\"header\" href=

time="2024-06-08T22:11:06+08:00" level=debug msg="102:Doc:{102 <h2 id=\"installation\"><a class=\"header\" href=\"#installation\">Installation</a></h2> The first step is to install Rust. We'll download Rust through , a\ncommand line tool for managing ... https://doc.rust-lang.org/book/ch01-01-installation.html 2024-05-31}Score:0.20090786228222085"
time="2024-06-08T22:11:06+08:00" level=debug msg="41:Doc:{41 <h2 id=\"extensible-concurrency-with-the-sync-and-send-traits\"><a class=\"header\" href=\"#extensible-concurrency-with-the-sync-and-send-traits\">Extensible Concurrency with the <code>Sync</code> and <code>Send</code> Traits</a></h2> Interestingly, the Rust language has few concurrency features. Almost\nevery concurrency feature we\Xe2... https://doc.rust-lang.org/book/ch16-04-extensible-concurrency-sync-and-send.html 2024-05-31}Score:0.006258576034104325"
time="2024-06-08T22:11:06+08:00" level=debug msg="43:Doc:{43 <h2 id=\"characteristics-of-object-oriented-languages\"><a class=\"header\" href=\"#characteristics-of-object-oriented-languages\">Characteristics of Object-Oriented Languages</a></h2> There is no consensus in the programming community about what features a\nlanguage must have to be co... https://doc.rust-lang.org/book/ch17-01-what-is-oo.html 2024-05-31}Score:0.029741694042025146"
time="2024-06-08T22:11:06+08:00" level=debug msg="<<< scoreMap"
[GIN] 2024/06/08 - 22:11:06 | 200 | 291.097ms | :1 | POST | "/api/v1/search_by_image"
[GIN] 2024/06/08 - 22:11:12 | 200 | 652.3µs | 10.29.94.205 | GET | "/api/v1/search?q=test&page=1&limit=10"
[GIN] 2024/06/08 - 22:11:17 | 200 | 880.5µs | 10.29.94.205 | GET | "/api/v1/document?id=78"
time="2024-06-08T22:11:17+08:00" level=debug msg="Extract info for doc 78 entities: map[8:3 Chapter 5:3 one:3 two:6 's:3] hot_words: map[code:24 function:36 test:89 tests:35 's:63]"
[GIN] 2024/06/08 - 22:11:17 | 200 | 475.6573ms | 10.29.94.205 | GET | "/api/v1/extract_info?id=78"
[GIN] 2024/06/08 - 22:11:41 | 200 | 1.4874ms | 10.29.94.205 | GET | "/api/v1/document?id=65"
time="2024-06-08T22:11:42+08:00" level=debug msg="Extract info for doc 65 entities: map[Listing:3 TDD:3 one:2] hot_words: map[We:9 function:19 return:15 test:19 's:43]"
[GIN] 2024/06/08 - 22:11:42 | 200 | 256.7817ms | 10.29.94.205 | GET | "/api/v1/extract_info?id=65"
[GIN] 2024/06/08 - 22:11:59 | 200 | 520.5µs | 10.29.94.205 | GET | "/api/v1/search?q=test&page=1&limit=10"
[GIN] 2024/06/08 - 22:12:05 | 200 | 261.7µs | 10.29.94.205 | GET | "/api/v1/document?id=69"
time="2024-06-08T22:12:05+08:00" level=debug msg="Extract info for doc 69 entities: map[1:3 12:4 Listing:6 one:5 two:5] hot_words: map[code:29 error:29 function:51 value:29 's:95]"
[GIN] 2024/06/08 - 22:12:05 | 200 | 527.8517ms | 10.29.94.205 | GET | "/api/v1/extract_info?id=69"
```



```
10.29.12.98 - - [08/Jun/2024 22:10:06] "POST /extract_info HTTP/1.1" 200 -
[2024-06-08 22:11:07,734] INFO in main: photo.png
1/1 [=====] - 0s 27ms/step
10.29.12.98 - - [08/Jun/2024 22:11:07] "POST /image_to_keywords HTTP/1.1" 200 -
[2024-06-08 22:11:18,740] INFO in main: Data language: en
entities: [{'text': 'three', 'label': 'CARDINAL'}, {'text': 'Rust', 'label': 'GPE'}, {'text': 'one', 'label': 'CARDINAL'}, {'text': 'Chapter 5', 'label': 'LAW'}, {'text': 'Cargo', 'label': 'ORG'}, {'text': 's', 'label': 'NORP'}, {'text': 'two', 'label': 'CARDINAL'}, {'text': 'Listing 11-1.For', 'label': 'FAC'}, {'text': 'two', 'label': 'CARDINAL'}, {'text': '2', 'label': 'CARDINAL'}, {'text': '4', 'label': 'CARDINAL'}, {'text': 'Listing', 'label': 'GPE'}, {'text': '11-2.Cargo', 'label': 'CARDINAL'}, {'text': 'n't', 'label': 'GPE'}, {'text': 'Chapter 14', 'label': 'LAW'}, {'text': 'First', 'label': 'ORDINAL'}, {'text': 'Chapter 9', 'label': 'LAW'}, {'text': 'Listing 11-3.Run', 'label': 'WORK_OF_ART'}, {'text': '11', 'label': 'CARDINAL'}, {'text': 'Two', 'label': 'CARDINAL'}, {'text': 'first', 'label': 'ORDINAL'}, {'text': '10', 'label': 'CARDINAL'}, {'text': 'one', 'label': 'CARDINAL'}, {'text': 'one', 'label': 'CARDINAL'}, {'text': 's', 'label': 'NORP'}, {'text': 'Boolean', 'label': 'GPE'}, {'text': 'Chapter 5', 'label': 'LAW'}, {'text': 'Listing 11-5', 'label': 'FAC'}, {'text': 'Boolean', 'label': 'GPE'}, {'text': '11-6', 'label': 'CARDINAL'}, {'text': '8', 'label': 'CARDINAL'}, {'text': '7', 'label': 'CARDINAL'}, {'text': '5', 'label': 'CARDINAL'}, {'text': '1.Note', 'label': 'CARDINAL'}, {'text': 'Chapter 7', 'label': 'LAW'}, {'text': 'two', 'label': 'CARDINAL'}, {'text': 'Two', 'label': 'CARDINAL'}, {'text': '8', 'label': 'CARDINAL'}, {'text': '5', 'label': 'CARDINAL'}, {'text': '8', 'label': 'CARDINAL'}, {'text': 'less than', 'label': 'CARDINAL'}, {'text': 'two', 'label': 'CARDINAL'}, {'text': 'two', 'label': 'CARDINAL'}, {'text': 's', 'label': 'NORP'}, {'text': 'two', 'label': 'CARDINAL'}, {'text': 'the day', 'label': 'DATE'}, {'text': 'the week', 'label': 'DATE'}, {'text': 'Chapter 5', 'label': 'LAW'}, {'text': 'Chapter 8', 'label': 'LAW'}, {'text': 'Chapter 9', 'label': 'LAW'}, {'text': 'between 1 and 100', 'label': 'CARDINAL'}, {'text': '11-8', 'label': 'CARDINAL'}, {'text': 'greater than 100', 'label': 'CARDINAL'}, {'text': 'Listing 11-8', 'label': 'FAC'}, {'text': 'Listing 11-9', 'label': 'FAC'}, {'text': 'Listing 11-1', 'label': 'FAC'}]
[2024-06-08 22:11:19,194] DEBUG in main: {"entities": {"two": 6, "one": 3, "Chapter 5": 3, "\u2019s": 3, "8": 3}, "hot_words": {"test": 89, "\u2019": 63, "function": 36, "tests": 35, "code": 24}}
```

