باسمه تعالی

درس: بازیابی اطلاعات

نام و نام خانوادگی: تینا توکلی -هادی امینی

شماره دانشجویی: 9922762220 - 9912762370

شماره تمرین : 01

---

1. در ابتدا کتابخانه های مناسب را import میکنیم .

```
pip install numpy pandas scikit-learn nltk
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import string
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
nltk.download('stopwords')
nltk.download('punkt')
```

2. برای استفاده از دیتاست دستورات زیر را اجرا می کنیم.

```
!unzip '/content/FUM_IR_1402-02_HW#1_Resources.zip'
dataset_path='/content/FUM_IR_1402-02_HW#1_Resources'
```

3. در تابع load_dataset(path) مراحل زیر را انجام می دهیم:
- مسیر مربوط به دیتاست و کوئری و همچنین کوئری جدید ساخته شده را در متغیری می ریزیم.
- برای هر کدام از آنها ،در حلقه ای محتویات فایل را خوانده و در لیستی ذخیره میکنیم
- یک دیکشنری شامل این 3 مورد را برمی گردانیم.

4. تابع load_relevace(folder_path) مراحل زیر را انجام میدهیم:
- برای فایل relevanceاست که محتویات داخل آن را خوانده و split میکنیم .
- یک دیکشنری میسازیم که برای هر کوئری، اسنادی مرتبط را مشخص می کنیم.

5. در تابع preprocess_document(doc) مراحل زیر را برای پیش پردازش انجام می دهیم:
- در ابتدا تمام حروف رو تبدیل به حروف کوچک می کنیم .
- جداسازی title و text با استفاده ازحذف SGML tags .
- حذف اعداد و punctuation marks ها .
- توکن سازی با استفاده از تابع آماده word_tokenize(doc) انجام می دهیم.

- تابع اماده stopwords.words('english')) لیست stopwordsها را به ما می دهد،آنها را حذف میکنیم .
- با استفاده از الگوریتم porter ،توکن های باقی مانده را ریشه یابی می کنیم تا در آخر توکن ها را به ما برگرداند.

6. ماتریس tf.idf را با استفاده از TfidfVectorizer()می سازیم.
7. در تابعpreprocess_query(query) مراحل زیر را انجام میدهیم :

تمام مراحل پیش پردازش در کوئری مانند بخش 5 (به غیر از حذف tags) انجام میشود.

8. در تابع related_docs(n,queries=dataset['queries']) مراحل زیر را انجام می دهیم:

- در ابتدا برای تمام کوئری ها بخش 6و7 را انجام میدهیم .
- با استفاده از cosine_similarity بین هر کوئری و تمام اسناد،میزان شباهت آنها را محاسبه میکنیم.
- سپس آنها را به صورت نزولی مرتب میکنیم و n تای برتر را ذخیره می کنیم
- براساس ایندکس ها id و میزان شباهت را ذخیره میکنیم.
- و سپس نتایج را به فرمت (query_id,doc_id,Score) نمایش می دهیم.

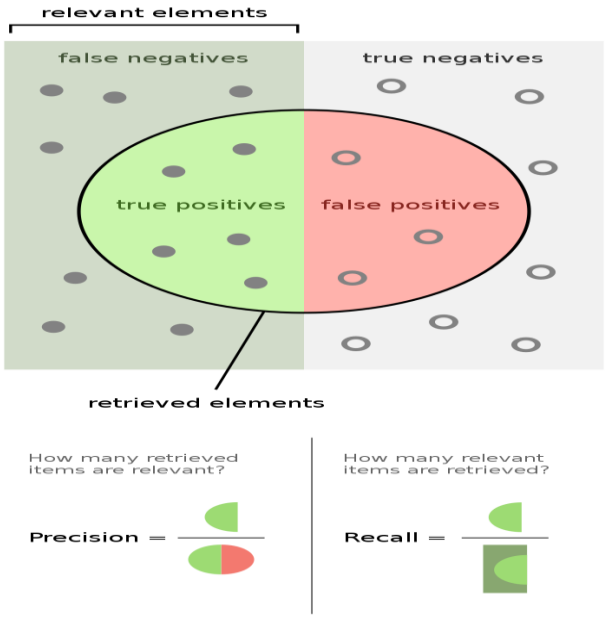
9. در تابع calculate_metrics(top_lists) مراحل زیر را انجام می دهیم:

- محاسبه precision و recall با استفاده از مقایسه بین لیست اسناد مرتبط به دست آمده به یک کوئری و لیست اسناد مرتبط داده شده برای همان کوئری صورت میگیرد.
- محاسبه میانگین precision و recall انجام می شود.

تعاریف:

Presision: به نسبتی از نتایج جستجو که در واقع مرتبط با پرس و جو کاربر هستند، اشاره دارد .به عبارت دیگر، نشان می‌دهد که چه تعداد از نتایج بازیابی شده مفید و مرتبط هستند.

