گزارش بخش امتیازی پروژه بازیابی اطلاعات

نام: صبا
نام خانوادگی: سحبان
شماره دانشجویی: ۹۹۳۱۰۹۶

آدرس لینک google colab notebook پروژه:
https://colab.research.google.com/drive/1sIgrJXrGOJD_DItTkX7fFgA4IscyYhT_?usp=sharing
آدرس لینک google drive directory خروجی هر فاز از پروژه:
https://drive.google.com/drive/folders/1a-d6UODHfYYMVc9gDXwJeHzy_xBd-jxs?usp=sharing

موارد تحویلی در گزارش

توجه شود که در هر مرحله، علاوه بر درج مقادیر خواسته شده در گزارش، تصویر کد مربوطه تکلیف خواسته شده، توضیح مختصری از کد، تصویر خروجی مورد نظر با اجرای کد و آدرس file یا directory ذخیره نتایج (در صورت نیاز به ذخیره) در گزارش ذکر شود.

کوئری انتخاب شده فازها = بهترین بازیکن فوتبال جهان

فاز ۱: تعبیه برت

https://drive.google.com/drive/folders/1kgBLK3mbUPAiDTxCV_asKhenbTjVHt6q?usp=drive_link

الف) در این مرحله مشخص کنید که از چه مدل زبانی‌ای برای محاسبه تعبیه برت استفاده کرده اید.

نام مدل زبانی:ParsBERT

لینک مدل زبانی:

https://huggingface.co/HooshvareLab/bert-fa-zwnj-base

ب) سه جمله زیر را به مدل زبانی انتخابی خود بدهید، تعبیه جمله آن سه را گرفته و بین هر دوتایی از آن‌ها شباهت کسینوسی را حساب کنید.

```python
from sklearn.metrics.pairwise import cosine_similarity

# Sentences
sentence_1 = "هنگام طلوع خورشید، پرندگان با آوازهای دلنشین خود روز را آغاز می‌کنند."
sentence_2 = "سپیده دم، مرغان با نغمه‌های دلپذیرشان شروع صبح را نوید می‌دهند."
sentence_3 = "طلوع خورشید همراه با صدای دلنشین پرندگان بهترین زمان روز است."

# Generate embeddings
embedding_1 = model.encode(sentence_1)
embedding_2 = model.encode(sentence_2)
embedding_3 = model.encode(sentence_3)

# Calculate cosine similarities
similarity_1_2 = cosine_similarity([embedding_1], [embedding_2])[0][0]
similarity_1_3 = cosine_similarity([embedding_1], [embedding_3])[0][0]
similarity_2_3 = cosine_similarity([embedding_2], [embedding_3])[0][0]

# Print results
print("\nCosine Similarities:")
print(f"Similarity (Sentence 1 & 2): {similarity_1_2:.4f}")
print(f"Similarity (Sentence 1 & 3): {similarity_1_3:.4f}")
print(f"Similarity (Sentence 2 & 3): {similarity_2_3:.4f}")
```

```
Cosine Similarities:
Similarity (Sentence 1 & 2): 0.7586
Similarity (Sentence 1 & 3): 0.7291
Similarity (Sentence 2 & 3): 0.6485
```

در این کد ابتدا با استفاده از مدل زبانی‌مان و sentence transformer امبدینگ کلمات را به دست می‌آوریم و شباهت را حساب میکنیم. اینکه جمله اول و دوم بیشترین شباهت را دارد نشانه درست کار کردن است چون از نظر tf-idf شباهت زیادی ندارند ولی معنایی بسیار شبیهند.

شباهت کسینوسی تعبیه برت جمله اول و دوم: ۰.۷۵۸۶

شباهت کسینوسی تعبیه برت جمله دوم و سوم: ۰.۷۲۹۱

شباهت کسینوسی تعبیه برت جمله اول و سوم: ۰.۶۴۸۵

ج) مانند مرحله قبل، سه جمله ارائه دهید که شباهت معنایی جمله اول و دوم بیشتر از شباهت معنایی جمله اول و سوم باشد، اما به نظر برسد که tf-idf شباهت بیشتری بین جمله اول و سوم نسبت به جمله اول و دوم محاسبه خواهد کرد. سپس مانند بخش ب، شباهت کسینوسی هر دو جمله را حساب کنید.

```python
# Sentences
sentence_1 = "I love programming."
sentence_2 = "Coding is my passion."
sentence_3 = "I think that programming involves debugging"

# Generate embeddings
embedding_1 = model.encode(sentence_1)
embedding_2 = model.encode(sentence_2)
embedding_3 = model.encode(sentence_3)

# Calculate cosine similarities
similarity_1_2 = cosine_similarity([embedding_1], [embedding_2])[0][0]
similarity_1_3 = cosine_similarity([embedding_1], [embedding_3])[0][0]
similarity_2_3 = cosine_similarity([embedding_2], [embedding_3])[0][0]

# Print results
print("\nCosine Similarities:")
print(f"Similarity (Sentence 1 & 2): {similarity_1_2:.4f}")
print(f"Similarity (Sentence 1 & 3): {similarity_1_3:.4f}")
print(f"Similarity (Sentence 2 & 3): {similarity_2_3:.4f}")
```

```
Cosine Similarities:
Similarity (Sentence 1 & 2): 0.7956
Similarity (Sentence 1 & 3): 0.7685
Similarity (Sentence 2 & 3): 0.7379
```

جمله اول:

I love programming

جمله دوم:

Coding is my passion

جمله سوم:

I think that programming involves debugging

شباهت کسینوسی تعبیه برت جمله اول و دوم: 0.7956

شباهت کسینوسی تعبیه برت جمله دوم و سوم: 0.7685

شباهت کسینوسی تعبیه برت جمله اول و سوم: 0.7379

د) کد ذخیره هر سند به شکل خواسته شده زیر در یک directory به نام phase_1_result را نمایش دهید.
هر سند به صورت فایلی با نام به صورت زیر ذخیره می‌شود:

{doc_id}.json

و محتوای آن نیز به صورت زیر است:

{

“doc_id”: DOCUMENT_ID,

“content”: CONTENT,

“embedding”: [//embedding entries]

