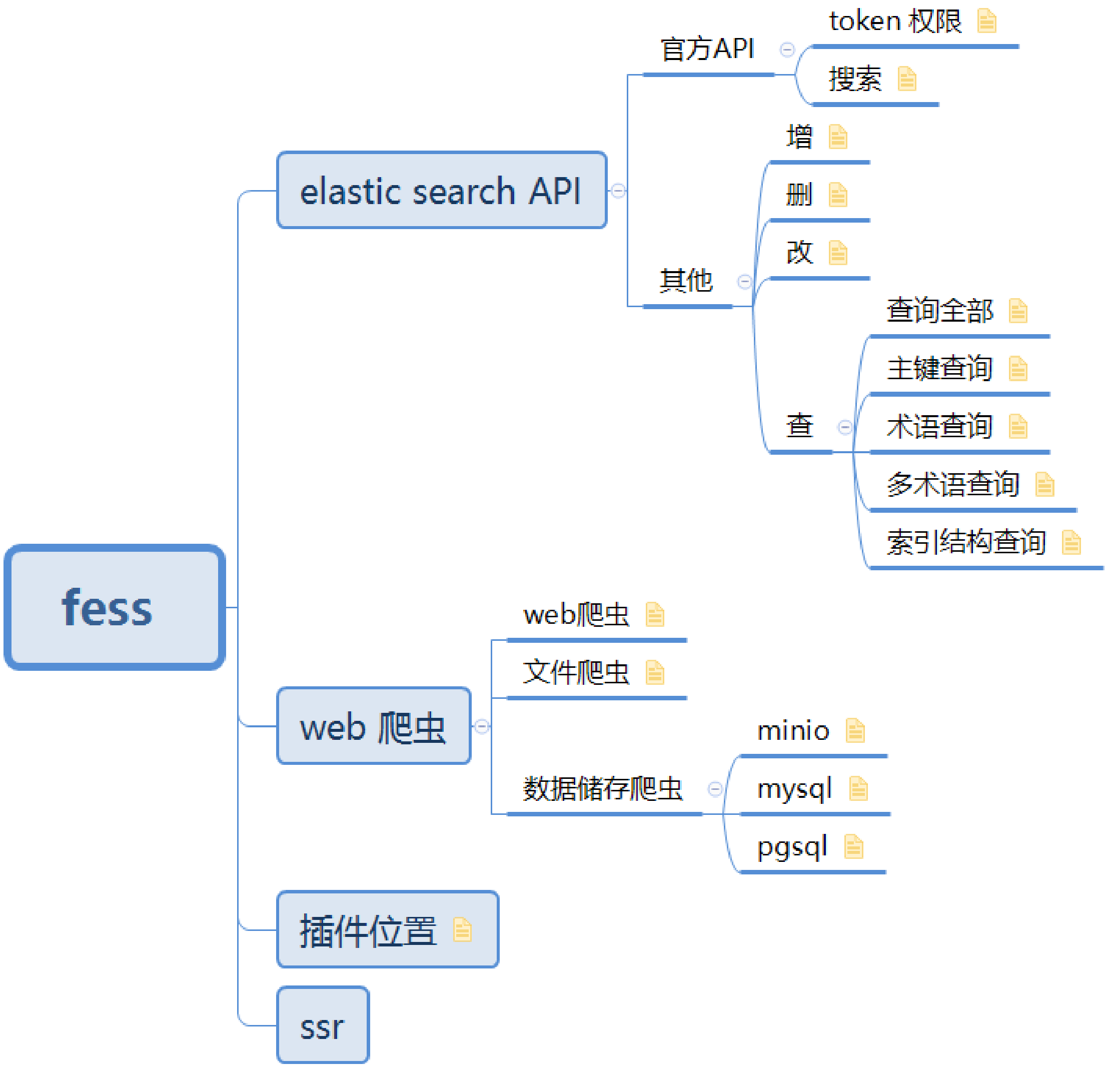
# fess



# 1. elastic search API

## 1.1. 官方API

### 1.1.1. token 权限

1. 生成一个 token

2. header 加入

Authorization: access_token

### 1.1.2. 搜索

调用方式1

http://localhost:8080/search/?as.q=&as.epq=&as.oq=&as.nq=&num=20&sort=&as.timestamp=&as.filetype=&as.occt=&as.sitesearch=