Recall: به نسبتی از اسناد مرتبط در مجموعه اسناد که توسط سیستم بازیابی شده‌اند، اشاره دارد .به عبارت دیگر، نشان می‌دهد که سیستم چه درصدی از اسناد مرتبط را پیدا کرده است.



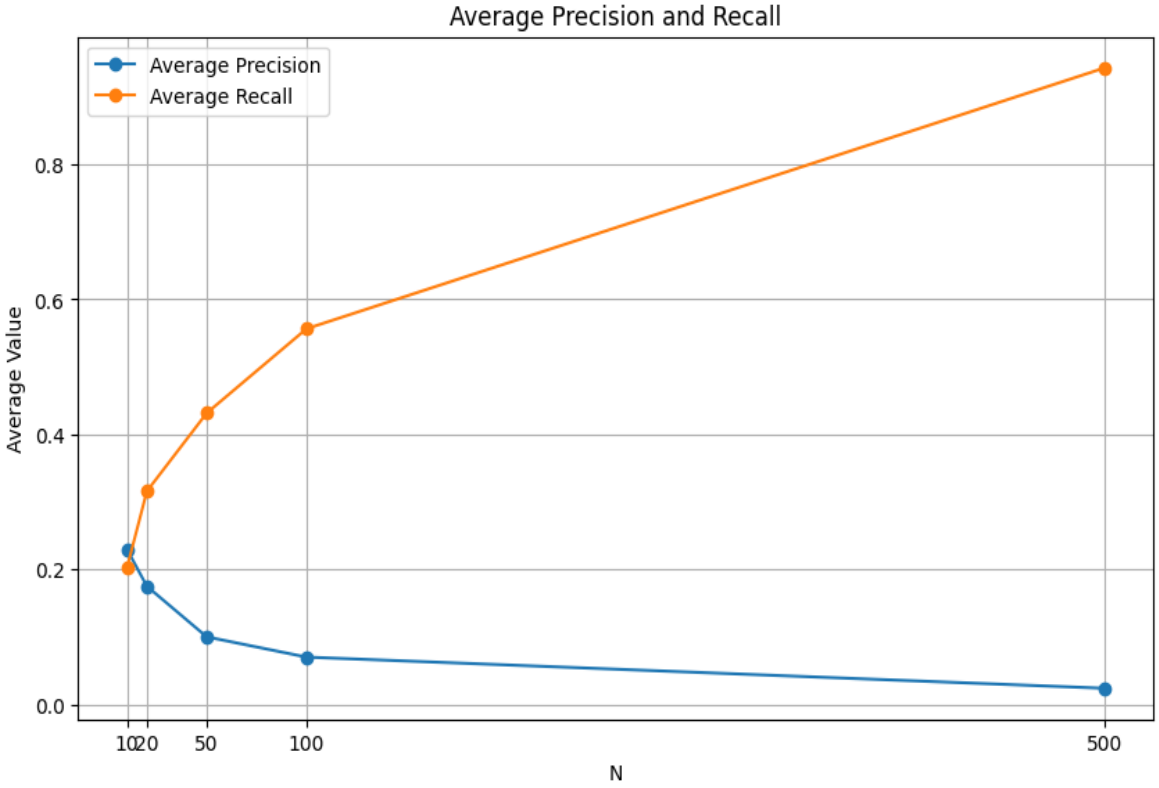
خروجی:

{10: {'precision_average': 0.23000000000000004,

```
 'recall_average': 0.20190058479532164,
 'queries': {1: (0.0, 0.0),
  2: (0.2, 0.13333333333333333),
  3: (0.3, 0.2),
  4: (0.1, 0.05555555555555555),
  5: (0.3, 0.15789473684210525),
  6: (0.4, 0.2222222222222222),
  7: (0.6, 0.6666666666666666),
  8: (0.2, 0.5),
  9: (0.0, 0.0),
  10: (0.2, 0.08333333333333333)}},
20: {'precision_average': 0.175,
 'recall_average': 0.3157748538011696,
 'queries': {1: (0.0, 0.0),
  2: (0.2, 0.26666666666666666),
  3: (0.25, 0.3333333333333333),
  4: (0.1, 0.1111111111111111),
  5: (0.2, 0.21052631578947367),
  6: (0.2, 0.2222222222222222),
  7: (0.4, 0.8888888888888888),
  8: (0.15, 0.75),
  9: (0.1, 0.25),
  10: (0.15, 0.125)}},
50: {'precision_average': 0.1,
 'recall_average': 0.43157894736842106,
 'queries': {1: (0.0, 0.0),
  2: (0.14, 0.4666666666666667),
  3: (0.16, 0.5333333333333333),
  4: (0.04, 0.1111111111111111),
  5: (0.12, 0.3157894736842105),
  6: (0.12, 0.3333333333333333),
  7: (0.16, 0.8888888888888888),
  8: (0.06, 0.75),
  9: (0.12, 0.75),
  10: (0.08, 0.16666666666666666)}},
100: {'precision_average': 0.06999999999999999,
 'recall_average': 0.5564766081871345,
 'queries': {1: (0.0, 0.0),
  2: (0.11, 0.7333333333333333),
  3: (0.12, 0.8),
  4: (0.04, 0.2222222222222222),
  5: (0.13, 0.6842105263157895),
  6: (0.08, 0.4444444444444444),
  7: (0.08, 0.8888888888888888),
  8: (0.03, 0.75),
  9: (0.07, 0.875),
  10: (0.04, 0.16666666666666666)}},
500: {'precision_average': 0.0238,
 'recall_average': 0.9430555555555555,
 'queries': {1: (0.002, 1.0),
  2: (0.03, 1.0),
  3: (0.03, 1.0),
  4: (0.032, 0.8888888888888888),
```

```
5: (0.038, 1.0),
6: (0.03, 0.8333333333333334),
7: (0.018, 1.0),
9: (0.016, 1.0),
10: (0.034, 0.7083333333333334)}}}
```