}

که DOCUMENT_ID شناسه سند، CONTENT محتوای سند، و embedding بردار تعبیه سند مورد نظر است.
تعبیه برت را از مدل زبانی مورد استفاده خود به دست آورید.

```
import json
import os

dataset_path = 'IR_bonus_dataset.json'
with open(dataset_path, 'r', encoding='utf-8') as f:
    dataset = json.load(f)

print("Sample Document:", dataset["0"])
```

```python
for doc_id, doc_data in dataset.items():
    content = doc_data["content"]

    embedding = model.encode(content).tolist()

    output_file = os.path.join(output_dir, f"{doc_id}.json")
    with open(output_file, 'w', encoding='utf-8') as f:
        json.dump({
            "doc_id": doc_id,
            "content": content,
            "embedding": embedding
        }, f, ensure_ascii=False, indent=4)

    print(f"Processed doc_id: {doc_id}, Saved to: {output_file}")
```

در ابتدا، با استفاده از یک حلقه، تمام اسناد موجود در دیتاست پیمایش می‌شوند. هر سند دارای یک شناسه منحصربه‌فرد (doc_id) و محتوای متنی (content) است که از دیتاست استخراج می‌شود. سپس مدل زبانی که قبلاً بارگذاری شده است، با استفاده از متد encode تعبیه متن سند را به‌صورت بردار عددی محاسبه می‌کند.

تعبیه محاسبه‌شده به یک لیست تبدیل می‌شود تا بتوان آن را در قالب JSON ذخیره کرد. برای هر سند، یک مسیر فایل جدید در پوشه مشخص‌شده (output_dir) ساخته می‌شود که نام فایل آن برابر با شناسه سند (doc_id) است. اطلاعات سند شامل شناسه، محتوای متن، و بردار تعبیه در قالب JSON در این فایل ذخیره می‌شود.

```python
from sentence_transformers import SentenceTransformer

model_name = "HooshvareLab/bert-fa-zwnj-base"
model = SentenceTransformer(model_name)
```

مدل‌های BERT (Bidirectional Encoder Representations from Transformers) از معماری **ترنسفورمر** استفاده می‌کنند که به صورت دوطرفه (Bidirectional) متن را پردازش می‌کند. این مدل‌ها با تحلیل همزمان کلمات و زمینه آن‌ها در جمله، تعبیه‌هایی (Embedding) دقیق تولید می‌کنند. **SentenceTransformers** یک افزونه برای BERT است که قابلیت محاسبه تعبیه‌های سطح جمله را فراهم می‌کند. این تعبیه‌ها بردارهای عددی هستند که مفهوم معنایی جملات را در فضای برداری نمایش می‌دهند و برای وظایفی مانند محاسبه شباهت کسینوسی بین جملات، دسته‌بندی متن، و بازیابی اطلاعات کاربرد دارند.

```
}
"doc_id": "0",
"content": "باشگاه فوتبال آرسنال (به انگلیسی: Arsenal Football Club
برتر انگلستان، ۱۴ قهرمانی در جام حذفی فوتبال انگلستان ، ۱۶ قهرمانی در
صدرنشینی بدون وقفه در لیگ فوتبال انگلیس، بیشترین بازی بدون باخت پیاپی
لیگ برتر باشند که جام طلایی را بدست می‌آورند.",
"embedding": [
-0.01528053730726242,
-0.008420940488576889,
-0.5131236910820007,
0.16778865456581116,
-0.5135084390640259,
-0.37456828355789185,
-0.3614235520362854,
0.3500715494155884,
-0.10261653363704681,
0.5794336199760437,
-0.4020284414291382,
-0.018176885321736336,
0.22908315062522888,
0.056980282068252563,
-0.4038555324077606,
-0.3446156978607178,
-0.24892480671405792,
```

```
}
"doc_id": "32",
"content": "سیارک‌ها (به انگلیسی: Asteroid) اجسام کوچکی هستند که
"embedding": [
-0.02288219891488552,
-0.2868300974369049,
0.04903075471520424,
-0.3253119885921478,
-0.5120662450790405,
-0.3145887553691864,
-0.24702100455760956,
-0.425175279378891,
0.07507181912660599,
0.2652963697910309,
-0.6684086918830872,
-0.3624887466430664,
0.44382837414741516,
0.23430097103118896,
-0.8307507634162903,
-0.06440320611000061,
0.3895355761051178,
0.4078782796859741,
-0.6818297505378723,
0.03478926420211792,
-0.7255710959434509,
0.6284453868865967,
0.3594525158405304,
0.5018863081932068,
0.10109331458806992,
0.9907159805297852,
```

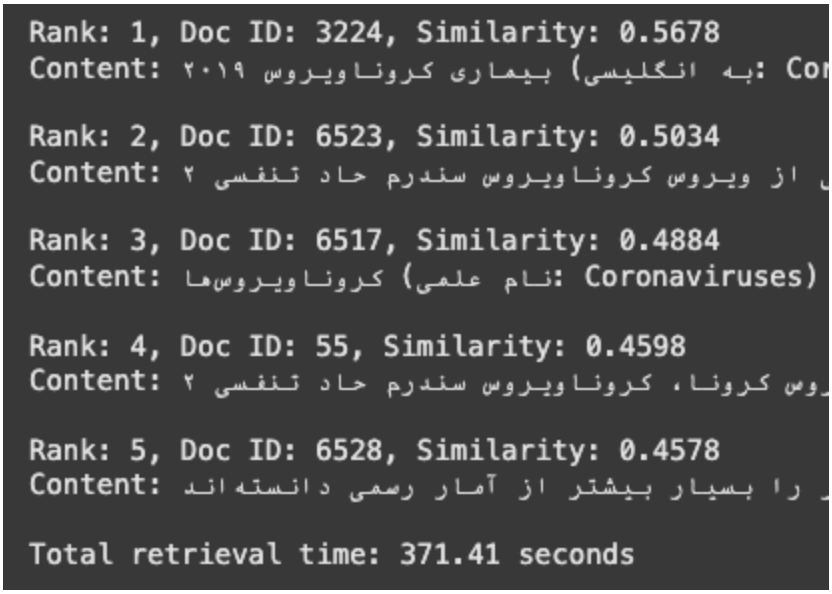

فاز ۲: جست‌وجوی همه اسناد

بررسی اندازه بردار تعبیه‌ها:
برای مدل زبانی انتخابی (مانند ParsBERT)، اندازه بردار تعبیه‌ها برای تمامی اسناد یکسان خواهد بود. این موضوع به این دلیل است که ParsBERT و سایر مدل‌های مبتنی بر BERT بردارهای تعبیه‌ای با ابعاد ثابت تولید می‌کنند (به طور پیش‌فرض 768 بُعدی). بنابراین، برای تمامی اسناد، طول بردار تعبیه‌ها ثابت است.