```
检查 STL 容器中元素的对象, \xe5... https://oi-wiki.org/lang/csl/iterator/ 2024-05-31}Score:66.18760740784518"
time="2024-06-27T17:33:25+08:00" level=debug msg="397:Doc:{397 <h1>C++ 语法基础</h1> C++ 语法基础代码框架如果你不想深究背后的原理, 初学时可以直接将这个「... https://oi-wiki.org/lang/basic/ 2024-05-31}Score:281.1842896027342"
time="2024-06-27T17:33:25+08:00" level=debug msg="41:Doc:{41 <h2 id=\"extensible-concurrency-with-the-sync-and-send-traits\"><a class=\"header\" href=\"#extensible-concurrency-with-the-sync-and-send-traits\">Extensible Concurrency with the <code>Sync</code> and <code>Send</code> Traits</a></h2> Interestingly, the Rust language has few concurrency features. Almost\never a concurrency feature we\xe2... https://doc.rust-lang.org/book/ch16-04-extensible-concurrency-sync-and-send.html 2024-05-31}Score:67.41524672206333"
time="2024-06-27T17:33:25+08:00" level=debug msg="5:Doc:{5 <h2 id=\"what-is-ownership\"><a class=\"header\" href=\"#what-is-ownership\">What Is Ownership?</a></h2> is a set of rules that govern how a Rust program manages memory.\nAll programs have to manage the wa... https://doc.rust-lang.org/book/ch04-01-what-is-ownership.html 2024-05-31}Score:5.668672065026064"
time="2024-06-27T17:33:25+08:00" level=debug msg="407:Doc:{407 <h1>通用</h1> 通用本页面介绍 Testlib checker/interactor/validator 的一些通用状态/对象/函数、一... https://oi-wiki.org/tools/testlib/general/ 2024-05-31}Score:15.246285156104527"
time="2024-06-27T17:33:25+08:00" level=debug msg="110:Doc:{110 <h1>拆点</h1> 拆点拆点是一种图论建模思想, 常用于, 用来处理的问题, 也常用于。结... https://oi-wiki.org/graph/node/ 2024-05-31}Score:13.584758032588448"
time="2024-06-27T17:33:25+08:00" level=debug msg="192:Doc:{192 <h1>牛顿迭代法</h1> 牛顿迭代法引入本文介绍如何用牛顿迭代法 (Newton's method for finding roots) 求\xe6... https://oi-wiki.org/math/numerical/newton/ 2024-05-31}Score:63.78599976079868"
time="2024-06-27T17:33:25+08:00" level=debug msg="360:Doc:{360 <h1>Java 进阶</h1> Java 进阶以下内容均基于 Java JDK 8 版本编写, 不排除在更高版本中有部分改\xe5\x8a... https://oi-wiki.org/lang/java-pro/ 2024-05-31}Score:1256.354467227266"
time="2024-06-27T17:33:25+08:00" level=debug msg="<<< scoreMap"
[GIN] 2024/06/27 - 17:33:25 | 200 | 2m6s | 10.29.23.17 | GET | "/api/v1/search?q=%E4%BA%A4%E4%BA%92%E9%A2%98%E3%80%8AP5473[NOI2019]I%20%E5%90%9B%E7%9A%84%E6%8E%A2%E9%99%A9%E3%80%8B&page=1&limit=10"
[GIN] 2024/06/27 - 17:36:15 | 200 | 546.7µs | 10.29.94.205 | GET | "/api/v1/search?q=%E4%BA%A4%E4%BA%92%E9%A2%98%E3%80%8AP5473[NOI2019]I+%E5%90%9B%E7%9A%84%E6%8E%A2%E9%99%A9%E3%80%8B&page=1&limit=10"
[GIN] 2024/06/27 - 17:36:41 | 200 | 2.1667ms | 10.29.94.205 | GET | "/api/v1/document?id=428"
time="2024-06-27T17:36:41+08:00" level=debug msg="Extract info for doc 428 entities: map[3k:2 NOI:2] hot_words: map[I:2 NOI:2 UOJ:4 bfs:6 的:2]"
[GIN] 2024/06/27 - 17:36:41 | 200 | 136.3396ms | 10.29.94.205 | GET | "/api/v1/extract_info?id=428"
time="2024-06-27T17:37:10+08:00" level=debug msg="Extract info for doc 428 pattern: .*OI word_class: ORG words: [IOI NOI NOI]"
[GIN] 2024/06/27 - 17:37:10 | 200 | 223.1369ms | 10.29.94.205 | GET | "/api/v1/extract_info_regex?id=428&pattern=.*OI&word_class=ORG"
```

# 详细设计

# 数据爬取

本次实验要求不少于 100 篇文档, 所以我们结合自身情况爬取了比较常用的全中文的Oiwiki和最近在学习的全英文的The Rust Programming Language, 最终爬取文章数量为中文 440 篇, 英文 104 篇。



爬虫框架选取了我们最为熟悉的 Scrapy，使用该框架可以快速爬取网页内容，并且可以方便的进行数据处理。

## OiWiki 数据爬取

对于 OiWiki，我们首先爬取文章列表：

```
def parse(self, response):
    sections = response.xpath(
        "//li[@class='md-nav__item']/a[@class='md-nav__link']"
    )
    hrefs = sections.xpath("@href").getall()
    texts = sections.xpath("text()").getall()
    texts = [t.strip() for t in texts]

    for href, section in zip(hrefs, texts):
        url = response.urljoin(href)
        yield scrapy.Request(
            url=url,
            callback=self.parse_section,
            cb_kwargs={"section": section},
        )
```

然后爬取每篇文章内容：

```
def parse_section(self, response, section="Unknown"):
    content = response.xpath(
        '//div[@class="md-content"]//blockquote[1]/preceding-sibling::*[not(self::a)]'
    ).getall()
    keywords = response.xpath(
        '//div[@class="md-content"]//*[self::h1 or self::h2 or self::h3 or self::h4 or self::li]'
    ).getall()

    self.id = self.id + 1
    yield {
        "id": str(self.id),
        "title": content[0],
        "content": "".join(para for para in content),
        "keywords": "".join(para for para in keywords),
        "url": response.url,
        "date": datetime.date.today().strftime("%Y-%m-%d"),
    }
```

其中，我们将全文内容作为文章内容用于前端展示，文章中的所有文本内容作为关键字用于索引和检索。

## The Rust Programming Language 数据爬取

对于 The Rust Programming Language，我们同样地也是先爬取文章所有章节，然后再爬取每个章节内的详细内容：

```
def parse(self, response):
    chapters = response.xpath(
        '//ol[@class="chapter"]//li[@class="chapter-item expanded " or @class="chapter-item expi
    )
    hrefs = chapters.xpath("@href").getall()
    texts = chapters.xpath("text()").getall()

    for href, chapter in zip(hrefs, texts):
        url = response.urljoin(href)
        yield scrapy.Request(
            url=url,
            callback=self.parse_chapter,
            cb_kwargs={"chapter": chapter},
        )

def parse_chapter(self, response, chapter="Unknown"):
    content = response.xpath("//main/*")
    keywords = content.xpath("text()").getall()
    content = content.getall()

