TED数据爬取

TED数据爬取

- 工具说明
- 与TED视屏相关数据的爬取
  - 爬取目标
  - 数据库建表语句
  - 字段
  - 程序主方法
- 关键技术点
  - 网页检查数据与Postman的get数据不一致
    - 正则表达式
  - 网页具有反爬虫机制（未解决）

工具说明

- 电脑：MacOS High Sierra 10.13.1
- Java版本

  java version "1.8.0_131"

  Java(TM) SE Runtime Environment (build 1.8.0_131-b11)

  Java HotSpot(TM) 64-Bit Server VM (build 25.131-b11, mixed mode)
- MySql版本

  Server version: 5.7.22 MySQL Community Server (GPL)
- Postman版本

与TED视屏相关数据的爬取

爬取目标

TED里所有演讲视屏的相关数据

对于每一个视屏爬取的数据，包括，以下图为例：

数据库建表语句

```
CREATE TABLE `CHNRecallAutoNews` (
  `threadId` VARCHAR(100) NOT NULL,
  `current_num` varchar(100) DEFAULT NULL,
    `related_tag` varchar(1000) DEFAULT NULL,
    `speaker_tag` VARCHAR(4000) DEFAULT NULL,
    `views` varchar(2000) DEFAULT NULL,
    `comment_num` varchar(2000) DEFAULT NULL,
    `speaker` varchar(2000) DEFAULT NULL,
    `title` varchar(2000) DEFAULT NULL,
    `rated` varchar(2000) DEFAULT NULL,
    `posted` varchar(2000) DEFAULT NULL,
    `description` varchar(2000) DEFAULT NULL,
    `about_speaker` varchar(2000) DEFAULT NULL,
  PRIMARY KEY (`threadId`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

字段

表名`teds.csv`：

- `threadId`:主键
- `current_num`：因为在爬取数据时，不能确定那一个当做主键，所以都爬取下来了。
- `related_tag`：ted视屏的标签，该标签指的是该视频与什么领域相关。
- `speaker_tag` ：演讲者的标签，比如：作家，等
- `views`：该视频观看的人数
- `comment_num`：该视频评论数
- `speaker`：演讲者的姓名
- `title`：演讲标题
- `rated`：视屏标签，比如：鼓舞人心，有趣，等
- `posted`：演讲发布时间，如： Jun 2018
- `description`:视屏内容的描述
- `about_speaker` ：关于演讲者的信息

程序主方法

`Crawlers.main.TEDmain`

关键技术点

网页检查数据与Postman的get数据不一致

网页中包含`javascript`，因此，只能获取如下标签和内容:

```
<script>q("talkPage.init", {
    "el": "[data-talk-page]",
    "__INITIAL_DATA__": {"comments":
{"id":27984,"count":30,"talk_id":17239},"threadId":2798
4,"current_talk":"17239","description":"What happens
when technology knows more about us than we do? Poppy
Crum studies how we express emotions -- and she
suggests the end of the poker face is near, as new tech
makes it easy to see the signals that give away how
we're feeling. In a talk and demo, she shows how
\"empathetic technology\" can read physical signals
like body temperature and the chemical composition of
our breath to inform on our emotional state. For better
or for worse. \"If we recognize the power of becoming
technological empaths, we get this opportunity where
technology can help us bridge the emotional and
cognitive divide,\" Crum
says.","event":"TED2018","language":"en","name":"Poppy
Crum: Technology that knows what you're
feeling","slug":"poppy_crum_technology_that_knows_what_
you_re_feeling","speakers":
[{"id":"3971","slug":"poppy_crum","is_published":true,"
firstname":"Poppy","lastname":"Crum","middleinitial":""
,"title":"","description":"Neuroscientist,
technologist","photo_url":"https://pe.tedcdn.com/images
/ted/284f28a25245a04794355c7f7059ae3f5860b3f6_254x191.j
pg","whatotherssay":"","whotheyare":"Poppy Crum builds
technologies that best leverage human physiology to
enhance our experiences and how we interact with the
world."}],"time":
    </script>
```

使用`Jsoup`不能解析`script`内容，考虑使用正则表达式匹配。

正则表达式

对几个关键正则表达式举例说明

匹配目标：`"threadId":27984`

正则表达式：`String threadId_par = "(talk_id\":)(\\d+)";`

匹配目标：`"description":"Neuroscientist, technologist"`

正则表达式：`String speaker_tags_par = "(\"speakers\":\\[\\{)(.*)(\\}\\])";`和`String speaker_par = "(\"description\":\")(\\D*)(\",\")";`

匹配目标：`"whotheyare":"Poppy Crum builds technologies that best leverage human physiology to enhance our experiences and how we interact with the world."`

正则表达式：`String about_speakers_par = "(\"whotheyare\":\")(\\D*)(\",\")";`

匹配目标：``

网页具有反爬虫机制（未解决）

在多次请求页面后，无法获取信息，考虑是反爬虫机制，考虑以下两个解决方案：

- 买代理IP
- 设置随机休息时间