دانشگاه اصفهان
دانشکده مهندسی کامپیوتر

تمرین اول: فاز یک پروژه سخنرانی‌های TED | **استاد:** دکتر احمد براآنی
درس: بازیابی پیشرفته اطلاعات | **دستیار:** الهام اسماعیلی
نام و نام خانوادگی: سید عمید اسدالهی مجد | **شماره دانشجویی:** ۴۰۰۳۶۱۴۰۰۴

آدرس گیت: https://github.com/amidmajd/ted-talk-classification

نحوه راه‌اندازی

این پروژه از دو بخش کرالر[۱] و الستیک‌سرچ[۲] تشکیل شده‌است. برای راه‌اندازی الستیک‌سرچ از داکر[۳] استفاده شد که یک فایل مخصوص برای راه‌اندازی محفظه[۴]های مورد نیاز به‌نام docker-compose.yml (موجود در گیت‌هاب) ایجاد شد. پس از نصب داکر و واردشدن به پوشه اصلی پروژه، با دستور docker-compose up -d می‌توان الستیک‌سرچ را به‌همراه کیبانا[۵] راه‌اندازی نمود. الستیک‌سرچ در پورت ۹۲۰۰ و کیبانا در پورت ۵۶۰۱ قابل دسترسی هستند. همچنین در این فایل تنظیمات مربوط به امنیت الستیک‌سرچ نسخه ۸ به‌بعد غیرفعال می‌شوند تا مشکلی برای دسترسی‌های بعدی به‌وجود نیاید.

برای اجرای فایل اصلی (ted-talk-indexer.py) ابتدا باید کتاب‌خانه‌های موردنیاز با استفاده از فایل requirement.txt و دستور pip install -r requirements.txt نصب شوند. سایت TED بخش مربوط به زیرنویس‌ها را پس از بارگیری صفحه اصلی سایت بارگیری می‌کند. بنابراین نمی‌توان به‌سادگی با کتاب‌خانه requests سایت را کرال کرد زیرا بخشی از سایت به‌صورت async بارگیری می‌شود که توسط کتاب‌خانه requests قابل دسترسی نیست. بنابراین با استفاده از کتاب‌خانه selenium که یک کتاب‌خانه پیشرفته در زمینه کرال کردن وبسایت‌ها می‌باشد می‌توان بخش به بخش به زیرنویس‌ها در فایل html وبسایت که به‌طور کامل بارگیری شده، دست یافت. درنتیجه باید دو متغیر CHROME_PATH و CHROME_DRIVER_PATH به‌طور مناسب مقداردهی شوند زیرا این دو متغیر برای استفاده از کتاب‌خانه selenium ضروری می‌باشند. مقدار متغیر CHROME_PATH باید آدرس فایل اجرایی[۶] مرورگر گوگل‌کروم[۷] (مثلا "C:\Program Files\Google\Chrome\Application\chrome.exe") باشد. مقدار متغیر CHROME_DRIVER_PATH نیز باید آدرس درایور مرورگر کروم باشد. این درایور از این سایت قابل دریافت است که نسخه این درایور باید مشابه ورژن مرورگر کروم باشد. پس از دانلود این درایور متناسب با سیستم‌عامل، می‌توان آن را کنار فایل پایتون[۸] پروژه کپی کرد و آدرس آن را تنظیم نمود. (مثلا "./chromedriver.exe")

[1] Crawler
[2] Elasticsearch
[3] Docker
[4] Container
[5] Kibana
[6] Executable
[7] Google-Chrome
[8] Python

فایل ted-talk-indexer.py

با اجرای این فایل (پس از ورود به پوشه src) ابتدا تمام زیرنویس‌ها از سایت TED دانلود می‌شوند و سپس ایندکس[9] موردنظر در الستیک‌سرچ ساخته می‌شود. برای خواندن فایل CSV و ایجاد تغییرات برروی آن، از کتاب‌خانه Pandas استفاده شده‌است که یک کتاب‌خانه بسیار معروف در زمینه داده‌کاوی می‌باشد. در ابتدا داده‌ها توسط کتاب‌خانه Pandas خوانده‌شده و به یک دیتافریم[10] تبدیل می‌شوند؛ سپس تنظیمات مربوط به مرورگر بدون محیط گرافیکی[11] که توسط کتاب‌خانه selenium ایجاد می‌شود انجام می‌شود. سپس با استفاده از کتاب‌خانه موازی‌سازی درونی پایتون (concurrent) یک استخر شامل ۵ کارگر[12] ساخته می‌شود.

