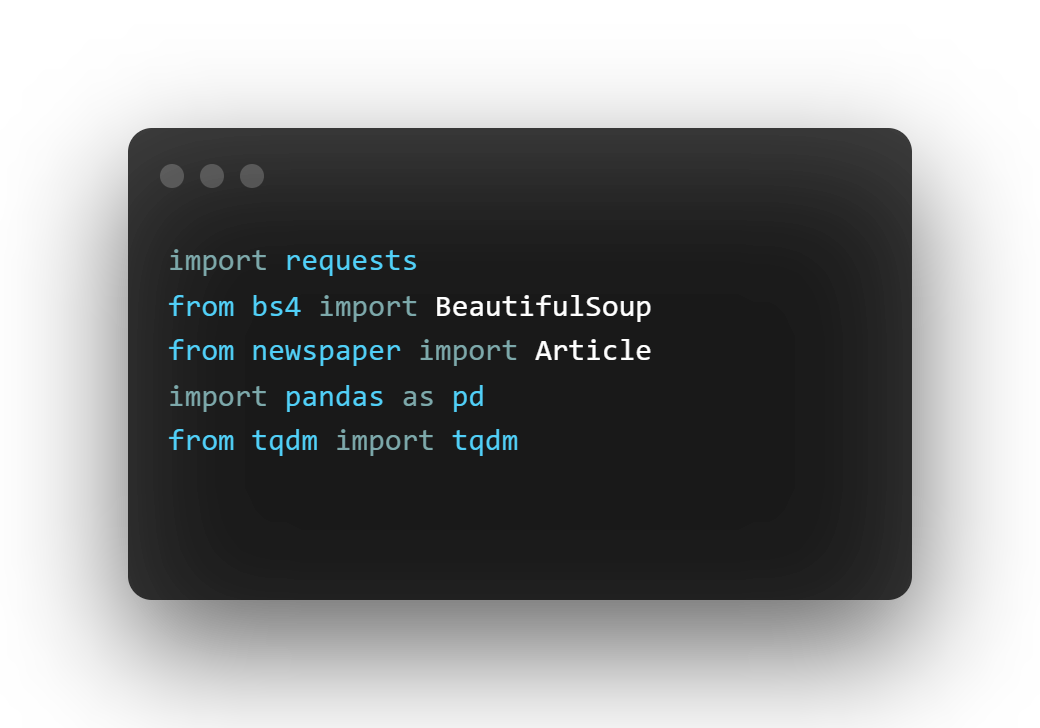
**بسم الله الرحمن الرحیم**

علیرضانظری

9713200080در این قسمت کتابخانه های مورد نیاز خودمان را import میکنیم .

کتابخانه ها:

Request: برای دریافت html صفحات .

Beautiful soup: برای استخراج تگ های Html صفحات استفاده میشود .

Newspaper: ابجکت artile برای parse کردن صفحات url های بدست آمده.

با استفاده از pandas دیتا های بدست آمده را روی فایل مورد نظر مینویسیم .



آبجکت csv برای خواندن فایلهای اکسل استفاده میشود و همچنین TfidfVictorizer برای برداری کردن داده ها استفاده میشود هم داکیومنت ها و هم کوئری ها.

Cosine similirity برای مقایسه کوئری های داکیومنت ها استفاده میشود .

Tqdm یک نوار ابزاری برای کارهای زمانبندی است .

Numpy یک کتابخانه برای استفاده از داده های عددی و همچنین array ها است .

در ابتدا فایل csv را باز کرده و با csv-reader دیتای آنرا میخوانیم .

متغیر docs را تعریف کرده و با استفاده از یک حلقه داده های خوانده شده راتوسط csv-reader را داخل docs میریزیم .

اول یک متغیرvictorizer از کلاس tfidfVictorizer درست میکنیم و سپس متغیر docs را توسط victorizer به روی داده fit میکنیم.

Tfidf ِ docs هارا با استفاده از متد fit\_transform بدست می آوریم.

Victorizer.vocabulary را میسازیم و type و shape آن را تعریف میکتیم.

Query processing با استفاده از victorizer transform میتوانیم tfidf کوئری را بدست بیاوریم.

متغیر coaine برای اینکه کوسینوس کوئری با تمام داکیومنت هارا نگه داریم .و با استفاده از یک for این مقایسه را انجام میدهیم و کوسینوس بدست آمده را در cosine نگه میداریم و با استفاده از numpy میاییم k داکیومنت با بیشترین کسینوس را برمیگردانیم