روش محاسبه فاصله:
اگر اندازه بردارها یکسان باشد: از شباهت کسینوسی (Cosine Similarity) یا فاصله اقلیدسی (L2 Distance) برای محاسبه شباهت یا فاصله میان بردارها استفاده می‌کنیم.
شباهت کسینوسی: معیاری برای سنجش شباهت زاویه‌ای بین دو بردار، مناسب برای سنجش شباهت معنایی.
فاصله اقلیدسی: مناسب برای سنجش فاصله فیزیکی بین دو نقطه در فضای برداری.
در این پروژه، برای کاهش حساسیت به مقیاس بردارها و تأکید بر شباهت معنایی، از شباهت کسینوسی استفاده می‌کنیم.
دلیل انتخاب شباهت کسینوسی:
مزایا:
مستقل از اندازه و مقیاس بردارها.
تمرکز بر شباهت معنایی به جای فاصله فیزیکی.
تناسب با مدل ParsBERT: بردارهای تولید شده توسط ParsBERT معمولاً به‌صورت نرمال نشده ارائه می‌شوند، اما با نرمال‌سازی طول بردارها (طول برابر ۱)، شباهت کسینوسی دقیق‌ترین نتایج را در زمینه شباهت معنایی ارائه می‌دهد.

الف) بر روی پرسش "ویروس کرونا" شناسه سند و محتوای ۵ سند که بیشترین ارتباط را با پرسش دارند به صورت مرتب شده بر اساس میزان ارتباطشان (اولین سند مشابه ترین سند، سپس دومین سند و...) را ذکر کنید:

```
d = 5  # Number of top similar documents to retrieve
phase_1_dir = "phase_1_result"

def load_document_embeddings(directory):
    documents = []
    for file_name in os.listdir(directory):
        file_path = os.path.join(directory, file_name)
        with open(file_path, 'r', encoding='utf-8') as f:
            doc = json.load(f)
            documents.append(doc)
    return documents

def compute_similarity(query_embedding, document_embedding):
    return 1 - cosine(query_embedding, document_embedding)

def find_top_d_similar(query_embedding, documents, d=5):
    similarities = []
    for doc in documents:
        sim = compute_similarity(query_embedding, doc['embedding'])
        similarities.append((doc['doc_id'], doc['content'], sim))

    similarities = sorted(similarities, key=lambda x: x[2], reverse=True)

    return similarities[:d]

start_time = time.time()

# Step 1: Convert query to embedding
query_embedding = model.encode(query_text).tolist()

# Step 2: Load all document embeddings from disk
documents = load_document_embeddings(phase_1_dir)

# Step 3: Compute similarities and retrieve top d similar documents
top_documents = find_top_d_similar(query_embedding, documents, d)

end_time = time.time()

for rank, (doc_id, content, score) in enumerate(top_documents, start=1):
    print(f"Rank: {rank}, Doc ID: {doc_id}, Similarity: {score:.4f}")
    print(f"Content: {content}\n")

execution_time = end_time - start_time
print(f"Total retrieval time: {execution_time:.2f} seconds")
```



ابتدا کوئری مورد نظر ("ویروس کرونا") به یک بردار تعبیه تبدیل می‌شود. سپس تمامی فایل‌های موجود در دایرکتوری مشخص‌شده (`phase_1_result`) که شامل تعبیه‌های اسناد است، بارگذاری می‌شوند. برای هر سند، شباهت کسینوسی میان تعبیه کوئری و تعبیه سند با استفاده از متد `cosine` محاسبه می‌شود. این شباهت‌ها به همراه شناسه و محتوای سند ذخیره می‌شوند.

در ادامه، اسناد بر اساس میزان شباهت به ترتیب نزولی مرتب شده و تعداد d سند مرتبط‌تر (در اینجا 5 سند) به عنوان خروجی بازگردانده می‌شوند. در پایان، شناسه، شباهت و محتوای این اسناد نمایش داده شده و زمان کل پردازش (شامل محاسبه تعبیه کوئری، بارگذاری اسناد و محاسبه شباهت) گزارش می‌شود.

شماره سند ۱:
3224

محتوا:
Content: بیماری کروناویروس ۲۰۱۹ (به انگلیسی: Coronavirus disease 2019) یا کووید-۱۹ (انگلیسی: COVID-19) که به آن بیماری تنفسی حاد ان‌کاو-۲۰۱۹ یا به‌شکل عمومی به آن کرونا نیز می‌گویند

امتیاز:
0.5678

شماره سند ۲:

6523

محتوا:
Content: دنیاگیری کووید-۱۹ یک دنیاگیری در جریان از کووید-۱۹، ناشی از ویروس کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2) است. این بیماری برای نخستین بار در دسامبر ۲۰۱۹ در شهر ووهان استان هوبئی، کشور چین شناسایی شد. در ۱۱ مارس، سازمان جهانی بهداشت شیوع بیماری را دنیاگیری اعلام کرد.

امتیاز:
0.5034

شماره سند ۳:
6517

محتوا:
Content: کروناویروس‌ها (نام علمی: Coronaviruses) خانواده بزرگی از ویروس‌ها و عضو خانوادهٔ ویروسی کروناویریده هستند که از ویروس سرماخوردگی معمولی تا عامل بیماری‌های شدیدتری همچون سارس، مرس و کووید ۱۹ را شامل می‌شود. کروناویروس‌ها در دههٔ ۱۹۶۰ کشف شدند و مطالعه بر روی آن‌ها به‌طور مداوم تا اواسط دهه ۱۹۸۰ ادامه داشت. این ویروس‌ها به‌طور طبیعی در پستانداران و پرندگان شیوع پیدا می‌کنند، با این حال تاکنون هفت کروناویروس منتقل شده به انسان، کشف شده‌است. آخرین نوع آن‌ها، کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2)، در دسامبر ۲۰۱۹ در شهر ووهان چین با همه‌گیری در انسان شیوع پیدا کرد. این کروناویروس پس از مدت کوتاهی تمام جهان را درگیر کرد

امتیاز:
0.4884

شماره سند ۴:
55

محتوا:
Content: از نوامبر ۲۰۰۲ تا ژوئیهٔ ۲۰۰۳ شیوع سارس در استان‌های جنوبی چین باعث بروز ۸٬۰۹۸ مورد بیماری احتمالی شد و در پایان مرگ ۷۷۴ تن در ۱۷ کشور گزارش شد. آخرین نوع ویروس کرونا، کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2)، در دسامبر ۲۰۱۹ در شهر ووهان چین با همه‌گیری در انسان توسط خوردن خفاش

شیوع پیدا کرد. کرونا ویروس پس از مدت کوتاهی موجب دنیاگیری کروناویروس در جهان شد. ضربات کرونا به امور حوزه دیپلماسی و تصور جهانیان از چینی‌ها شد. به دلیل عدم شفافیت‌ها، در نهایت آثار منفی کرونا بر موقعیت جهانی چین تا جایی پیش رفت که کشورهای گوناگونی چین را به مخفی‌کاری دربارهٔ انتشار کرونا و همدستی با سازمان بهداشت جهانی برای مخفی‌کاری خطاب کردند. به دلیل این، هم سازمان بهداشت جهانی کمک سالانه ۴۰۰ میلیون دلار واشینگتن را از دست داد.