    self.id = self.id + 1
    yield {
        "id": str(self.id),
        "title": content[0],
        "content": "".join(p for p in content),
        "keywords": "".join(p for p in keywords),
        "url": response.url,
        "date": datetime.date.today().strftime("%Y-%m-%d"),
    }
```

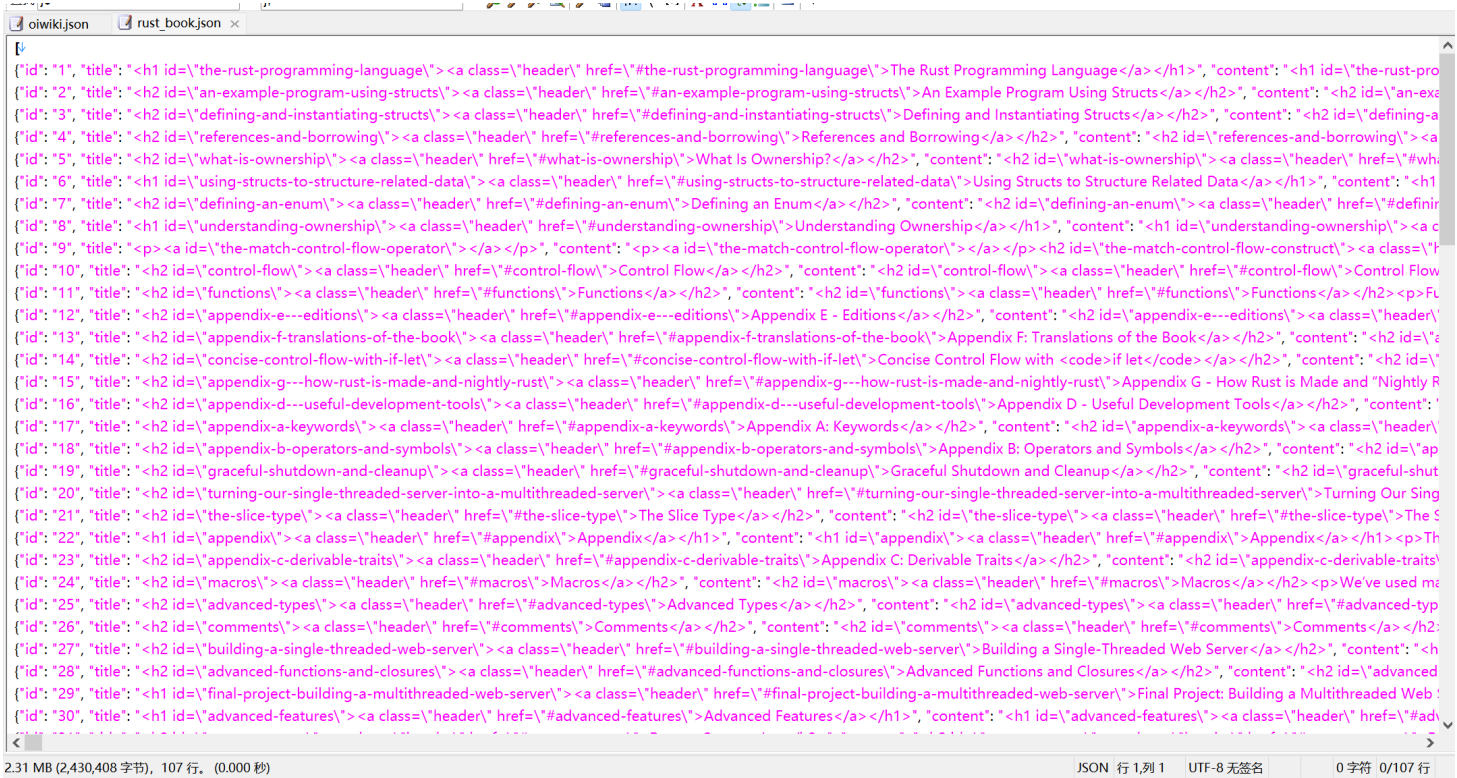
## 数据存储

为了方便后续索引和检索，我们将所有爬取到的数据都存储成 json 文件，每一条的数据的形式如下，保证所有要求的必要信息都会被存储：

```
{
  "id": "1",
  "title": "The Rust Programming Language",
  "content": "The Rust Programming Language",
  "keywords": "The Rust Programming Language",
  "url": "https://doc.rust-lang.org/book/",
  "date": "2022-04-27"
}
```

最终的爬取结果如下：





# 后端设计与实现

因为 Go 在各种测试中表现出了优秀的性能水平，所以本次实验后端我使用 Go 语言进行开发，框架使用了 Gin 这一高性能且比较主流的 Web 框架，本部分将介绍本项目的后端接口设计以及相关逻辑实现。

不过 Go 令人最为诟病的一点就是其 err 的判断机制，几乎每一次函数调用都要判断函数返回的 err 是否需要处理，所以考虑到报告的篇幅长度，我在后续展示 Go 代码时都将删去 err 的判断部分，以便带来更加良好的阅读体验。

## 接口展示

首先展示一下我们此次实验完成的所有接口（包括作业二和作业三）

查询接口



GET	/document 获取文档详细信息	▼
GET	/search 分页查询	▼
POST	/search_by_image 上传图片查询	▼

反馈接口



POST	/entity_feedback 实体反馈	▼
POST	/extract_info_regex_feedback 正则提取反馈 (正则+词性)	▼
POST	/feedback 结果反馈	▼
POST	/hotword_feedback 热词反馈	▼

提取接口



GET	/extract_info 提取关键信息	▼
GET	/extract_info_regex 提取关键信息	▼

```
r.GET("/swagger/*any", ginSwagger.WrapHandler(swaggerFiles.Handler))

v1 := r.Group("/api/v1")
{
    // Search with keywords
    v1.GET("/search", controller.Search)
    // Fetch SearchResult content details
    v1.GET("/document", controller.GetDocument)
    // Search by image
    v1.POST("/search_by_image", controller.SearchByImage)

    // Entities and hot words
    v1.GET("/extract_info", controller.ExtractInfo)
    // Entity and hot word feedback
    v1.POST("/extract_info_regex", controller.ExtractInfoRegex)

    // Feedback
    v1.POST("/feedback", controller.Feedback)
    // Entity Feedback
    v1.POST("/entity_feedback", controller.EntityFeedback)
    // Hotword Feedback
    v1.POST("/hotword_feedback", controller.HotwordFeedback)
    // Regex Feedback
    v1.POST("/extract_info_regex_feedback", controller.ExtractInfoRegexFeedback)
}
```

## 默认信息提取

默认每篇文章我们会提取出该篇文章的热词和部分实体进行展示，在执行过程中我收到请求后在后端查询该篇文章详细内容，然后将其作为参数发送给 Python 接口，具体提取逻辑见[热词和实体抽取](#)。

后端代码如下：

```

func ExtractInfo(doc_id string) (model.DocumentAbstract, error) {
    var result model.DocumentAbstract

    doc, ok := idDocMap[doc_id]
    if !ok {
        log.Error("Error getting doc ", doc_id)
    }
    data := map[string]string{"text": doc.Keywords, "language": doc.Lang.String()}
    jsonData, _ := json.Marshal(data)

    resp, err := http.Post(model.PYTHON_SERVER_URL+"/extract_info", "application/json", bytes.NewReader(jsonData))
    if err != nil {
        return model.DocumentAbstract{}, err
    }
    defer resp.Body.Close()

    json.NewDecoder(resp.Body).Decode(&result)

    if resp.StatusCode != http.StatusOK {
        body, err := ioutil.ReadAll(resp.Body)
        if err != nil {
            log.Error(err.Error())
            return model.DocumentAbstract{}, err
        }
        return model.DocumentAbstract{}, errors.New(string(body))
    }

    log.Debug("Extract info for doc ", doc_id, " entities: ", result.Entities, " hot_words: ", result.HotWords)

    return result, nil
}

```

## 正则和信息点信息提取

正则和信息点抽取主要是我们可以让用户自定义**正则表达式**和想要抽取的**信息点**（如：人名、法律、序数词、日期、量词、地理位置、产品名以及组织名等），后端接收到参数后也是先查询该篇文章详细内容，然后将其作为参数发送给 Python 接口，具体提取逻辑见[正则和指定信息点抽取](#)。

后端代码如下：



