دانشگاه اصفهان
دانشکده مهندسی کامپیوتر

تمرین سوم: آشنایی با الاستیک سرچ
درس: تحلیل داده‌های حجیم

استاد: دکتر محمدعلی نعمت‌بخش
دستیاران: فرزانه طاهری، محمدعلی شهرام‌پور، پریسا لطیفی

نام و نام‌خانوادگی:
شماره دانشجویی:
لینک گیت:

- لطفا پاسخ تمرین خود را در سامانه کوئرا ارسال کنید.
- لطفا از هر منبعی که در گزارش خود استفاده می‌کنید، آن را در پایان همان سوال ذکر کنید.
- فایل‌های مورد نیاز به صورت زیر در دسترس می‌باشند:
  دیتاست مورد بررسی در این لینک قابل دسترسی می‌باشد.
  فایل تمرین نیز در این لینک قابل دسترس می‌باشد.
- نام سند ارسالی شما HW-{homework number}-{Last Name}-{Student number}
- خروجی از هر مرحله تمرین را در گزارش ارسالی خود بارگذاری کنید.

هدف از تعریف این تمرین، آشنایی با الاستیک سرچ و بررسی آن و همچنین بررسی اتصال آن با آپاچی کاساندرا می‌باشد.

Elastic Search یک موتور جستجوی متن باز و توزیع یافته بر پایه Lucence می‌باشد و به شما اجازه می‌دهد تا حجم عظیمی از داده‌ها را ذخیره سازی، جستجو و آنالیز کنید. موتور جستجوی الاستیک سرچ به دلیل نوع بازیابی آن، بسیار سریع بوده و در این موتور جستجو به جای اینکه جستجو به صورت مستقیم و براساس متن ذخیره شده انجام شود، از سیستم ایندکس‌گذاری برای اینکار استفاده می‌شود.
در این تمرین دیتاستی در اختیار شما قرار خواهد گرفت که شامل اطلاعاتی در رابطه با کتاب‌ها می‌باشد که این اطلاعات به صورت زیر بوده و لازم است تا با بررسی این مجموعه داده، تمرین تعریف شده را در گام‌های زیر بررسی کنید:

Books Cataloge dataset items:

Title: title of the book

url: link to the metadata record in Trove

contributurs: pip-seprated names of contributors

date: publication date

format: The type of the work, eg ‘book’ or ‘government publication’, can have multiple values (pip-seprated)

full_text_url: link to the digital version;

trove_id: unique identifire of the digital version

language: mail language of the work

rights: copyright status

pages: number of pages

form: work format, generally one of ‘Book’, ‘Multi volume book’, or ‘Digital publication’

volume: volume/ part number

children: pip-seprated ids of any child work

parent: id of parent work (if any)

text_downloaded: file name of the downloaded OCR text

text_file: True/ False is there any OCRd text

گام اول:

پس از دانلود و اجرای الاستیک سرچ، داده‌های موجود در مجموعه داده در دسترس را در الاستیک سرچ ذخیره کنید. سپس به سوالات زیر پاسخ دهید:

۱) بررسی کنید چند کتاب کودک در این مجموعه جمع آوری شده است و سپس میانگین تعداد صفحات کتاب‌های کودکان را بدست آورید.

۲) در چه سالی بیشترین تعداد کتاب چاپ شده است؟

گام دوم:

الاستیک سرچ و کیبانا با یکدیگر نصب و اجرا می‌شوند. بنابراین داشبوردی در کیبانا طراحی کنید که موارد خواسته شده در زیر را به شما نمایش دهد:

۱) از هر فرمت چند کتاب در این مجموعه داده جمع‌آوری شده است؟

۲) با کمک نمودارها تعداد کتاب‌های موجود به هر زبان را نشان دهید.

گام سوم:

پس از بررسی داده‌های ذخیره شده در الاستیک سرچ، داده‌ها را بر اساس موارد خواسته شده که در ادامه درج شده، در کاساندرا ذخیره کنید.
(لازم است توجه داشته باشید که تکرار داده‌ها یک اصل کاملا پذیرفته شده بوده و بنابراین به دنبال نرمال سازی داده‌ها نباشید.)
در این قسمت دو جدول در کاساندرا ایجاد کنید که بر اساس گروه سنی کتاب‌ها این تقسیم بندی انجام شود، و در هر دو جدول از فیلد trove_id به عنوان کلید اصلی جداول استفاده کنید.

- ✓ سپس یک جدول برای ذخیره سازی اطلاعات کتاب‌های کودک با استفاده از فیلدهای زیر ایجاد کنید:
  نام کتاب، نام نویسندگان، و آدرس اینترنتی در دسترس برای محتوای الکترونیک و نوع فایل در دسترس کتاب (format).
- ✓ جدول دومی برای ذخیره سازی اطلاعات کتاب‌های parent با استفاده از فیلدهای زیر ایجاد کنید:
  نام کتاب، نام نویسندگان، آدرس اینترنتی در دسترس برای محتوای الیکترونیک و نوع فایل در دسترس کتاب (format).

گام چهارم:

به کوئری‌های زیر بر اساس جداول ذخیره شده در گام سوم پاسخ دهید.

۱) میانگین تعداد صفحات کتاب‌های موجود در هر دو جدول را محاسبه کنید.
۲) کتاب‌هایی را نشان دهید که در سال‌های قبل از ۲۰۰۰ چاپ شده اند را نمایش دهد.

نکات مهم:

- ➢ برای آشنایی بیشتر با موتور جستجوی الاستیک سرچ، می‌توانید از لینک زیر برای آشنایی بیشتر، استفاده کنید.
https://vrgl.ir/KoSXA

- ➢ حتما در سند ارسالی خود، نمودارهای ترسیمی و تحلیل‌های خود را درج کنید.
- ➢ تمامی کدهای آماده موجود در ایترنت جمع‌آوری شده است و قطعا کپی کردن شما مشخص می‌شود، بنابراین کپی نکنید.
- ➢ فایل کد خود را در گیت هاب آپلود کرده و لینک آن را در سند ارسالی خود درج کنید.