امتیاز:
0.4598

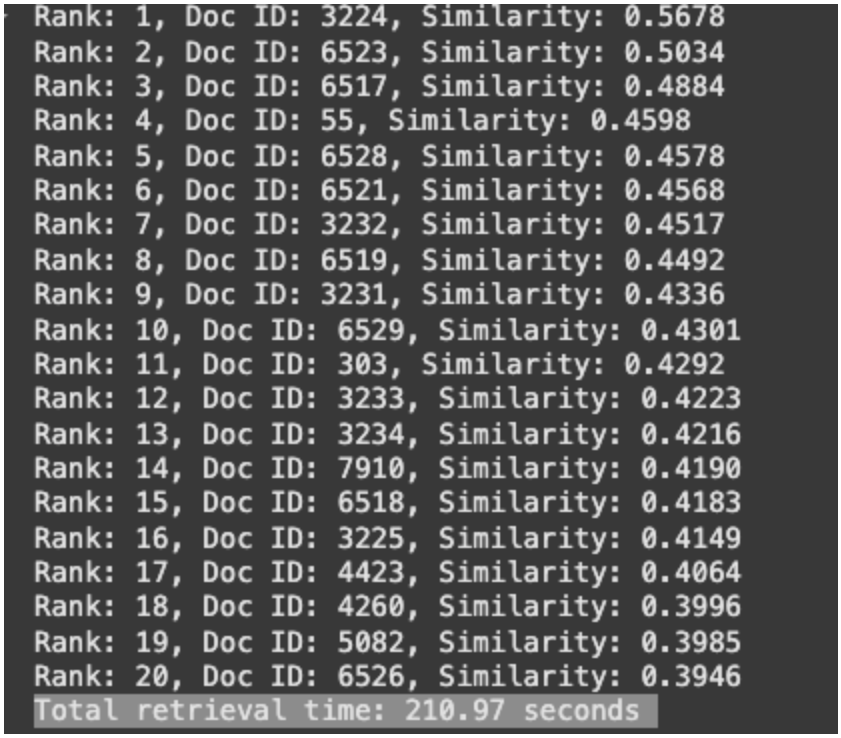
شماره سند ۵:
6528

محتوا:
Content: در پی دنیاگیری ۲۰–۲۰۱۹ کروناویروس در نقاط مختلف جهان، دنیاگیری کروناویروس در ایران رسماً در تاریخ ۲۹ بهمن ۱۳۹۸ تأیید شد. روز چهارشنبه ۱۴ اسفند حسن روحانی، رئیس‌جمهور ایران گفت، شیوع ویروس جدید کرونا تقریباً همه استان‌های کشور را دربرگرفته است. به گفته روابط عمومی وزارت بهداشت، درمان و آموزش پزشکی ایران تا ظهر یکشنبه ۱۰ شهریور ۱۳۹۹، شمار ۳۷۵۲۱۲ بیمار مبتلا به کووید-۱۹ در کشور شناسایی شده که از این شمار ۲۱۵۷۱ نفر جان باخته‌اند و همچنین ۳۲۳۲۳۳ نفر از مبتلایان به این ویروس تاکنون بهبود یافته‌اند.این آمار تا تاریخ ۱۰ شهریور تنها بر مبنای موارد مثبت تست کرونا طبق استاندارد سازمان جهانی بهداشت و از این تاریخ برمبنای عوارض بالینی سی‌تی اسکن ریه در کنار تست بوده‌است. تخمین‌هایی که توسط پژوهشگران و صاحب‌نظران زده شده، از جمله پژوهش‌هایی در دانشگاه تورنتو، میزان ابتلا و مرگ و میر را بسیار بیشتر از آمار رسمی دانسته‌اند.

امتیاز:
0.4578

ب) بخش الف را برای ۲۰ سند برای پرسش "ویروس کرونا" انجام دهید اما این بار محتوای اسناد را ذکر نکنید.

شماره اسناد و امتیاز هرکدام:



ج) زمان اجرای بخش‌های (الف) و (ب) را محاسبه کرده و در گزارش خود بیاورید.
منظور از زمان بازیابی، مدت زمان محاسبه بردار تعبیه برای کوئری، خواندن دانه دانه سندها از فایل‌ها، محاسبه شباهت هر سند با کوئری، و در نهایت مرتب سازی اسناد بر اساس شباهت و برگرداندن شبیه ترین سندها به کوئری است.

الف) Total retrieval time: 209.71 seconds

ب) Total retrieval time: 210.97 seconds

د) مورد (الف) را برای یک کوئری دلخواه انجام دهید.

query_text = "بهترین بازیکن فوتبال جهان"

Rank: 1, Doc ID: 2366, Similarity: 0.5294

Content: علی دایی (زادهٔ ۱۴ بهمن ۱۳۴۸) بازیکن فوتبال بازنشستهٔ تیم ملی ایران و باشگاه پرسپولیس است که علاوه بر بازی در لیگ برتر ایران، سابقه حضور در لیگ‌های ستارگان قطر، امارات و بوندسلیگا آلمان را هم در کارنامه دارد. او اکنون مربی فوتبال، تجارت‌پیشه، کارآفرین و بنیانگذار و مدیرعامل شرکت پوشاک ورزشی دایی است. دایی که بهترین گلزن تاریخ تیم ملی ایران به‌شمار می‌رود؛ در نظرسنجی‌های سایت ای‌اف‌سی و برنامه نود با کسب اکثریت آرا به ترتیب به عنوان بهترین مهاجم تاریخ جام ملت‌های آسیا و نیز بهترین مهاجم بعد از انقلاب در ایران انتخاب شد. همچنین نام وی از سوی ای‌اس‌پی‌ان در لیست ده بازیکن برتر تاریخ فوتبال آسیا قرار

گرفته‌است. او با زدن ۱۰۹ گل در ۱۴۹ بازی ملی، رکورد بیشترین گلِ زده در بازی‌های ملی فوتبال مردان جهان و با زدن ۱۴ گل، رکورد بیشترین گلزنی را در ادوار جام ملت‌های آسیا در اختیار دارد.