محاسبه دقت و بازیابی برای کوئری های متفاوت به ازای nهای خواسته شده



تصویر 1- نمودار دقت و بازیابی به ازای nهای خواسته شده

توضیحات مرتبط با نمودار:
این نمودار میانگین دقت و میانگین بازیابی را در مقابل تعداد n نشان می‌دهد.
تحلیل نمودار:
1. میانگین دقت:
- دقت با نقاط آبی نشان داده شده است.
- با افزایش تعداد n، میانگین دقت کاهش می‌یابد.
- در مقادیر کمتر n (بین 10 تا 20)، میانگین دقت به حدود 0.2 می‌رسد و سپس با افزایش n به تدریج کاهش می‌یابد تا نهایتاً به حدود 0 در n=500 برسد.

2. میانگین بازیابی:
- بازیابی با نقاط نارنجی نشان داده شده است.
- با افزایش تعداد n، میانگین بازیابی افزایش می‌یابد.
- در مقادیر کمتر n، میانگین بازیابی حدود 0.2 است و سپس با افزایش n به سرعت افزایش می‌یابد و نهایتاً در n=500 به حدود 0.8 می‌رسد.

بنابراین :
- رابطه دقت و بازیابی: با افزایش تعداد n، میانگین دقت کاهش و میانگین بازیابی افزایش می‌یابد. این نشان‌دهنده یک مبادله (trade-off) بین دقت و بازیابی است؛ یعنی با افزایش تعداد داده‌ها، سیستم توانایی پیدا کردن موارد بیشتر (بازیابی) را دارد اما دقت آن کاهش می‌یابد.
-به طور کلی، این نمودار نشان می‌دهد که با افزایش تعداد n، سیستم تمایل به بازیابی بیشتر داده‌ها دارد اما دقت آن کاهش می‌یابد.

کوئری های نوشته شده برای دیتاست:

a. What are the comparative efficiencies of ballistic, glide, and skip vehicles in converting velocity into range for hypervelocity flight?
b. How do turbulence-intensity profiles vary between the wing and forward-fuselage section of a fighter jet?
c. What is the solution for the temperature gradient in the slip region of a moving rarefied gas
d. What were the key findings regarding boundary layer transition and skin friction on an insulated flat plate in a hypersonic wind tunnel at Mach 5.8?
e. What was the purpose of the experimental study of a wing in a propeller slipstream

برای تحلیل این کوئری ها همانند بخش 8از تابع related_docs(n,queries=dataset['my_query'] بهره برده ایم تا اسناد مرتب شده مرتبط به هر کوئری را بیابیم.
نمونه ای از خروجی کوئری های نوشته شده برای n=3 (برای تمام nهای خواسته شده در صورت تمرین نیز انجام شده و در کد مشاهده میشود).

N=3

(query_id:1,doc_id:0077, Score: 0.6033222356713115)

(query_id:1,doc_id:1379, Score: 0.377105810167954)

(query_id:1,doc_id:0598, Score: 0.18377891540771252)

(query_id:2,doc_id:0076, Score: 0.3865098235319819)

(query_id:2,doc_id:0218, Score: 0.2862240759983179)

(query_id:2,doc_id:0997, Score: 0.21690629619171814)

(query_id:3,doc_id:0022, Score: 0.5582857058748145)

(query_id:3,doc_id:0326, Score: 0.2826658881980399)

(query_id:3,doc_id:0550, Score: 0.28232328281777436)

(query_id:4,doc_id:0568, Score: 0.42366096779144896)

(query_id:4,doc_id:0009, Score: 0.3855276825066247)

(query_id:4,doc_id:0525, Score: 0.28827515953457006)

(query_id:5,doc_id:0453, Score: 0.5041313990469968)

(query_id:5,doc_id:0001, Score: 0.45986543182688294)

(query_id:5,doc_id:1144, Score: 0.43338965199499146)

گزا

	تینا توکلی	هادی امینی
مشارکت بخش یک	10	10
مشارکت بخش دو	10	10