======================

参数　对应搜索页面上的字段　解释　可选值

======================

as.q　all these words　搜索的关键字　字符串 空格分隔

======================

as.epq　phrase search　短语搜索　字符串空格分隔

======================

as.oq　any of these words　包含这些词中的一个　字符串空格分隔

======================

as.nq　none of these words　不包含这些词中的任意一个　字符串空格分隔

======================

num　results per page　每页显示的数量　数字

======================

sort　sort　排序

score.desc

filename.asc

其他字段

created.asc

content_length.asc

last_modified.asc
click_count.asc

=====================

lang　语言　语言

all，ar 阿拉伯语…

=====================

as.timestamp last update 最后更新时间

[now-1d/d To *] 24小时

[now-1w/d To *] 一周内

[now-1M/d To *] 一个月内

[now-1y/d To *] 一年内

=====================

as.filetype file type文件类型

html
…

=====================

as.occt　terms appearing 出现在哪？

allintitle 在标题中

allinurl　在url中

或者为空。在页面中任何位置

=====================

as.sitesearch　site or domain　site 或者域名　字符串

======================

调用方式2

http://localhost:8080/search/?q=%E4%BA%BA%E6%B0%91&num=20&sort=

q:　搜索的内容

num: 和上面的一样

sort: 和上面的一样

lang : 语言

示例

调用接口http://localhost:8080/search/?

mothods : **GET**

传参方式 1

http://localhost:8080/search/?as.q=%E4%BA%BA%E6%B0%91+&as.epq=%E4%B8%AD%E5%9B%BD%E4%BE%A8%E8%81%94+%E5%AD%A6%E4%B9%A0%E5%BC%BA%E5%9B%BD&as.oq=%E4%BA%BA%E6%B0%91+%E6%B5%99%E6%B1%9F+%E8%9C%82%E8%9C%9C&as.nq=%E9%A6%96%E9%A1%B5&num=10&sort=click_count.asc&lang=ar&as.timestamp=&as.filetype=&as.occt=&as.sitesearch=

解析:

```
{
 "as.q": "人民",
 "as.epq": "中国侨联 学习强国",
 "as.oq": "人民 浙江 蜂蜜",
 "as.nq": "首页",
 "num": 10,
 "sort": "click_count.asc",
 "lang": "ar",
 "as.timestamp": "",
 "as.filetype": "",
 "as.occt": "",
 "as.sitesearch": ""
}
```

传参方式 2

http://localhost:8080/search/?q=%E4%BA%BA%E6%B0%91+%22%E4%B8%AD%E5%9B%BD%E4%BE%A8%E8%81%94+%E5%AD%A6%E4%B9%A0%E5%BC%BA%E5%9B%BD%22+%28%E4%BA%BA%E6%B0%91+OR+%E6%B5%99%E6%B1%9F+OR+%E8%9C%82%E8%9C%9C%29+NOT+%E9%A6%96%E9%A1%B5+sort%3Aclick_count.asc+site%3Ablog.csdn.net+sort%3Aclick_count.asc&num=10&sort=click_count.asc&lang=ar

解析:

```
{
 "q": "人民 \"中国侨联 学习强国\" (人民 OR 浙江 OR 蜂蜜) NOT 首页 sort:click_count.asc site:blog.csdn.net sort:click_count.asc",
 "num": 10,
 "sort": "click_count.asc",
 "lang": "ar"
}
```

## 1.2. 其他

### 1.2.1. 增

```
POST /fess.20230712/_doc
{
  "lang": "zh",
  "anchor": [
    "http://www.swahili.people.cn/index.html"
    "http://www.swahili.people.cn/416664/index.html"
    "http://www.swahili.people.cn/416665/index.html"
    "http://www.swahili.people.cn/img/FOREIGN/2021/05/313146/static/css/mob.css
  ],
  "boost": "1.0",
  "cache": """""",
  "click_count": 0,
  "config_id": "W_x0oSYkBOlYYuW5v-k1z",
  "content": """这是手动插入的数据""",
  "content_length": "28486",
  "created": "2023-08-08T11:15:29.148Z",
  "digest": "这是手动插入的数据",
  "doc_id": "6a3ac829444246deaf3f230183770896",
  "favorite_count": 0,
  "filetype": "html",
  "has_cache": "true",
  "host": "www.swahili.people.cn",
  "important_content": "Lugha Nyingine",
  "label": [],
  "last_modified": "2023-08-08T08:17:07.000Z",
  "mimetype": "text/html",
  "parent_id":
"7f68b53f718d6eb56ad9d62cf05efa40974279fc8da3a1d33fabbab4f4e3f34e36060aef7161e5cf2a1b456a0e16e65a
907baab357e902c9405a0ea16894341d",
  "role": [
   "Rguest"
  ],
  "segment": "_x0oSYkBOlYYuW5v_k1z",
  "site": "www.swahili.people.cn/",
  "thumbnail":
"http://www.swahili.people.cn/NMediaFile/2023/0808/FOREIGN16914809878616I1I2A4EQ5.jpg"
  "timestamp": "2023-08-08T08:17:07.000Z",
  "title": "Tovuti ya Gazeti la Umma--People's Daily Online这是手动插入的数据",
  "url": "http://www.swahili.people.cn/"
  "virtual_host": []
 }
```