Rank: 2, Doc ID: 0, Similarity: 0.5156
Content: باشگاه فوتبال آرسنال (به انگلیسی: Arsenal Football Club) یک باشگاه فوتبال انگلیسی در شمال شهر لندن است که موفق به کسب ۱۳ عنوان قهرمانی در لیگ دسته اول و لیگ برتر انگلستان، ۱۴ قهرمانی در جام حذفی فوتبال انگلستان ، ۱۶ قهرمانی در جام خیریه انگلستان و دو قهرمانی در جام اتحادیه فوتبال انگلستان شده‌است. آن‌ها رکورددار طولانی‌ترین مدت صدرنشینی بدون وقفه در لیگ فوتبال انگلیس، بیشترین بازی بدون باختِ پیاپی (۴۹ بازی) و همچنین قهرمانی بدون شکست در یک فصل (۰۴–۲۰۰۳) می‌باشند و توانستند اولین و تنها تیمی در تاریخ لیگ برتر باشند که جام طلایی را بدست می‌آورند.

Rank: 3, Doc ID: 4081, Similarity: 0.5069
Content: منچستر یونایتد با داشتن بیشترین هوادار، باشگاه‌های هواداری و بیشترین میانگین تماشاگر برای هر بازی خانگی، پرطرفدارترین تیم جهان به حساب می‌آید. این باشگاه بالغ بر ۲۰۰ باشگاه هواداری رسمی را در ۲۴ کشور دنیا اداره می‌کند. باشگاه منچستر یونایتد به خاطر تورهای تابستانی خود و سفر به نقاط مختلف جهان در تعطیلات، محبوبیت ویژه‌ای در نزد مردم جهان دارد. نتایج یک نظرسنجی نیز در سال ۲۰۱۲ نشان می‌دهد که منچستر یونایتد با ۶۵۹ میلیون هوادار در سرتاسر دنیا، پرطرفدارترین تیم فوتبال دنیاست. نتایج یک تحقیق در سال ۲۰۱۴ نشان داد که هواداران یونایتد، پرسروصداترین هواداران در لیگ برتر انگلستان هستند.

Rank: 4, Doc ID: 3152, Similarity: 0.5064
Content: باشگاه فوتبال استون ویلا (به انگلیسی: Aston Villa F.C.) یک باشگاه حرفه‌ای فوتبال در لیگ برتر فوتبال انگلستان است که در شهر بیرمنگام در کشور انگلستان قرار دارد. این باشگاه در سال ۱۳۶۸ تأسیس شد و ورزشگاه خانگی آن‌ها از سال ۱۸۹۷ ورزشگاه ویلا پارک است. این باشگاه یکی از اعضای مؤسس لیگ فوتبال در سال ۱۸۸۸ و لیگ برتر فوتبال انگلستان در سال ۱۹۹۲ است. استون ویلا یکی از پنج باشگاه انگلیسی است که موفق به قهرمانی در ۸ جام باشگاه‌های اروپا شده‌است؛ آن‌ها در فصل ۸۲–۱۹۸۱ فاتح این رقابت‌ها شدند. آن‌ها هم‌چنین موفق شده‌اند هفت بار فاتح سطح اول لیگ فوتبال انگلستان، هفت بار فاتح جام حذفی فوتبال انگلستان، پنج بار فاتح جام اتحادیه باشگاه‌های انگلستان و یک بار قهرمان سوپرجام اروپا شوند.

Rank: 5, Doc ID: 3415, Similarity: 0.5039
Content: لیگ برتر با پخش شدن در بیش از ۲۱۲ سرزمین جهان و در ۶۴۳ میلیون خانه با بینندگانی که حداکثر به تعداد ۴/۷ میلیارد نفر می‌رسند، پربیننده‌ترین لیگ ورزشی در دنیاست. در فصل ۱۵–۲۰۱۴، هر بازی در لیگ برتر به صورت میانگین ۳۶,۰۰۰ تماشاگر داشت که پس از بوندسلیگا با ۴۳,۵۰۰ نفر، پرتماشاگرترین لیگ حرفه‌ای در

دنیا بود. در بیشتر بازی‌های لیگ برتر، استادیوم‌ها به صورت تقریباً کامل پر می‌شوند. لیگ برتر در جدول ضریب یوفا که بر پایهٔ عملکرد باشگاه‌ها در رقابت‌های اروپایی در پنج سال گذشته تنظیم می‌شود، در جایگاه سوم قرار دارد..

Total retrieval time: 207.11 seconds

توجه: در موارد (الف)، (ب) و (د) اگر تعداد اسناد بازیابی شده کمتر از تعداد خواسته شده بود، همه اسناد بازیابی شده را خروجی دهید.

فاز ۳: جست و جو در خوشه‌ها

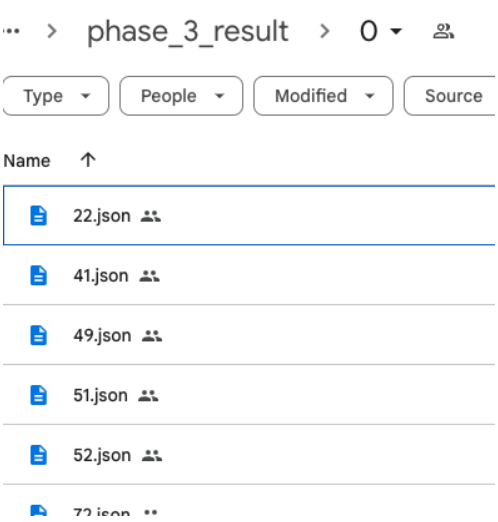
در این بخش چون خوشه بندی به خوبی انجام شده بود نتایج بسیار شبیه به فاز ۱ بود ولی از طرفی سرعت بازیابی چون فقط در آن خوشه مرتبط با کوئری جست و جو میشود بسیار کاهش میابد.