```

func ExtractInfoRegex(doc_id, pattern, word_class string) (model.DocumentExtractRegex, error) {
    var result model.DocumentExtractRegex

    doc, ok := idDocMap[doc_id]
    if !ok {
        log.Error("Error getting doc ", doc_id)
    }
    data := map[string]string{"text": doc.Keywords, "language": doc.Lang.String(), "pattern": pattern, "word_class": word_class}
    jsonData, _ := json.Marshal(data)

    resp, err := http.Post(model.PYTHON_SERVER_URL+"/extract_info_regex", "application/json", bytes.NewReader(jsonData))
    if err != nil {
        return model.DocumentExtractRegex{}, err
    }
    defer resp.Body.Close()

    json.NewDecoder(resp.Body).Decode(&result)

    if resp.StatusCode != http.StatusOK {
        body, err := ioutil.ReadAll(resp.Body)
        if err != nil {
            log.Error(err.Error())
            return model.DocumentExtractRegex{}, err
        }
        return model.DocumentExtractRegex{}, errors.New(string(body))
    }

    log.Debug("Extract info for doc ", doc_id, " pattern: ", pattern, " word_class: ", word_class)

    return result, nil
}

```

## 缓存优化

为了提高系统性能，我额外增加了缓存优化（第二次作业也有涉及，但是没有详细说明），缓存使用 LRU 策略，保证高命中率的同时也可以将占用内存限制在有限大小内，具体代码如下：

```

type Cache struct {
    Mu      sync.Mutex
    Cache    map[string]*list.Element
    evictList *list.List
    capacity int
}

type Entry struct {
    key    string
    value []model.SearchResult
}

func NewCache(capacity int) *Cache {
    return &Cache{
        Cache:    make(map[string]*list.Element),
        evictList: list.New(),
        capacity: capacity,
    }
}

func (c *Cache) Get(key string) ([]model.SearchResult, bool) {
    c.Mu.Lock()
    defer c.Mu.Unlock()

    if ent, ok := c.Cache[key]; ok {
        c.evictList.MoveToFront(ent)
        return ent.Value.(*Entry).value, true
    }

    return nil, false
}

func (c *Cache) Set(key string, val []model.SearchResult) {
    c.Mu.Lock()
    defer c.Mu.Unlock()

    if ent, ok := c.Cache[key]; ok {
        c.evictList.MoveToFront(ent)
        ent.Value.(*Entry).value = val
        return
    }

    if c.evictList.Len() >= c.capacity {

```

```

        ent := c.evictList.Back()
        if ent != nil {
            c.removeElement(ent)
        }
    }

    ent := &Entry{key: key, value: val}
    element := c.evictList.PushFront(ent)
    c.Cache[key] = element
}

func (c *Cache) removeElement(e *list.Element) {
    c.evictList.Remove(e)
    kv := e.Value.(*Entry)
    delete(c.Cache, kv.key)
}

```

搜索功能也是本系统必不可少的一部分，但由于篇幅限制就不再赘述，详情请见[实验二报告](#)

## 前端设计与实现

部分在实验二中已有体现，主要集中在实验三的部分。



## Knowledge Acquisition


**关键词搜索**

**图片检索** 

### 关键词、图片搜索

- 关键词搜索**  
rust test organization run

**关键词搜索**  
mask comic\_book book\_jacket

**图片检索**   
photo.png

- **组件：** `<v-text-field>` 和 `<v-btn>`
- **布局：** 关键词输入框和搜索按钮在同一行显示，使用 Vuetify 的布局系统优化空间利用。
- **接口函数：** `searchByKeyword()`, `searchByImage()`

```

searchByKeyword() {
  const params = { q: this.searchText };
  axios
    .get(`api/v1/search`, { params })
    .then((response) => {
      this.searchResults = response.data;
    })
    .catch((error) => {
      console.error("Error during keyword search:", error);
    });
},

```

```

  searchByImage() {
    if (!this.imageFile) {
      alert("Please upload an image.");
      return;
    }
    const formData = new FormData();
    formData.append("image", this.imageFile);
    axios
      .post(`api/v1/search_by_image`, formData, {
        headers: { "Content-Type": "multipart/form-data" },
      })
      .then((response) => {
        this.searchResults = response.data.results;
        this.searchText = response.data.keywords;
      })
      .catch((error) => {
        console.error("Error during image search:", error);
      });
  },

```

Knowledge Acquisition

关键词搜索

输入关键字

mask comic\_book book\_jacket

搜索

图片检索

点击上传图片

photo.png

搜索

Introduction

2024-05-31

介绍: Welcome to , an introductory book about Rust. The Rust programming language helps you write faster...

相关度: 3.8353177405018393

链接: <https://doc.rust-lang.org/book/ch00-00-introduction.html>

输入正则表达式

选择词性

搜索

☆☆☆☆☆

Installation

2024-05-31

介绍: The first step is to install Rust. We'll download Rust through , a command line tool for managing ...

相关度: 0.20090786228222082

链接: <https://doc.rust-lang.org/book/ch01-01-installation.html>

输入正则表达式

选择词性

搜索

☆☆☆☆☆

- 组件： <v-card>
- 功能：
  - 动态展示搜索结果，每个结果为一个卡片，展示包括文档标题、相关度、介绍信息、链接（可点击跳转）， 右下角有对每个搜索结果的评分反馈，用户可以进行评分。
  - 用户可点击每张卡片的标题， 点击标题将根据该结果的文章 id，从后端调取本结果的详细信息。（再次点击即可隐藏）
  - 详细信息包括：整篇文章的所有内容、文章关键字、文章实体表格、文章热词表格。

## How to Write Tests

2024-05-31

介绍: Tests are Rust functions that verify that the non-test code is functioning in the expected manner. T...

相关度: 3162.068716442747

链接: <https://doc.rust-lang.org/book/ch11-01-writing-tests.html>

### How to Write Tests

Tests are Rust functions that verify that the non-test code is functioning in the expected manner. The bodies of test functions typically perform these three actions:

1. Set up any needed data or state.
2. Run the code you want to test.
3. Assert the results are what you expect.

Let's look at the features Rust provides specifically for writing tests that take these actions, which include the `test` attribute, a few macros, and the `should_panic` attribute.

### The Anatomy of a Test Function

At its simplest, a test in Rust is a function that's annotated with the `test` attribute. Attributes are metadata about pieces of Rust code; one example is the `derive` attribute we used with structs in Chapter 5. To change a function into a test function, add `#[test]` on the line before `fn`. When you run your tests with the `cargo test` command, Rust builds a test runner binary that runs the annotated functions and reports on whether each test function passes or fails.

Whenever we make a new library project with Cargo, a test module with a test function in it is automatically generated for us. This module gives you a template for writing your tests so you don't have to look up the exact structure and syntax every time you start a new project. You can add as many additional test functions and as many test modules as you want!

We'll explore some aspects of how tests work by experimenting with the template test before we actually test any code. Then we'll write some real-world tests that call some code that we've written and assert that its behavior is correct.

Let's create a new library project called `adder` that will add two numbers:

```
$ cargo new adder --lib
Created library `adder` project
$ cd adder
```

The contents of the `src/lib.rs` file in your `adder` library should look like Listing 11-1.

Filename: `src/lib.rs`

```
pub fn add(left: usize, right: usize) -> usize {
    left + right
}

#[cfg(test)]
mod tests {
    use super::add;

