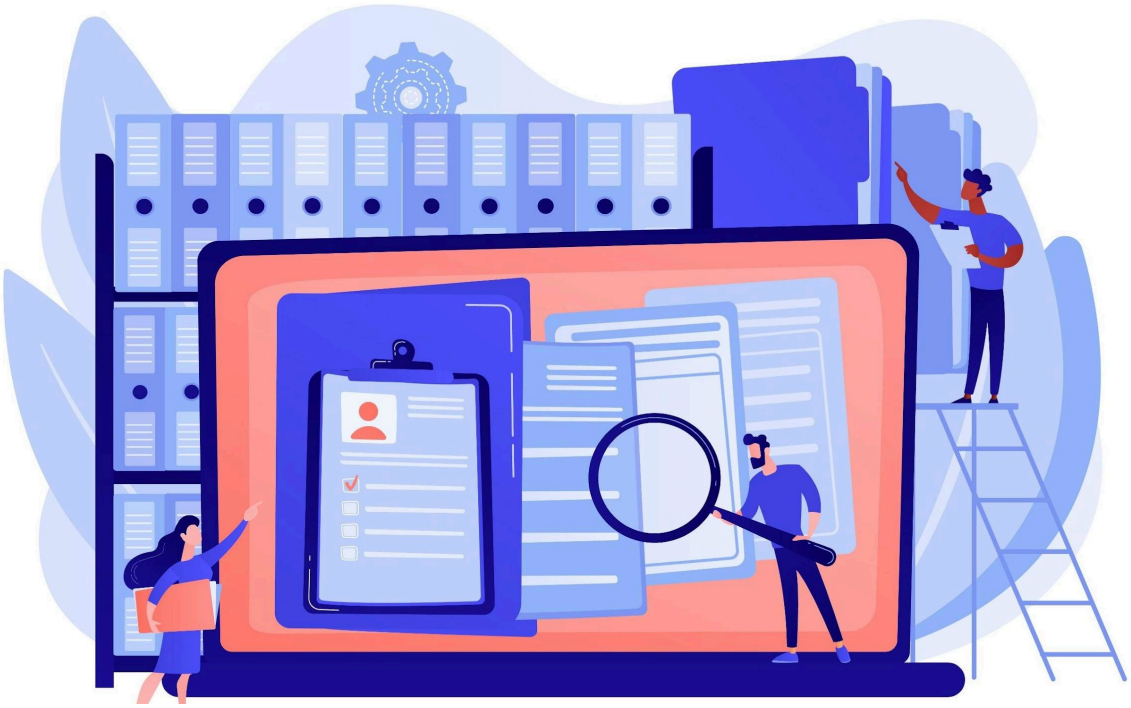

مبانی بازیابی اطلاعات و جستجوی وب

نیم سال دوم 1402-1403

تمرین 3

اهداف:

آشنایی بیشتر با فرایندهای استخراج اطلاعات

مهلت تحویل:

دوشنبه ۲۸خرداد ۱۴۰۳

توضیحات تمرین

در این تمرین می‌خواهیم از سایت IEEE Xplore، دیتای مرتبط با مقالات یک حوزه را crawl کنیم. از اطلاعات استخراج‌شده در این تمرین، در تمرین بعدی یعنی Elasticsearch استفاده خواهد شد.

مراحل دریافت اطلاعات

1. ابتدا به سایت IEEE Xplore مراجعه کنید.
2. در نوار سرچ وب‌سایت، یک حوزه تحقیقاتی را جستجو کنید. برای مثال `Blockchain`
3. اطلاعات زیر را در فرمت `JSON` برای مقالاتی که در ۵ صفحه اول نتایج جستجو هستند (هم به ترتیب `Relevance` و هم به ترتیب `Newest`) استخراج کنید.

فیلد	نوع دیتا	توضیحات
title	String	عنوان مقاله
Page(s)	Integer	تعداد صفحات
Cites in Papers	Integer	تعداد ارجاعات در مقالات دیگر
Cites in Patent	Integer	تعداد ارجاعات در پتنت‌ها
Full Text Views	Integer	تعداد مشاهده‌های کامل متن
Publisher	String	ناشر مقاله
DOI	String	(DOI) شناسه دیجیتال شی
Date of Publication	String	تاریخ انتشار مقاله
abstract	String	چکیده مقاله
Published in	List of Objects	منتشر شده در: اطلاعات کنفرانس یا مجله‌ای که مقاله در آن منتشر شده است
Authors	List of Objects	نویسندگان: اسامی نویسندگان مقاله و سازمان‌های مربوطه
IEEE Keywords	List of Strings	کلمات کلیدی IEEE: کلمات کلیدی تعریف شده توسط IEEE برای مقاله
Author Keywords	List of Strings	کلمات کلیدی نویسنده: کلمات کلیدی تعریف شده توسط نویسندگان مقاله