الف) دلیل بیاورید که با داشتن n سند و با فرض یکنواخت بودن توزیع بردار تعبیه آن‌ها، چه تعداد خوشه‌ای برای کاهش زمان پرسش مناسب است؟

با داشتن n سند و فرض یکنواخت بودن توزیع بردار تعبیه‌ها، تعداد مناسب خوشه‌ها برای کاهش زمان پرسش sqrt(n) = k است. دلیل این انتخاب آن است که در خوشه‌بندی، فرآیند جستجو شامل دو مرحله اصلی است: یافتن خوشه مرتبط با کوئری و جستجو در میان اسناد موجود در آن خوشه. پیچیدگی زمانی یافتن خوشه مرتبط O(k) است و جستجو در خوشه شامل n/k سند پیچیدگی زمانی O(n/k) دارد. مجموع این دو، به حداقل مقدار خود می‌رسد با مقدار sqrt(n) = k این مقدار تعادلی بهینه بین تعداد خوشه‌ها و اندازه هر خوشه ایجاد می‌کند و منجر به کاهش زمان کل پرسش می‌شود.

ب) کد و آدرس directory مربوط به خوشه‌بندی اسناد و ذخیره آن‌ها در directory به نام خوشه مربوطه را نشان دهید.

https://drive.google.com/drive/folders/1zK35hnm3AhrFiOvD-kcDLtx7Pn_eyuEc?usp=drive_link




```python
from sklearn.cluster import KMeans
import joblib
import math

num_clusters = math.ceil(math.sqrt(len(embeddings)))

kmeans = KMeans(n_clusters=num_clusters, random_state=42).fit(embeddings)

output_dir = "phase_3_result"
os.makedirs(output_dir, exist_ok=True)
kmeans_model_path = os.path.join(output_dir, "kmeans_model.pkl")
joblib.dump(kmeans, kmeans_model_path)

print(f"KMeans model saved to {kmeans_model_path}.")
```

```
KMeans model saved to phase_3_result/kmeans_model.pkl.
```

```python
for cluster_id in range(num_clusters):
    cluster_dir = os.path.join(output_dir, str(cluster_id))
    os.makedirs(cluster_dir, exist_ok=True)

for doc, cluster in zip(documents, kmeans.labels_):
    source_file = os.path.join(phase_1_dir, f"{doc['doc_id']}.json")
    target_file = os.path.join(output_dir, str(cluster), f"{doc['doc_id']}.json")
    os.system(f"cp {source_file} {target_file}")

print("Documents distributed into cluster directories.")
```

```
Documents distributed into cluster directories.
```

ابتدا تعداد خوشه‌ها (k) بر اساس ریشه دوم تعداد اسناد محاسبه می‌شود تا توزیع بهینه‌ای ایجاد شود. سپس الگوریتم KMeans با این تعداد خوشه اجرا می‌شود و مدل خوشه‌بندی به‌صورت یک فایل در مسیر مشخص‌شده ذخیره می‌شود. برای هر خوشه، یک پوشه جداگانه در دایرکتوری `phase_3_result` ایجاد می‌شود. در ادامه، اسناد با توجه به خوشه‌بندی KMeans به پوشه‌های متناظر منتقل می‌شوند.

ج)گزارش کنید که خوشه ای که کمترین تعداد اسناد را دارد، تعداد اسنادش چند است؟ همینطور خوشه ای که بیشترین تعداد اسناد را دارد؟

```
[24] from collections import Counter

     cluster_counts = Counter(kmeans.labels_)
     min_cluster = min(cluster_counts, key=cluster_counts.get)
     max_cluster = max(cluster_counts, key=cluster_counts.get)

     print(f"Cluster with the minimum documents: {min_cluster}, Count: {cluster_counts[min_cluster]}")
     print(f"Cluster with the maximum documents: {max_cluster}, Count: {cluster_counts[max_cluster]}")

     Cluster with the minimum documents: 75, Count: 19
     Cluster with the maximum documents: 69, Count: 165
```

```python
import time
from scipy.spatial.distance import cosine

start_time = time.time()

query_text = "ویروس کرونا"
query_embedding = model.encode(query_text).tolist()

kmeans = joblib.load(kmeans_model_path)

# Predict query cluster
query_cluster = kmeans.predict([query_embedding])[0]
print(f"Query belongs to cluster: {query_cluster}")

# Retrieve documents from the query's cluster
cluster_dir = os.path.join(output_dir, str(query_cluster))
cluster_files = os.listdir(cluster_dir)

# Compute similarities within the cluster
d = 5
similarities = []

for file_name in cluster_files:
    file_path = os.path.join(cluster_dir, file_name)
    with open(file_path, 'r', encoding='utf-8') as f:
        doc = json.load(f)
        sim = 1 - cosine(query_embedding, doc['embedding'])
        similarities.append((doc['doc_id'], doc['content'], sim))

# Sort by similarity score and retrieve top `d`
similarities = sorted(similarities, key=lambda x: x[2], reverse=True)[:d]

end_time = time.time()

# Display results
for rank, (doc_id, content, score) in enumerate(similarities, start=1):
    print(f"Rank: {rank}, Doc ID: {doc_id}, Similarity: {score:.4f}")
    print(f"Content: {content}\n")

# Report retrieval time
print(f"Retrieval time: {end_time - start_time:.2f} seconds")
```

ابتدا کوئری متنی "ویروس کرونا" به یک بردار تعبیه تبدیل می‌شود. سپس مدل KMeans که قبلاً ذخیره شده بود، بارگذاری شده و خوشه مربوط به کوئری با استفاده از متد `predict` تعیین می‌شود. پس از مشخص شدن خوشه، تمامی فایل‌های مربوط به اسناد موجود در آن خوشه بارگذاری می‌شوند. برای هر سند در این خوشه، شباهت کسینوسی میان تعبیه کوئری و تعبیه سند محاسبه شده و در لیستی همراه با شناسه و محتوای سند ذخیره می‌شود. لیست شباهت‌ها بر اساس امتیاز به ترتیب نزولی مرتب شده و d سند مرتبط‌تر (در اینجا ۵ سند) به‌عنوان خروجی نمایش داده می‌شود. در نهایت، زمان کل پردازش (از تبدیل کوئری به تعبیه تا بازیابی و مرتب‌سازی اسناد) گزارش می‌شود. این کد کارایی جستجو را با محدود کردن جستجو به یک خوشه خاص بهینه می‌کند.

د) موارد (الف)، (ب)، (ج)، (د) فاز ۲ را این بار با نتیجه خوشه‌بندی انجام دهید.

نتیجه بخش (الف) با خوشه بندی:

Query belongs to cluster: 57

Rank: 1, Doc ID: 6523, Similarity: 0.5034

Content: دنیاگیری کووید-۱۹ یک دنیاگیری در جریان از کووید-۱۹، ناشی از ویروس کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2) است. این بیماری برای نخستین بار در دسامبر ۲۰۱۹ در شهر ووهان استان هوبئی، کشور چین شناسایی شد. در ۱۱ مارس، سازمان جهانی بهداشت شیوع بیماری را دنیاگیری اعلام کرد.