    #[test]
    fn it_adds() {
        let sum = add(2, 2);
        assert_eq!(sum, 4);
    }
}
```



### Accessing Fields of a Union

The final action that works only with `unsafe` is accessing fields of a `union`. A `union` is similar to a `struct`, but only one declared field is used in a particular instance at one time. Unions are primarily used to interface with unions in C code. Accessing union fields is unsafe because Rust can't guarantee the type of the data currently being stored in the union instance. You can learn more about unions in [the Rust Reference](#).

### When to Use Unsafe Code

Using `unsafe` to take one of the five actions (superpowers) just discussed isn't wrong or even frowned upon. But it is trickier to get `unsafe` code correct because the compiler can't help uphold memory safety. When you have a reason to use `unsafe` code, you can do so, and having the explicit `unsafe` annotation makes it easier to track down the source of problems when they occur.

关键字: All the code we've discussed so far has had Rust's...

语言: 英文

实体	频率	反馈
ABI	3	☆☆☆☆☆
Listing	3	☆☆☆☆☆
five	5	☆☆☆☆☆
one	4	☆☆☆☆☆
two	5	☆☆☆☆☆
热词	频率	反馈
Rust	36	☆☆☆☆☆
code	41	☆☆☆☆☆
raw	27	☆☆☆☆☆
unsafe	45	☆☆☆☆☆
,	95	☆☆☆☆☆

☆☆☆☆☆

- **布局：**结果以列表形式排列，通过 Vuetify 的响应式布局确保在不同设备上的显示效果。
- **接口函数：**获取文档详细信息和实体热词表格。

```

toggleDetail(id) {
  // 检查detailMap中是否存在该id且其visible属性为true
  if (this.detailMap[id] && this.detailMap[id].visible) {
    // 如果已经可见, 则设置为不可见
    this.$set(this.detailMap[id], "visible", false);
  }
  // 检查detailMap中是否存在该id且其visible属性为false
  else if (this.detailMap[id] && !this.detailMap[id].visible) {
    // 如果不可见, 则设置为可见
    this.$set(this.detailMap[id], "visible", true);
  } else {
    // 如果detailMap中没有该id的信息, 通过axios请求获取数据
    axios
      .all([
        axios.get(`api/v1/document`, { params: { id } }), // 请求文档详情
        axios.get(`api/v1/extract_info`, { params: { id } }), // 请求提取信息
      ])
      .then(
        axios.spread((DocRes, infoRes) => {
          // 处理响应数据

          const entitiesWithScore = Object.entries(infoRes.data.entities).reduce((acc, [key, value]) => {
            // 初始化每个实体的评分为0
            acc[key] = { value, score: 0 };
            return acc;
          }, {});

          const hotWordsWithScore = Object.entries(infoRes.data.hot_words).reduce((acc, [key, value]) => {
            // 初始化每个热词的评分为0
            acc[key] = { value, score: 0 };
            return acc;
          }, {});

          // 设置detailMap以包含获取的详情数据
          this.$set(this.detailMap, id, {
            visible: true, // 设置为可见
            content: DocRes.data.content, // 文档内容
            keywords: DocRes.data.keywords, // 关键词
            Lang: DocRes.data.Lang, // 语言
            entities: infoRes.data.entities, // 实体
            hot_words: infoRes.data.hot_words, // 热词
          });
        })
      )
    )
  }
}

```

```
.catch((error) => {  
    // 处理请求错误  
    console.error("Error fetching Document details:", error);  
});  
}  
}
```

## 正则和指定信息搜索

搜索 OI 作为后缀的组织名：

### 交互题

2024-05-31

介绍：交互题上个世纪的 IOI 就已涉及交互题。虽然交互题近年来没有在省选以下...  
相关度：843.489752205303  
链接：<https://oi-wiki.org/contest/interaction/>

输入正则表达式

.OI

选择词性

组织名

搜索

关键词	评分
IOI	☆☆☆☆
NOI	☆☆☆☆

☆☆☆☆☆

搜索所有数词：

### Programming a Guessing Game

2024-05-31

介绍：Let's jump into Rust by working through a hands-on project together! This chapter introduces you t...  
相关度：389893.77405271505  
链接：<https://doc.rust-lang.org/book/ch02-00-guessing-game-tutorial.html>

输入正则表达式

\*

选择词性

数词

搜索

关键词	评分
first	★★★★☆
second	★★★★☆
third	★★★★☆
First	☆☆☆☆

☆☆☆☆☆

- 功能：
  - 正则表达式搜索和词性选择：在每个搜索结果的底部添加了一个正则表达式输入框和一个下拉菜单用于选择词性。用户可以输入正则表达式并从下拉菜单中选择词性（如人名、法律、序数词、日期、量词、地理位置、产品名以及组织名等），这些词性对应后端的特定标记（如 PERSON, LAW 等）。

对应关系如下：

```
wordClasses: [
  { text: "人物", value: "PERSON" },
  { text: "法律", value: "LAW" },
  { text: "数词", value: "ORDINAL" },
  { text: "时间", value: "DATE" },
  { text: "量词", value: "QUANTITY" },
  { text: "地理位置", value: "GPE" },
  { text: "产品", value: "PRODUCT" },
  { text: "组织名", value: "ORG" },
],
```

- **动态请求与展示搜索结果**：用户填写完正则表达式和选择词性后，点击搜索按钮将发送请求到后端。后端返回匹配的词汇列表，前端接收并去重显示这些词汇。
- **展示结果和评分**：搜索得到的结果以表格形式展示，显示关键词和用户可进行评分的部分。
- **布局**：搜索框和下拉菜单并排布置，搜索结果在用户点击搜索后动态生成，结果以表格形式呈现在搜索框和下拉菜单下方。
- **接口函数**：`searchByRegex`：根据用户输入的正则表达式和选择的词性，向后端发送 GET 请求，获取匹配的关键词列表。

```

searchByRegex(id) {
  if (!this.regexMap[id]) {
    console.error("Regex search parameters not initialized.");
    return;
  }
  const params = {
    id,
    pattern: this.regexMap[id].pattern,
    word_class: this.regexMap[id].wordClass,
  };
  axios
    .get("api/v1/extract_info_regex", { params })
    .then((response) => {
      // 使用 Set 进行去重
      const uniqueWords = [...new Set(response.data.words)];
      // 将去重后的关键词转化为需要的格式
      const wordsWithScore = uniqueWords.reduce((acc, word) => {
        acc[word] = { score: 0 }; // 默认评分为0
        return acc;
      }, {});
      this.$set(this.regexMap[id], "results", wordsWithScore);
    })
    .catch((error) => {
      console.error("Error during regex search:", error);
    });
},

```

## 搜索结果反馈

### 实体反馈

- 组件: <v-rating>

```

<v-rating
  dense
  hover
  small
  v-model="detailMap[result.Doc.id].entities[key].score"
  @input="handleEntityFeedback(result.Doc.id, key, detailMap[result.Doc.id].entities[key])">
</v-rating>

```

实体	频率	反馈
8	3	★ ★ ★ ☆ ☆
Chapter 5	3	★ ★ ☆ ☆ ☆
one	3	★ ★ ★ ★ ★
two	6	☆ ☆ ☆ ☆ ☆
's	3	★ ★ ★ ☆ ☆

- **功能：**允许用户对搜索结果中每个实体的准确性进行评分。
- **属性：**
  - `dense` 和 `small` 使评分组件更紧凑、适合放置在搜索结果卡片中。
  - `hover` 允许用户在鼠标悬停时预览评分效果。
  - `v-model` 绑定到 `detailMap[result.Doc.id].entities[key].score`，实现数据的双向绑定。
  - `@input` 事件处理函数 `handleEntityFeedback` 发送用户的评分到后端。
- **布局：**每个搜索结果的详细信息区域均设有实体反馈评分组件，与实体信息并排展示。
- **接口函数：**

```
handleEntityFeedback(resultId, item, score) {
  const payload = {
    item,
    resultId,
    score
  };
  axios.post(`api/v1/entity_feedback`, payload)
    .then(response => {
      console.log("Entity Feedback sent successfully", response);
    })
    .catch(error => {
      console.error("Error sending entity feedback", error);
    });
},
```

## 热词反馈

- **组件：**`<v-rating>`

