بسمه تعالی

پروژه درس بازیابی اطلاعات  
سید محمد سجاد حجازی

9923115037

https://github.com/Zi-zo/IR-Proj

Part 1

Chosen Site is Mojanews.com

در ابتدا یک لینک از ارشیو صفحات خبری سایت میخوانیم و فرمت آن را مورد بررسی قرار می دهیم:

[https://www.mojnews.com/newsstudios/archive/?curp=1&order=order\_time&offset%5Bseen%5D=434800&offset%5Btime%5D={year}-06-02+{page}%3A30%3A58&dir=0#gsc.tab=0](https://www.mojnews.com/newsstudios/archive/?curp=1&order=order_time&offset%5Bseen%5D=434800&offset%5Btime%5D=%7byear%7d-06-02+%7bpage%7d%3A30%3A58&dir=0#gsc.tab=0)

وقتی به لینک یک خبر دیگر میرویم، متوجه تغییر در قیمت عدد پیج یا سال خبر منتشر شده یا غیره ... میشویم.

Part 2

هر خبر، تیتر Article دارد.

پس از بررسی های انجام شده در قسمت Source، مشخص میشود این سایت از تگ h2، برای نمایش اخبار استفاده کرده است..

در ادامه قسمت متغیر Article هارا به همراه قسمت اول ادرس سایت متصل میکنیم.

for link in tqdm(links):

news\_url = 'https://www.mojnews.com/' + link.a['href']

url\_list.append(link.a['href'])

بدین شکل میتوان خبر ها را استخراج و پارس کرد.

Part 3

فایل خروجی در یک فایل به فرمت CSV ذخیره میشود.

از کتابخانه pandas استفاده میشود و Word ها از Doc ها به متغیر Vocab ریخته میشود.

Part 4

از کلاس TfIdf-vectorizer یک Obj میسازیم Doc هارا بر آن Transform میکنیم که باعث میشود لغات را یاد گرفته و

میتوان بعد از ان، Tf-Idf را محاسبه کرد

حال ما میتوانیم فاصله Cosine لغات را محاسبه کنیم.

Part 5

در این مرحله، باید لغات یا شبه جمله ای که برای مقایسه شباهت نیاز داریم، قرار بدهیم.

یک کوئری میسازیم.

نمره tf-idf آن را محاسبه می کنیم.

حال tf-idf یک به یک مقایسه شده و در آخر به ترتیب نمرات شباهت ها، مرتب می شوند.