Rank: 2, Doc ID: 6517, Similarity: 0.4884

Content: کروناویروس‌ها (نام علمی: Coronaviruses) خانواده بزرگی از ویروس‌ها و عضو خانوادهٔ ویروسی کروناویریده هستند که از ویروس سرماخوردگی معمولی تا عامل بیماری‌های شدیدتری همچون سارس، مرس و کووید ۱۹ را شامل می‌شود. کروناویروس‌ها در دههٔ ۱۹۶۰ کشف شدند و مطالعه بر روی آن‌ها به‌طور مداوم تا اواسط دهه ۱۹۸۰ ادامه داشت. این ویروس‌ها به‌طور طبیعی در پستانداران و پرندگان شیوع پیدا می‌کنند، با این حال تاکنون هفت کروناویروس منتقل شده به انسان، کشف شده‌است. آخرین نوع آن‌ها، کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2)، در دسامبر ۲۰۱۹ در شهر ووهان چین با همه‌گیری در انسان شیوع پیدا کرد. این کروناویروس پس از مدت کوتاهی تمام جهان را درگیر کرد

Rank: 3, Doc ID: 55, Similarity: 0.4598

Content: از نوامبر ۲۰۰۲ تا ژوئیهٔ ۲۰۰۳ شیوع سارس در استان‌های جنوبی چین باعث بروز ۸٬۰۹۸ مورد بیماری احتمالی شد و در پایان مرگ ۷۷۴ تن در ۱۷ کشور گزارش شد. آخرین نوع ویروس کرونا، کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2)، در دسامبر ۲۰۱۹ در شهر ووهان چین با همه‌گیری در انسان توسط خوردن خفاش شیوع پیدا کرد. کرونا ویروس پس از مدت کوتاهی موجب دنیاگیری کروناویروس در جهان شد. ضربات کرونا به امور حوزه دیپلماسی و تصور جهانیان از چینی‌ها شد. به دلیل عدم شفافیت‌ها، در نهایت آثار منفی کرونا بر موقعیت جهانی چین تا جایی پیش رفت که کشورهای گوناگونی چین را به مخفی‌کاری دربارهٔ انتشار کرونا و همدستی با سازمان بهداشت جهانی برای مخفی‌کاری خطاب کردند. به دلیل این، هم سازمان بهداشت جهانی کمک سالانه ۴۰۰ میلیون دلار واشینگتن را از دست داد.

Rank: 4, Doc ID: 6528, Similarity: 0.4578

Content: در پی دنیاگیری ۲۰–۲۰۱۹ کروناویروس در نقاط مختلف جهان، دنیاگیری کروناویروس در ایران رسماً در تاریخ ۲۹ بهمن ۱۳۹۸ تأیید شد. روز چهارشنبه ۱۴ اسفند حسن روحانی، رئیس‌جمهور ایران گفت، شیوع ویروس جدید کرونا تقریباً همه استان‌های کشور را دربرگرفته است. به گفته روابط عمومی وزارت بهداشت، درمان و آموزش پزشکی ایران تا ظهر یکشنبه ۱۰ شهریور ۱۳۹۹، شمار ۳۷۵۲۱۲ بیمار مبتلا به کووید-۱۹ در کشور شناسایی

شده که از این شمار ۲۱۵۷۱ نفر جان باخته‌اند و همچنین ۳۲۳۲۳۳ نفر از مبتلایان به این ویروس تاکنون بهبود یافته‌اند.این آمار تا تاریخ ۱۰ شهریور تنها بر مبنای موارد مثبت تست کرونا طبق استاندارد سازمان جهانی بهداشت و از این تاریخ برمبنای عوارض بالینی سی‌تی اسکن ریه در کنار تست بوده‌است. تخمین‌هایی که توسط پژوهشگران و صاحب‌نظران زده شده، از جمله پژوهش‌هایی در دانشگاه تورنتو، میزان ابتلا و مرگ و میر را بسیار بیشتر از آمار رسمی دانسته‌اند.

Rank: 5, Doc ID: 6521, Similarity: 0.4568
Content: بسته به نوع کروناویروس، روش‌های انتقال آن متفاوت است. در برخی از موارد روش‌های انتقال بیماری از انسان به انسان شبیه بیماری آنفلوانزا از طریق سرفه و عطسه است. با این حال احتمال انتقال بیماری در فضای باز بسیار محدود بوده و موارد انتقال انسان به انسان در مواردی رخ داده‌است که افراد به مدت طولانی در فضای بسته در کنار فرد بیمار بوده‌اند مانند افرادی که در بیمارستان‌ها با بیماران در ارتباط هستند. هنوز مشخص نیست که این بیماری نخستین بار به‌طور مستقیم از طریق جانوران به انسان منتقل شده‌است یا از طریق سطوح آلوده به ویروس.

Retrieval time: 3.07 seconds

نتیجه بخش (ب) با خوشه بندی (شماره اسناد و امتیاز هرکدام)

Query belongs to cluster: 57
Rank: 1, Doc ID: 6523, Similarity: 0.5034
Rank: 2, Doc ID: 6517, Similarity: 0.4884
Rank: 3, Doc ID: 55, Similarity: 0.4598
Rank: 4, Doc ID: 6528, Similarity: 0.4578
Rank: 5, Doc ID: 6521, Similarity: 0.4568
Rank: 6, Doc ID: 3232, Similarity: 0.4517
Rank: 7, Doc ID: 6519, Similarity: 0.4492
Rank: 8, Doc ID: 3231, Similarity: 0.4336
Rank: 9, Doc ID: 6529, Similarity: 0.4301
Rank: 10, Doc ID: 3233, Similarity: 0.4223
Rank: 11, Doc ID: 3234, Similarity: 0.4216
Rank: 12, Doc ID: 6526, Similarity: 0.3946
Rank: 13, Doc ID: 961, Similarity: 0.3775

Rank: 14, Doc ID: 6522, Similarity: 0.3658
Rank: 15, Doc ID: 3235, Similarity: 0.3572
Rank: 16, Doc ID: 2423, Similarity: 0.3436
Rank: 17, Doc ID: 397, Similarity: 0.3424
Rank: 18, Doc ID: 291, Similarity: 0.3322
Rank: 19, Doc ID: 3226, Similarity: 0.3216
Rank: 20, Doc ID: 1522, Similarity: 0.3200
Retrieval time: 0.85 seconds