در ابتدای این بخش (خط ۷۹) ابتدا صفی شامل لینک‌ها که باید تابع get_transcript برروی آن‌ها اعمال شود ساخته می‌شود و به کتاب‌خانه موازی‌ساز داده می‌شوند. سپس با فراخوانی as_completed برروی هریک از عناصر این صف، می‌توان عملی را به‌هنگام پایان کار هریک از عناصر صف (در اینجا دریافت زیرنویس مربوط به هر لینک با استفاده تابع get_transcript) انجام داد. به‌هنگام پایان کار هر لینک، سعی می‌شود تا مقدار فیلد جدید transcript برای هر سخنرانی در دیتافریم برابر با زیرنویس دریافت شده از آن لینک (خروجی تابع get_transcript) قرار داده‌شود. در انتهای این بخش دیتافریم جدید شامل فیلد زیرنویس برای استفاده‌های بعدی ذخیره نیز می‌شود.

در بخش دوم از برنامه یک ارتباط جدید با الستیک‌سرچ راه‌اندازی شده با استفاده از کتاب‌خانه elasticsearch در پایتون برقرار می‌شود. متغیر es_indices_client نیز برای استفاده از آنالیزرهای داخلی الستیک‌سرچ مورد استفاده قرار خواهد گرفت. در ادامه در خط ۱۰۰ از برنامه، به‌ازای هر سخنرانی که مقدار فیلد زیرنویس آن خالی نباشد ابتدا با استفاده از آنالیزر stop، عمل نرمال‌سازی انجام می‌شود. در این عملیات نرمال‌سازی stop wordsها و علائم سجاوندی حذف می‌شوند و تمام کلمات به حالت حروف کوچک تبدیل می‌شوند و هر سخنرانی به‌صورت توکن‌شده برگردانده می‌شود. برای برخی از سخنرانی‌ها زیرنویسی در سایت TED وجود ندارد که مقدار زیرنویس آن‌ها در دیتافریم برابر با Null قرار گرفت و در الستیک‌سرچ نیز وارد نشدند.

پس از چسباندن این توکن‌ها به‌هم، سخنرانی نرمال‌شده به ایندکس ساخته‌شده اضافه می‌شود. به‌هنگام اضافه کردن داده به ایندکس جدید اگر آن ایندکس وجود نداشته باشد، توسط الستیک‌سرچ ساخته می‌شود. در نهایت یک نمونه داده از ایندکس ساخته‌شده دریافت می‌شود تا به عنوان نمونه به کاربر نمایش داده‌شود.

بخش اول از این پروژه، یعنی کرال‌کردن، به دلیل زیاد بودن حجم داده‌ها، در محیط google colab اجرا شده‌است که فایل ipython notebook مربوطه نیز در پوشه src در گیت‌هاب قرار دارد. این بخش هم‌چنان کاملا قابل اجرا برروی سیستم شخصی نیز می‌باشد (اگر اینترنت یاری کنه!). دیتافریم جدید شامل زیرنویس‌ها نیز در گیت‌هاب ذخیره شده‌است.

---

[9] Index
[10] Data frame
[11] Headless
[12] Worker

تابع get_transcript

در این تابع شماره سخنرانی و لینک مربوط به آن به عنوان ورودی دریافت می‌شوند و شماره سخنرانی و زیرنویس آن به‌عنوان خروجی برگردانده می‌شوند. به هر لینک transcript/ نیز اضافه می‌شود تا بخش زیرنویس سخنرانی بدون نیاز به کلیک برروی دکمه مربوط به آن در سایت، بارگیری شود. سپس یک مرورگر بدون محیط گرافیکی ساخته می‌شود و سعی می‌شود تا فایل HTML کامل هر آدرس دریافت شود. درواقع کتاب‌خانه selenium با ورود آدرس در این مرورگر بدون محیط گرافیکی به‌طور کامل تمام بخش‌های سایت را دریافت می‌کند و سپس فایل HTML نهایی را برمی‌گرداند.

