

بِسْمِ تَعَالٰی

گزارش پروژه

درس بازیابی اطلاعات

استاد امیرخانی

علی، عموری ۹۸۱۳۲۰۰۰۵۴

هدف پروژه:

قصد ما در این پروژه‌ی ساده این است که در ابتدا باید سایت انتخابی را crawl کنیم،

<https://f1iran.com>

سایت انتخابی:

بدین منظور لینک های URL ، Title و Text این سایت خبری را از آرشیو استخراج می کنیم و از آن به صورت یک فایل json خروجی می گیریم.

در مرحله آخر باید یک query به عنوان پارامتر ورودی وارد کنیم تا TF-IDF و مقدار شباهت (similarities) را بیابیم.

لینک پروژه در git :

<https://github.com/AliAmoori/main.git>

code:

فایل F1Iran.py شامل کد crawling می باشد.

تابع `scraping` را به این منظور می‌سازیم. مقدار `page` را برابر ۵۴۰ وارد می‌کنیم. (به دلیل صفحات زیاد سایت، صفحات ۵۴۰ به بعد `crawl` شده است.)

موارد مورد نظر سایت در تگ `h2` قرار دارند. تگ های شامل `href` را برای `url` ها جداسازی می کنیم. برای `Title` و `Text` با استفاده از کتابخانه `Article` ، بعد از دانلود و `parse`، خروجی را – به منظور استفاده برای کد `TFIDF` – در فایل `json` قرار

می‌دهیم. کتابخانه pandas به همین دلیل add شده است.

نمونه ای از خروجی فایل در json:

"title": "شُد؟ - فرمول یک ایران DTM برند اشنایدر چطور بهترین راننده نارنجی",
"text": "**فصل دوم علاقه اشنایدر**
تی ام به حساب می آید. من برای هیچ فصلی به اندازه فصل ۲۰۰۸ خودم را (برای کسب قهرمانی) آماده نکرده بودم
بهترین خودروی برند اشنایدر
به آن را بفروشد. اما حداقل در رویداد های کلاسیک هر از چندگاهی آنها به من اجازه رانندگی با آن را می دهند
سر سخت ترین رقیبی برند اشنایدر
به یاد داشت که آنها حتی (قهرمان جهان دو دوره از فرمول یک) یعنی میکا مکتین را نیز به خدمت گرفته بودند"
url": "https://firilan.com/149154/xd8xa8dx8a1kxd90x86xd8xaf-xd8xa7xd8xb4%ld9d8x6xd8xa7xd8xb8cxd8xa7xd8xb1-1xd8xb8"

نکته:

- شرط stop تابع repetitive می باشد که نمونه ای از آن را در کلاس مشاهده کردیم.

- سایت انتخابی، فیلتر سال ندارد و فقط مطابق با پارامتر page قابل فیلتر است.

فایل TF-IDF.py شامل کد TF-IDF می باشد که query در اینجا پاس داده می شود.

(نکته: k در sorting به صورت بیش فرض برابر ۱۰ است.)

با وارد کردن query خروجی را به صورت کامل می‌توانید مشاهده کنید.

نمونه: با وارد کردن کلمه کلیدی "تکنولوژی" این نمونه خبر را که مربوط به تکنولوژی است را مشاهده می‌کنیم. کلمه "تکنولوژی" در این متن خبری نیست؛ اما موضوع خبر، فناوری و تکنولوژی ماشین‌های بدون سرنشین در فرمول یک می‌باشد. (به دلیل درهم‌ریختن کلمات در ترمینال، آن را به صورت کامنت در کد قرار دادم. خروجی را با اجرای دوباره می‌توانید مشاهده کنید.)

'یکی از هزاران استفاده از اینترنت نسل پنجم طراحی ماشین‌های بدون سرنشین است. 0.3129232850921755'