بسمه تعالی

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،

سایت انتخابی: https://f1iran.com

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

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

لینک پروژه درgit :

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

code:

فایل F1Iran.py شامل کد crawling می‌باشد.

تابع scraping را به این منظور می‌سازیم. مقدار page را برابر ۵۴۰ وارد می‌کنیم. (به دلیل صفحات زیادِ 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 می‌کنیم. 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:

نکته:

- شرط stop تابع repetitive می‌باشد 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 می‌باشد که query در اینجا پاس داده می‌شود.

(نکته : k در sorting به صورت پیش‌فرض 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 متنِ خبری نیست؛ اما موضوعِ خبر،فناوری و تکنولوژی ماشین های 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'
```