### 1.2.2. 删

```
DELETE /fess.20230712/_doc/{_id}
```

```
DELETE
/fess.20230712/_doc/3e49d941b960b16e0523b27d064ba5f89906f3a02fa7df0dfb020800d681fd5e773ea101d0dc8
6a953f465bf7936f93cb9f8c5e20bb24438732ad3f95908795d
```

### 1.2.3. 改

```
POST fess.20230713/_update/{_id}
{
 "doc": {
"{字段}":"{值}"
 }
}
```

```
POST
fess.20230713/_update/4e16c6d3a86c98381fcf9c0e8c8174e803b2ecd2f2d7c31cc1df7c86dc63eaae885b1c016d18
ff4d75da88403160fca301df77ba7409084633b90f68d78ce2be
{
  "doc": {
    "url":"http://10.17.1.37:9000/fess/doc/doc1%20-%20%E5%89%AF%E6%9C%AC%20-
%20%E5%89%AF%E6%9C%AC.docx
  }
}
```

### 1.2.4. 查

#### 1.2.4.1. 查询全部

```
GET /fess.20230712/_search
```

#### 1.2.4.2. 主键查询

```
POST /fess.20230712/_search
{
  "query": {
    "ids": {
      "values":
"2d42f2715fd3f3276b233f107b16d426f2a153e603615e36875e6a733089b6a2aa195b13e56a8031249cb148a70
9d4fb2e3cc3c6990cb41cbe5b64b3fd4c40b8"
    }
  }
}
```

或者

```
GET fess.20230713/_doc/{_id}
```

### 1.2.4.3. 术语查询

```
POST /fess.20230712/_search
{
  "query": {
    "term": {
      "title": {
        "value": "人民"
      }
    }
  }
}
```

### 1.2.4.4. 多术语查询

```
POST /fess.20230712/_search
{
  "query": {
    "terms": {
      "title":  [
          "营养",
          "浙江"
          ]
    }
  }
}
```

### 1.2.4.5. 索引结构查询

```
GET /fess.20230712/_mapping
```

# 2. web 爬虫

## 2.1. web爬虫

各项参数解释

Name 爬虫名称 解释

URLs 目标地址 要爬取的起始地址

Included URLs For Crawling 爬取的URl（java 正则）

Excluded URLs For Crawling 不爬取的URl（java 正则）

Included URLs For Indexing 爬取的这个url 要不要存入 elasticsearch

Excluded URLs For Indexing 爬取的这个url 要不要存入 elasticsearch

Config Parameters 自定义赋值(grovvy语法)

Depth

Max Access Count 最大数量(可以存入 elasticsearch 的最大数量 )

User Agent agent

The number of Thread 几个县城

Interval time 周期访问

Boost elasticsearch 中的 权重

Permissions {role}secondarySite

Virtual Hosts

Status Enabled

Description

示例

IDESwO8okBjJ3PH9r9QOwg

Name http://10.17.1.37:57833/index.html

URLs http://10.17.1.37:57833/index.html

Included URLs For Crawling

Excluded URLs For Crawling (?i).*(css|js|jpeg|jpg|gif|png|bmp|wmv|xml|ico|exe)

Included URLs For Indexing

Excluded URLs For Indexing .*index\.html.*

Config Parameters

Depth　2

Max Access Count 5

User Agent　Mozilla/5.0 (compatible; Fess/14.9; +http://fess.codelibs.org/bot.html)

The number of Thread　1

Interval time 5 ms

Boost　1.0

Permissions　{role}secondarySite

Virtual Hosts

Status　Enabled

Description

## 2.2. 文件爬虫

示例

Paths file:///opt/rsts/

## 2.3. 数据储存爬虫

### 2.3.1. minio

示例 =》

Handler Name　==》

AmazonS3DataStore

Parameter　==》

```
region="111"
access_key_id=AKIAIOSFODNN7EXAMPLE
secret_key={cipher}aa05e5ba93401bdc5b08249be6e4cf5bfdf923a76c015cc8543022d0519407f883461fa44479f985d600869f564e8c88
endpoint=http://127.0.0.1:9000
buckets=fess
```

Script　==》

```
url=object.url
title=object.key
content=object.contents
mimetype=object.mimetype
filetype=object.filetype
```

```
filename=object.filename
content_length=object.size
last_modified=object.last_modified
```