```

<v-rating
  dense
  hover
  small
  v-model="detailMap[result.Doc.id].hot_words[key].score"
  @input="handleHotwordFeedback(result.Doc.id, key, detailMap[result.Doc.id].hot_words[ke
></v-rating>

```

热词	频率	反馈
code	24	☆☆☆☆☆
function	36	★★★★☆
test	89	★★★★☆
tests	35	★★★★☆
,	63	★★★★☆

- **功能：**允许用户对搜索结果中的热词进行评分。
- **属性：**
  - `dense` 和 `small` 使评分组件更紧凑、适合放置在搜索结果卡片中。
  - `hover` 允许用户在鼠标悬停时预览评分效果。
  - `v-model` 绑定到 `detailMap[result.Doc.id].hot_words[key].score`，实现数据的双向绑定。
  - `@input` 事件处理函数 `handleEntityFeedback` 发送用户的评分到后端。
- **布局：**与实体反馈类似，热词反馈组件与对应的热词信息并排展示。
- **接口函数：**



```

handleHotwordFeedback(resultId, item, score) {
  const payload = {
    item,
    resultId,
    score
  };
  axios.post(`api/v1/hotword_feedback`, payload)
    .then(response => {
      console.log("Hotword Feedback sent successfully", response);
    })
    .catch(error => {
      console.error("Error sending hotword feedback", error);
    });
},

```

## 信息抽取反馈

- **组件：** <v-rating>

```

<v-rating
  dense
  hover
  small
  v-model="regexMap[result.Doc.id].results[word].score"
  @input="handleRegexFeedback(result.Doc.id, word, regexMap[result.Doc.id].results[word].
></v-rating>

```

## Programming a Guessing Game

2024-05-31

介绍: Let's jump into Rust by working through a hands-on project together! This chapter introduces you t...

相关度: 389893.77405271505

链接: <https://doc.rust-lang.org/book/ch02-00-guessing-game-tutorial.html>

选择词性  
 数词

关键词	评分
first	★★★★☆
second	★★★★☆
third	★★★★☆
First	★☆☆☆☆

☆☆☆☆☆

- **功能：** 允许用户对正则表达式搜索结果中的每个关键词进行评分。
- **属性：**
  - `dense` 和 `small` 使评分组件更紧凑，适合放置在搜索结果卡片中。
  - `hover` 允许用户在鼠标悬停时预览评分效果。

- `v-model` 绑定到 `regexMap[result.Doc.id].results[word].score`，实现数据的双向绑定，保持界面和数据状态的同步。
- `@input` 事件处理函数 `handleRegexFeedback` 用于当用户修改评分时发送这一改变到后端。
- **布局**：评分组件在每个搜索结果的正则表达式搜索结果表格中，与关键词并排显示。
- **接口函数**：

```
handleRegexFeedback(resultId, word, score) {
  const payload = {
    item: word,
    resultId: resultId,
    score: score
  };
  axios.post(`api/v1/regex_feedback`, payload)
    .then(response => {
      console.log("Regex Feedback sent successfully", response);
    })
    .catch(error => {
      console.error("Error sending regex feedback", error);
    });
}
```

## 整体准确率评价反馈

- **组件**：<v-rating>

```
<v-rating
  dense
  hover
  v-model="result.Score"
  @input="handleOverallFeedback(result.Doc.id, result.Score)"
></v-rating>
```

### Controlling How Tests Are Run

2024-05-31

介绍: Just as compiles your code and then runs the resulting binary, compiles your code in test mode and...

相关度: 1509.4414788276588

链接: <https://doc.rust-lang.org/book/ch11-02-running-tests.html>



- **功能**：提供对整个搜索结果的整体准确率满意度评价。
- **属性**：
  - `hover` 和 `dense` 属性同上。

- v-model 绑定到 result.Score。
- @input 通过 handleOverallFeedback 方法发送整体评分数据到后端。
- **布局**：整体评价组件位于搜索结果卡片的底部，方便用户在查看完信息后给出整体准确率评价。
- **接口函数**：

```
handleOverallFeedback(resultId, Score) {  
  const payload = {  
    resultId,  
    Score  
  };  
  axios.post(`api/v1/feedback`, payload)  
    .then(response => {  
      console.log("Overall Feedback sent successfully", response);  
    })  
    .catch(error => {  
      console.error("Error sending overall feedback", error);  
    });  
}
```

## 信息抽取服务

信息检索服务是本次实验的核心内容，在实验中，我完成了三种情况的信息提取——用户自定义**正则表达式和信息点**提取、部分实体词以及高频词提取。其中用户自定义的信息点总体上有**人名、法律、序数词、日期、量词、地理位置、产品名以及组织名这七种**。在实际实验中，这部分的实际逻辑主要由 python 服务完成，go 服务中处理处理请求数据，然后调用 python 服务进行分词处理和提取。

## 热词和实体抽取

该部分是先接受 /extract\_info 路径的请求，然后调用 entity\_detect 函数通过不同的处理方式提取出所有实体，然后选取频率前五的实体词作为热词返回，具体的抽取逻辑 entity\_dectect 见[实体词检测和抽取](#)，下面展示路由的处理和热词提取的处理代码：

```

@app.route("/extract_info", methods=["POST"])
def extract_info():
    data = request.get_json()
    text = data.get("text")
    language = data.get("language")
    log.info("Data language: " + language)

    if not text or not language:
        return "Invalid request: no text or no language", 400

    if language not in ["en", "cn"]:
        return "Unsupported language: " + language, 400

    # Entity detection
    entities = entity_detection.entity_detect(text, language)

    # Extract hot words
    stop_words = (
        set(stopwords.words("english")) if language == "en" else set()
    )
    word_tokens = word_tokenize(text)
    words = [
        w for w in word_tokens if not w in stop_words and not w in punctuation
    ]

    hot_words = dict(Counter(words).most_common(5))

    entities = [e["text"] for e in entities]
    entities = dict(Counter(entities).most_common(5))
    entities = {k: v for k, v in entities.items() if v > 1}

    jsonResponse = json.dumps({"entities": entities, "hot_words": hot_words})
    log.debug(jsonResponse)
    return jsonResponse

```

## 正则和指定信息点抽取

在本次实验中我们提供了**人名、法律、序数词、日期、量词、地理位置、产品名、组织名**这七种信息点可以选择抽取，抽取的过程如下：

1. 检测实体并提取；
2. 对每个词语判断是否匹配正则表达式和符合规定信息点；

### 3. 组装并返回结果

具体代码如下：

```
@app.route("/extract_info_regex", methods=["POST"])
def extract_info_regex():
    data = request.get_json()
    text = data.get("text")
    pattern = data.get("pattern")
    language = data.get("language")
    word_class = data.get("word_class")

    if not text or not language:
        return "Invalid request: no text or no language", 400

    if language not in ["en", "cn"]:
        return "Unsupported language: " + language, 400

    # Entity detection
    entities = entity_detection.entity_detect(text, language)

    # Extract words with regex
    words = []
    entities = [{"text": item["text"], "label": item["label"]} for item in entities]
    for entity in entities:
        if word_class == entity.get("label") and bool(re.fullmatch(pattern=pattern, string=entity
            words.append(entity.get('text'))

    jsonResponse = json.dumps({"words": words})
    log.debug(jsonResponse)
    return jsonResponse
```

## 实体词检测和提取

如何从文章中抽取出实体词，这是此次实验的核心问题，由于有中英文两种不同的文本，所以在实验中我也采用了两种不同的方法分别进行处理：

- 对于中文文本：jieba 处理库对于中文文本已经有了非常好的处理和检测能力，它对于每个词都能准确地分割并识别其词性，所以我使用 jieba 库对中文文本进行处理；
- 对于英文文本：对于英文文本我使用了 spaCy 框架，该库可以使用模型对文本进行分词和词性标注，提高了结果的准确率。在模型方面，我选择了 en\_core\_web\_sm 模型。

具体我的处理代码如下：