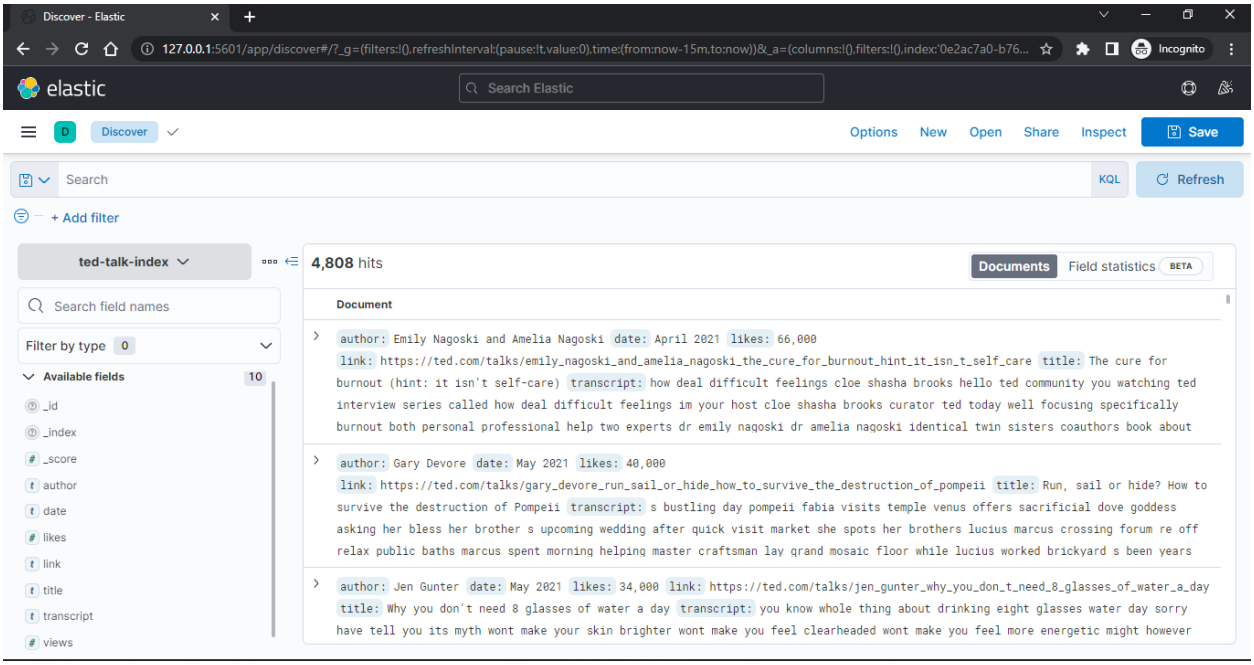
سپس با استفاده از کتاب‌خانه beautifulsoap فرمت فایل HTML به‌حالت مناسب تبدیل می‌شود و با استفاده از نام خاص کلاس‌های CSS مربوط به خطوط زیرنویس سعی می‌شود تا تمام بخش‌های زیرنویس دریافت شوند. در نهایت تمام بخش‌های زیرنویس که در یک آرایه قرار دارند باهم ترکیب می‌شوند و همراه با شماره آن سخنرانی برگردانده می‌شوند. این تابع درصورت عدم وجود زیرنویس برای یک سخنرانی، شماره آن سخنرانی را به همراه مقدار Null برمی‌گرداند.

تصاویری از نتایج اجرای برنامه

	title	author	date	views	likes	link	transcript
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	December 2021	404000	12000	https://ted.com/talks/ozawa_bineshi_albert_cli...	Yuchi F'as English Good afternoon I come from ...
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	February 2022	214000	6400	https://ted.com/talks/sydney_iaukea_the_dark_h...	It was January 16th 1895 Two men arrived at Li...
2	How play can spark new ideas for your business	Martin Reeves	September 2021	412000	12000	https://ted.com/talks/martin_reeves_how_play_c...	Have you ever trodden on a Lego brick in your ...
3	Why is China appointing judges to combat clima...	James K. Thornton	October 2021	427000	12000	https://ted.com/talks/james_k_thornton_why_is_...	Imagine a world in which China was an environm...
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	October 2021	2400	72	https://ted.com/talks/mahendra_singhi_cement_s...	None
...	...	...	...	...	...	...	...
5435	The best stats you've ever seen	Hans Rosling	February 2006	15000000	458000	https://ted.com/talks/hans_rosling_the_best_st...	About 10 years ago I took on the task to teach...
5436	Do schools kill creativity?	Sir Ken Robinson	February 2006	72000000	2100000	https://ted.com/talks/sir_ken_robinson_do_scho...	Good morning How are you Audience Good Its bee...
5437	Greening the ghetto	Majora Carter	February 2006	2900000	88000	https://ted.com/talks/majora_carter_greening_t...	If youre here today and Im very happy that you...
5438	Simplicity sells	David Pogue	February 2006	2000000	60000	https://ted.com/talks/david_pogue_simplicity_s...	Music The Sound of Silence Simon amp Garfunkel...
5439	Averting the climate crisis	Al Gore	February 2006	3600000	109000	https://ted.com/talks/al_gore_averting_the_cli...	Thank you so much Chris And its truly a great ...

5440 rows × 7 columns

دیتافریم نهایی شامل زیرنویس سخنرانی‌ها



سخنرانی‌های اضافه‌شده به ایندکس به همراه زیرنویس