其他 ==》

```
url=
    object.url.replaceFirst( object.bucket_name+ ".","")
        .substring(
            0,
            object.url.replaceFirst( object.bucket_name+ ".","").indexOf("/",7)
        )
    + "/"
    + object.bucket_name
    + object.url.replaceFirst( object.bucket_name+ ".","")
        .substring(
            object.url.replaceFirst( object.bucket_name+ ".","")
                .indexOf("/",7)
                )
```

如果 ip 是固定的，可以写成

```
url="http://IP:端口/"
    + object.bucket_name
    +”/” + object.key
object.url
```

例如

"http://10.17.1.37:9000/"+ object.bucket_name +"/" + object.ke

原来是 http://桶名.IP:端口/路徑

需要转成 http://IP:端口/桶名/路徑

爬取的object对象全部字段

```
{
        "ssekms_key_id": null,
      "delete_marker": null,
      "filetype": "word",
      "expires": null,
      "owner_id": "02d6176db174dc93cb1b899f7c6078f08654445fe8cf1b6ce98d8855f66bdbf4",
      "content_language": null,
      "accept_ranges": "bytes",
      "request_charged": null,
      "tag_count": null,
      "content_type": "application/vnd.openxmlformats-officedocument.wordprocessingml.document",
      "content_range": null,
      "content_disposition": null,
      "missing_meta": null,
      "website_redirect_location": null,
      "last_modified": "2023-08-22T05:37:19.044Z",
      "key": "a1/doc1.docx",
      "content_length": 12538,
      "object_lock_mode": null,
      "storage_class": "STANDARD",
      "management_url":
"https://s3.console.aws.amazon.com/s3/object/fess/a1/doc1.docx?region=%2522111%2522"
      "object_lock_legal_hold_status": null,
      "parts_count": null,
      "restore": null,
      "bucket_name": "fess",
      "content_encoding": null,
      "owner_display_name": "minio",
      "cache_control": null,
      "creation_date": "2023-08-22T05:35:39.348Z",
      "version_id": null,
      "url": "http://fess.10.17.1.37:9000/a1/doc1.docx"
      "sse_customer_key_md5": null,
      "server_side_encryption": null,
      "filename": "doc1.docx",
      "size": 12538,
      "contents": "Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1 Doc1
Doc1 Doc1 Doc1 Doc1",
      "sse_customer_algorithm": null,
      "expiration": null,
      "mimetype": "application/vnd.openxmlformats-officedocument.wordprocessingml.document",
      "e_tag": "\"49a5aa355da3308880ee287a47f398f1\"",
      "replication_status": null,
      "object_lock_retain_until_date": null
     }
```

### 2.3.2. mysql

准备 =》

1. 下载插件和 驱动

    DatabaseDataStore

    jdbc

2. 更改java配置，然后重启

    文件位置

    /opt/java/openjdk/conf/security/java.security

    修改内容

```
      # jdk.tls.disabledAlgorithms=SSLv3, TLSv1, TLSv1.1, RC4, DES, MD5withRSA, \
#     DH keySize < 1024, EC keySize < 224, 3DES_EDE_CBC, anon, NULL


jdk.tls.disabledAlgorithms=RC4, DES, MD5withRSA, \
    DH keySize < 1024, EC keySize < 224, 3DES_EDE_CBC, anon, NULL
```

爬虫配置=》

参数设置

```
driver=com.mysql.cj.jdbc.Driver
url=jdbc:mysql://10.17.1.37:3306/books?useUnicode=true&characterEncoding=UTF-8
username=root
password={cipher}17e8271562b4af1cf5915fbb75720795
sql=select * from Book
```

脚本设置（想不出来场景）

```
url="http://localhost/" + BookID
host="localhost"
site="localhost"
title=Title
content=ISBN
cache=ISBN
digest=ISBN
anchor=Author
content_length=
last_modified=new java.util.Date()
```

### 2.3.3. pgsql

准备 =》

1.在插件中安装驱动

爬虫配置 =》

参数

```
driver=org.postgresql.Driver
url=jdbc:postgresql://10.17.1.37:5432/books
username=postgres
password={cipher}17e8271562b4af1cf5915fbb75720795
sql=SELECT * from articles
```

脚本

```
url="http://localhost/" + id
host="localhost"
site="localhost"
title=title
content=content
cache=content
digest=content
anchor=author
content_length=content.length()
last_modified=new java.util.Date()
```

## 3. 插件位置

插件存放位置
/usr/share/fess/app/WEB-INF/plugin

注意:

在文件夹中放入插件，虽然在 system -》 plugin 中有显示。但是在 数据爬虫中仍然无法读取到此插件。需要再重新安装

## 4. ssr