```
def entity_detect(text: str, language: str) -> str:
    entities = ""
    if language == "en":
        entities = en_entity_detect(text)
    elif language == "cn":
        entities = cn_entity_detect(text)
    print(f"entities: {entities}")
    return entities

def en_entity_detect(text: str) -> str:
    entities = []
    doc = nlp_en(text)
    # Extract entities
    for entity in doc.ents:
        entities.append(
            {
                "text": entity.text,
                "label": entity.label_,
            }
        )

    return entities

def cn_entity_detect(text: str) -> str:
    entities = []
    words = pseg.cut(text)
    # Extract entities
    for word, flag in words:
        entities.append({"text": word, "label": flag})

    return entities
```

## 多媒体信息抽取服务

想要实现多媒体的信息抽取，首先就要解决从多媒体中提取出关键词的问题，我在这里沿用了实验二的代码进行抽取关键词信息，并将其和实验三中[信息抽取服务](#)进行融合。

具体流程是：

1. python 使用 flask 框架接受请求;
2. 将图片输入到 ResNet50 模型中进行对象识别;
3. 返回识别到的关键词;
4. 调用信息抽取的代码, 使用识别到的关键词进行抽取。

接口代码如下:

```
@app.route("/image_to_keywords", methods=["POST"])
def image_to_keywords():
    if "file" not in request.files:
        return "No file part", 400
    file = request.files["file"]

    if file.filename == "":
        return "No selected file", 400
    log.info(file.filename)
    result, code = image_detection.image_to_keywords(file)
    if code != 200:
        log.error(result)
        return result, code
    return result, code
```

模型处理代码如下 (错误处理代码则删去不再展示) :



```

config = tf.compat.v1.ConfigProto(
    gpu_options=tf.compat.v1.GPUOptions(allow_growth=True))
sess = tf.compat.v1.Session(config=config)

log = logging.getLogger("ImageToKeywords")

# Image object detection model
model = ResNet50(weights="imagenet")

def image_to_keywords(file: str) -> tuple[str, int]:
    if file.filename == "":
        return ("No selected file", 400)
    log.info(file.filename)

    img = (
        Image.open(io.BytesIO(file.read()))
        .convert("RGB")
        .resize((224, 224))
    )

    x = img_to_array(img)

    x = np.expand_dims(x, axis=0)
    x = preprocess_input(x)

    preds = model.predict(x)
    predictions = decode_predictions(preds, top=5)[0]

    if len(predictions) >= 3:
        keywords = [pred[1] for pred in predictions[:3]]
    else:
        keywords = [pred[1] for pred in predictions]
    keywords = " ".join(kw for kw in keywords)
    log.info(keywords)
    return (json.dumps({"keyword": keywords}), 200)

```

## 抽取结果评价

### 实体反馈

- 组件： <v-rating>

```

<v-rating
  dense
  hover
  small
  v-model="detailMap[result.Doc.id].entities[key].score"
  @input="handleEntityFeedback(result.Doc.id, key, detailMap[result.Doc.id].entities[key])"
></v-rating>

```

实体	频率	反馈
8	3	★ ★ ★ ☆ ☆
Chapter 5	3	★ ★ ☆ ☆ ☆
one	3	★ ★ ★ ★ ★
two	6	☆ ☆ ☆ ☆ ☆
's	3	★ ★ ★ ☆ ☆

- **功能：**允许用户对搜索结果中每个实体的准确性进行评分。
- **属性：**
  - `dense` 和 `small` 使评分组件更紧凑、适合放置在搜索结果卡片中。
  - `hover` 允许用户在鼠标悬停时预览评分效果。
  - `v-model` 绑定到 `detailMap[result.Doc.id].entities[key].score`，实现数据的双向绑定。
  - `@input` 事件处理函数 `handleEntityFeedback` 发送用户的评分到后端。
- **布局：**每个搜索结果的详细信息区域均设有实体反馈评分组件，与实体信息并排展示。
- **接口函数：**

```

handleEntityFeedback(resultId, item, score) {
  const payload = {
    item,
    resultId,
    score
  };
  axios.post(`api/v1/entity_feedback`, payload)
    .then(response => {
      console.log("Entity Feedback sent successfully", response);
    })
    .catch(error => {
      console.error("Error sending entity feedback", error);
    });
},

```

# 热词反馈

- 组件: <v-rating>

```
<v-rating
  dense
  hover
  small
  v-model="detailMap[result.Doc.id].hot_words[key].score"
  @input="handleHotwordFeedback(result.Doc.id, key, detailMap[result.Doc.id].hot_words[ke
></v-rating>
```

热词	频率	反馈
code	24	☆☆☆☆☆
function	36	★★★★☆
test	89	★★★★☆
tests	35	★★★★☆
,	63	★★☆☆☆

- 功能: 允许用户对搜索结果中的热词进行评分。
- 属性:
  - dense 和 small 使评分组件更紧凑、适合放置在搜索结果卡片中。
  - hover 允许用户在鼠标悬停时预览评分效果。
  - v-model 绑定到 detailMap[result.Doc.id].hot\_words[key].score , 实现数据的双向绑定。
  - @input 事件处理函数 handleEntityFeedback 发送用户的评分到后端。
- 布局: 与实体反馈类似, 热词反馈组件与对应的热词信息并排展示。
- 接口函数:

```

handleHotwordFeedback(resultId, item, score) {
  const payload = {
    item,
    resultId,
    score
  };
  axios.post(`api/v1/hotword_feedback`, payload)
    .then(response => {
      console.log("Hotword Feedback sent successfully", response);
    })
    .catch(error => {
      console.error("Error sending hotword feedback", error);
    });
},

```

## 信息抽取反馈

- 组件: `<v-rating>`

```

<v-rating
  dense
  hover
  small
  v-model="regexMap[result.Doc.id].results[word].score"
  @input="handleRegexFeedback(result.Doc.id, word, regexMap[result.Doc.id].results[word].
></v-rating>

```

### Programming a Guessing Game

2024-05-31

介绍: Let's jump into Rust by working through a hands-on project together! This chapter introduces you t...

相关度: 389893.77405271505

链接: <https://doc.rust-lang.org/book/ch02-00-guessing-game-tutorial.html>

选择词性  
 数词

关键词	评分
first	★★★★☆
second	★★★★☆
third	★★★★☆
First	☆☆☆☆

☆☆☆☆☆

- 功能: 允许用户对正则表达式搜索结果中的每个关键词进行评分。
- 属性:
  - `dense` 和 `small` 使评分组件更紧凑, 适合放置在搜索结果卡片中。

- `hover` 允许用户在鼠标悬停时预览评分效果。
- `v-model` 绑定到 `regexMap[result.Doc.id].results[word].score`，实现数据的双向绑定，保持界面和数据状态的同步。
- `@input` 事件处理函数 `handleRegexFeedback` 用于当用户修改评分时发送这一改变到后端。
- **布局**：评分组件在每个搜索结果的正则表达式搜索结果表格中，与关键词并排显示。
- **接口函数**：

```
handleRegexFeedback(resultId, word, score) {
  const payload = {
    item: word,
    resultId: resultId,
    score: score
  };
  axios.post(`api/v1/regex_feedback`, payload)
    .then(response => {
      console.log("Regex Feedback sent successfully", response);
    })
    .catch(error => {
      console.error("Error sending regex feedback", error);
    });
}
```

## 整体准确率评价反馈

- **组件**： `<v-rating>`

```
<v-rating
  dense
  hover
  v-model="result.Score"
  @input="handleOverallFeedback(result.Doc.id, result.Score)"
></v-rating>
```

### Controlling How Tests Are Run

2024-05-31

介绍: Just as compiles your code and then runs the resulting binary, compiles your code in test mode and...

相关度: 1509.4414788276588

链接: <https://doc.rust-lang.org/book/ch11-02-running-tests.html>



- **功能**：提供对整个搜索结果的整体准确率满意度评价。
- **属性**：

- hover 和 dense 属性同上。
- v-model 绑定到 result.Score。
- @input 通过 handleOverallFeedback 方法发送整体评分数据到后端。
- **布局**：整体评价组件通常位于搜索结果卡片的底部，方便用户在查看完信息后给出整体准确率评价。
- **接口函数**：