مثال

برای مثال از این مقاله (که اولین مقاله از نتایج جستجوی کلمه `Blockchain` بر حسب `Relevance` هست) اطلاعات زیر استخراج می‌شود:


```
{
 "title": "ArtChain: Blockchain-Enabled Platform for Art Marketplace",
 "Page(s)": 18,
 "Cites in Papers": 40,
 "Cites in Patent": 1,
 "Full Text Views": 4217,
 "Publisher": "IEEE",
 "DOI": "10.1109/Blockchain.2019.00068",
 "Date of Publication": "14-17 July 2019",
 "abstract": "Blockchain is an emerging technology that has the potential to revolutionize the global industry and create a trusted relationship in a multi-party business network. There are a number of practical use cases where blockchain has been applied. One specific area is the Art industry, where it is a natural fit in the way that art forensics and transactions are conducted, tracked and recorded. This motivates us to develop the ArtChain platform to assist the Art Industry. In this paper, we present ArtChain, which is an integrated trading system based on blockchain. It includes the front end, the back end, the services, the smart contract, the chain connection and the deployment scripts from the bottom to the top. To the best of our knowledge, this is the first deployed blockchain-enabled art trading platform in Australia. It provides a transparent yet privacy-preserving, and  tamper-proof transaction history for registration, provenance, and traceability of art assets. Our objective analysis and evaluation show that the ArtChain platform is applicable and practical. For the interest of other researchers, our system implementation related resources are open-sourced on Github.",
```

```
"Published in": [
  {
    "name": "2019 IEEE International Conference on Blockchain (Blockchain)",
    "link": "https://ieeexplore.ieee.org/xpl/conhome/8938397/proceeding"
  }
],
"Authors": [
  {
    "name": "Ziyuan Wang",
    "from": "Blockchain Innovation Centre, Swinburne University of Technology"
  },
  {
    "name": "Lin Yang",
    "from": "Blockchain Innovation Centre, Swinburne University of Technology"
  },
  {
    "name": "Qin Wang",
    "from": "Blockchain Innovation Centre, Swinburne University of Technology"
  },
  {
    "name": "Donghai Liu",
    "from": "Blockchain Innovation Centre, Swinburne University of Technology"
  },
  {
    "name": "Zhiyu Xu",
    "from": "Blockchain Innovation Centre, Swinburne University of Technology"
  },
  {
    "name": "Shigang Liu",
    "from": "Blockchain Innovation Centre, Swinburne University of Technology"
  }
],
```

```
 "IEEE Keywords": [
   "Art",
   "Blockchain",
   "Ecosystems",
   "Business",
   "History",
   "Distributed ledger"
 ],
 "Author Keywords": ["blockchain", "artwork", "provenance"]
}
```


شیوه تحویل

خروجی مورد نظر برای هر قسمت تمرین شامل بخش‌هایی‌ست که توضیحات آن به شرح زیر می‌باشد.

در بخش گزارشِ هر قسمت، اسامی اعضای گروه و مشارکت هر عضو که عددی بین 1 تا 10 هست را بنویسید.

گزارش

گزارشی از مراحل فنی انجام کار را بنویسید.

همچنین در صورت برخورد با چالش‌های مختلف فنی در زمان استخراج اطلاعات، چالش‌ها به همراه راه‌حل‌های پیاده‌شده یا پیشنهادی خود را نیز مکتوب کنید.

کد

فایلی با فرمت $ipnby$. که شامل کدهای شما می‌باشد.

اطلاعات استخراج‌شده

تمام اطلاعات استخراج شده را در دایرکتوری (فولدری) با نام *data* قرار دهید.

نکات

- تمامی مراحل کار - از باز کردن اولین صفحه وب‌سایت تا آخرین مرحله - باید به کمک Selenium یا کتاب‌خانه‌های مشابه (مانند Playwright) انجام شود و امکان انجام بخشی از کار بصورت دستی یا manual وجود ندارد. اجرای اسکریپت شما باید در انتها به خروجی نهایی از ساختار اطلاعات درخواست‌شده منتهی شود.
- لطفا همه اعضای تیم، خروجی‌هارا در VU آپلود نمایند.
- در نهایت فایل‌های خواسته‌شده در بخش "شیوه تحویل" را بصورت zip شده و با نام `IR_1402_02_HW3_GroupNumber.zip` ارسال کنید.

موفق باشید.