```
handleOverallFeedback(resultId, Score) {  
  const payload = {  
    resultId,  
    Score  
  };  
  axios.post(`api/v1/feedback`, payload)  
    .then(response => {  
      console.log("Overall Feedback sent successfully", response);  
    })  
    .catch(error => {  
      console.error("Error sending overall feedback", error);  
    });  
}
```

## 优化与创新性

在本次实验中，通过不断地打磨我们的项目，我们实现了以下优化和创新：

### 后端-算法优化

- 缓存优化：后端自己实现了 LRU 缓存，在内存中缓存每次查询的结果，减少了重复的查询计算，提高了查询的性能和效率；
- 并发优化：在检索时我通过 Go 的协程并行计算每篇文档的得分，这样充分利用了 Go 轻量级协程的优势，大大提高了检索的性能；
- 抽取优化：在抽取过程中，我分别使用了 spaCy 和 jieba 对中英文不同文本进行分词和词性识别处理，大大提高了抽取的准确性和效率；
- 基于用户反馈动态修改排名：对于每个查询结果，我们都设置用户可以对其进行反馈，并且通过用户反馈修改查询结果的权重，动态地调整查询结果的排名；
- 多媒体抽取：在项目中我们引入了图片识别模型，通过识别提取图片关键词，从而实现多媒体检索和信息抽取。

# 前端

## 优化措施

- **异步数据处理**：使用 Vue.js 的异步组件和 axios 进行数据请求，优化了页面加载速度和响应时间，避免了在请求数据时阻塞用户界面的问题。
- **组件化开发**：利用 Vue.js 的组件化能力，将前端界面分解为可重用的组件，如搜索栏、结果卡片、反馈评分等。这种方式不仅提高了代码的可维护性，也简化了功能的扩展。
- **用户体验优化**：
  - 在关键组件中实施了响应式设计，确保在不同设备上都能提供良好的用户体验。
  - 通过细致的动画和过渡效果增强了界面的交互性，如加载动画和按钮点击反馈。
  - 针对用户操作提供即时的反馈信息，比如在发送反馈评分后，通过控制台日志确认反馈已成功发送。
- **错误处理与数据验证**：
  - 在数据输入和网络请求中实施了全面的错误处理机制，确保了应用的稳定运行和用户的流畅体验。
  - 对用户输入进行验证，防止无效或恶意的数据被提交到后端。

## 创新性特点

- **集成的反馈机制**：引入了多层次的用户反馈系统，包括实体评分、热词评分和整体搜索结果评分。这不仅使用户能够直接参与改善搜索结果，也为算法优化提供了实时数据。
- **多模态搜索功能**：实现了关键词搜索与图像搜索的结合，提供了一种多模态的信息检索方式。用户可以通过文本或图像中的内容进行搜索，增强了搜索的灵活性和准确性。
- **实时动态更新**：通过 Vue.js 的双向数据绑定和组件状态管理，实现了搜索结果和用户反馈的实时动态更新。用户在界面上的任何操作都可以即时反映，提高了交互的实时性。

# 环境和社会可持续发展思考

## 环境影响

- **减少能源消耗**：信息检索系统通过自动化处理大量数据，可能会对服务器造成较大负荷，进而影响能源消耗。为了降低这种影响，我们采用了能效较高的数据处理算法和节能的服务器配置。此外，通过优化爬虫的爬取策略，减少不必要的请求，可以进一步减轻对源网站服务器的压力，间接减少整个网络中的能源消耗。
- **绿色技术选择**：在选择第三方服务和云平台时，优先考虑那些承诺使用可再生能源并具有良好碳足迹记录的供应商。例如，使用支持绿色能源的数据中心可以减少项目的环境负担。

## 社会效益

- **信息获取的公平性**：本信息检索系统支持中英文内容的处理，有助于跨语言和文化的信息流通，促进了知识的平等获取。这对于教育资源的公平分配尤其重要，可以帮助来自不同背景的用户访问和利用全球的信息资源。
- **促进知识共享和教育**：系统提供的多媒体关键词提取服务和自然语言查询功能，使用户能够更容易地找到所需信息，从而促进知识的传播和教育的普及。这种技术的应用尤其可以支持教育不发达地区的学习和研究，减少城乡之间的教育差异。
- **提高社会意识和参与**：通过提供高效的信息检索工具，可以增强公众对于重要社会问题的意识和理解，比如环境保护、公共健康和社会正义等。用户可以更容易地获取相关信息，从而在这些重要问题上做出更为明智的决策和参与。

## 技术与可持续性的结合

- 在技术实现方面，本项目特别重视环保和可持续性原则。通过采用最新的算法优化技术，我们显著提高了数据处理效率，从而减少了能耗。同时，项目在选择服务器和存储解决方案时，优先考虑那些采用可再生能源的服务提供商。这种策略不仅降低了系统的环境足迹，也体现了我们对环境保护的承诺。

## 促进包容性和平等

- 信息检索系统设计之初就考虑到了多样性和包容性，特别是在语言处理功能上。系统支持中英文的自然语言处理，确保了不同语言用户的信息获取需求得到满足。此外，我们通过用户界面的多语言支持和无障碍设计，使系统对不同文化和能力水平的用户都友好，从而推动了信息获取的平等性和包容性。

# 实验总结

本次实验的主旨在于设计并实施一款复合型信息抽取系统，旨在针对中英文文档执行高效检索，并精确展示检索结果。在实验的推进中，我们遭遇了多重挑战，同时收获了丰富的知识与实践经验。

- **数据获取与预处理**：我们依托 Scrapy 框架完成了网页数据的爬取工作，并将这些数据保存为便于后续处理的 JSON 格式。数据预处理环节，尤其是针对中文内容的分词，成为提升检索精确性的关键步骤。我们选用了 jieba 分词工具，它在处理中文文本时表现出优越的性能，但在面对特定领域术语时，需要我们精细调整参数以达到最佳分词效果；
- **信息点抽取与算法优化**：实现高效的信息点抽取功能，涉及设计并优化正则表达式匹配算法，以及开发针对特定信息点（如人名、地点等）的抽取逻辑。我们不仅在理论层面深入学习了信息抽取技术，还在实践中反复调试，确保信息抽取的准确性和效率。此外，引入用户反馈机制，动态调整搜索结果排序，进一步提升了系统的表现；
- **多模态信息处理**：在系统中整合了图片检索功能，利用 TensorFlow 进行图像识别，Flask 框架构建了图片检索服务接口。这一过程不仅涉及图像的大量处理，还包括从图像中提取关键词以供检



索，对计算资源和算法的智能程度提出了更高标准。

通过本实验，我们不仅深化了对信息抽取、自然语言处理以及机器学习技术的理解，还掌握了 Scrapy 框架的高级应用，jieba 分词的细微调优，以及如何在 TensorFlow 框架下实施图像识别和关键词提取。系统成功集成了从数据爬取至前端展示的完整流程，有效处理了用户查询，并依据相关性返回精确信息。特别是在处理中英文混合信息时，系统展示了高度的灵活性和准确性。总之，此实验不仅锻炼了我们的技术实践能力，也深化了我们对信息抽取领域前沿技术的认识。

## 实验分工

	郭晨旭	韩景锐
学号	2021211184	2021211176
代码	Go 后端，Scrapy 爬虫，Python 图片提取关键词服务，Python 信息抽取服务	Vue 前端
报告	项目概述，后端设计与实现，信息抽取服务， 多媒体信息抽取服务，优化与创新	前端设计与实现，优化与创新， 环境和社会可持续发展思考，实验总结