نتیجه بخش (ج) با خوشه بندی:
زمان الف - Retrieval time: 3.07 seconds
زمان ب - Retrieval time: 0.85 seconds

نتیجه بخش (د) با خوشه بندی: (بهترین بازیکن فوتبال جهان)
Query belongs to cluster: 79
Rank: 1, Doc ID: 2366, Similarity: 0.5294
Content: علی دایی (زادهٔ ۱۴ بهمن ۱۳۴۸) بازیکن فوتبال بازنشستهٔ تیم ملی ایران و باشگاه پرسپولیس است که علاوه بر بازی در لیگ برتر ایران، سابقه حضور در لیگ‌های ستارگان قطر، امارات و بوندسلیگا آلمان را هم در کارنامه دارد. او اکنون مربی فوتبال، تجارت‌پیشه، کارآفرین و بنیانگذار و مدیرعامل شرکت پوشاک ورزشی دایی است. دایی که بهترین گلزن تاریخ تیم ملی ایران به‌شمار می‌رود؛ در نظرسنجی‌های سایت ای‌اف‌سی و برنامه نود با کسب اکثریت آرا به ترتیب به عنوان بهترین مهاجم تاریخ جام ملت‌های آسیا و نیز بهترین مهاجم بعد از انقلاب در ایران انتخاب شد. همچنین نام وی از سوی ای‌اس‌پی‌ان در لیست ده بازیکن برتر تاریخ فوتبال آسیا قرار گرفته‌است. او با زدن ۱۰۹ گل در ۱۴۹ بازی ملی، رکورد بیشترین گلِ زده در بازی‌های ملی فوتبال مردان جهان و با زدن ۱۴ گل، رکورد بیشترین گلزنی را در ادوار جام ملت‌های آسیا در اختیار دارد.

Rank: 2, Doc ID: 0, Similarity: 0.5156
Content: باشگاه فوتبال آرسنال (به انگلیسی: Arsenal Football Club) یک باشگاه فوتبال انگلیسی در شمال شهر لندن است که موفق به کسب ۱۳ عنوان قهرمانی در لیگ دسته اول و لیگ برتر انگلستان، ۱۴ قهرمانی در جام حذفی فوتبال انگلستان ، ۱۶ قهرمانی در جام خیریه انگلستان و دو قهرمانی در جام اتحادیه فوتبال انگلستان شده‌است. آن‌ها رکورددار طولانی‌ترین مدت صدرنشینی بدون وقفه در لیگ فوتبال انگلیس، بیشترین بازی بدون باختِ پیاپی (۴۹ بازی) و همچنین قهرمانی بدون شکست در یک فصل (۰۴–۲۰۰۳) می‌باشند و توانستند اولین و تنها تیمی در تاریخ لیگ برتر باشند که جام طلایی را بدست می‌آورند.

Rank: 3, Doc ID: 4081, Similarity: 0.5069

Content: منچستر یونایتد با داشتن بیشترین هوادار، باشگاه‌های هواداری و بیشترین میانگین تماشاگر برای هر بازی خانگی، پرطرفدارترین تیم جهان به حساب می‌آید. این باشگاه بالغ بر ۲۰۰ باشگاه هواداری رسمی را در ۲۴ کشور دنیا اداره می‌کند. باشگاه منچستر یونایتد به خاطر تورهای تابستانی خود و سفر به نقاط مختلف جهان در تعطیلات، محبوبیت ویژه‌ای در نزد مردم جهان دارد. نتایج یک نظرسنجی نیز در سال ۲۰۱۲ نشان می‌دهد که منچستر یونایتد با ۶۵۹ میلیون هوادار در سرتاسر دنیا، پرطرفدارترین تیم فوتبال دنیاست. نتایج یک تحقیق در سال ۲۰۱۴ نشان داد که هواداران یونایتد، پرسروصداترین هواداران در لیگ برتر انگلستان هستند.

Rank: 4, Doc ID: 3152, Similarity: 0.5064

Content: باشگاه فوتبال استون ویلا (به انگلیسی: Aston Villa F.C.) یک باشگاه حرفه‌ای فوتبال در لیگ برتر فوتبال انگلستان است که در شهر بیرمنگام در کشور انگلستان قرار دارد. این باشگاه در سال ۱۳۶۸ تأسیس شد و ورزشگاه خانگی آن‌ها از سال ۱۸۹۷ ورزشگاه ویلا پارک است. این باشگاه یکی از اعضای مؤسس لیگ فوتبال در سال ۱۸۸۸ و لیگ برتر فوتبال انگلستان در سال ۱۹۹۲ است. استون ویلا یکی از پنج باشگاه انگلیسی است که موفق به قهرمانی در ۸ جام باشگاه‌های اروپا شده‌است؛ آن‌ها در فصل ۸۲–۱۹۸۱ فاتح این رقابت‌ها شدند. آن‌ها هم‌چنین موفق شده‌اند هفت بار فاتح سطح اول لیگ فوتبال انگلستان، هفت بار فاتح جام حذفی فوتبال انگلستان، پنج بار فاتح جام اتحادیه باشگاه‌های انگلستان و یک بار قهرمان سوپرجام اروپا شوند.

Rank: 5, Doc ID: 3415, Similarity: 0.5039

Content: لیگ برتر با پخش شدن در بیش از ۲۱۲ سرزمین جهان و در ۶۴۳ میلیون خانه با بینندگانی که حداکثر به تعداد ۴٫۷ میلیارد نفر می‌رسند، پربیننده‌ترین لیگ ورزشی در دنیاست. در فصل ۱۵–۲۰۱۴، هر بازی در لیگ برتر به صورت میانگین ۳۶,۰۰۰ تماشاگر داشت که پس از بوندسلیگا با ۴۳,۵۰۰ نفر، پرتماشاگرترین لیگ حرفه‌ای در دنیا بود. در بیشتر بازی‌های لیگ برتر، استادیوم‌ها به صورت تقریباً کامل پر می‌شوند. لیگ برتر در جدول ضریب یوفا که بر پایهٔ عملکرد باشگاه‌ها در رقابت‌های اروپایی در پنج سال گذشته تنظیم می‌شود، در جایگاه سوم قرار دارد..

Retrieval time: 3.61 seconds

فاز ۴: جست‌وجوی بهینه در فضای برداری

FAISS در مقایسه با خوشه‌بندی (مثل KMeans) نیازی به مرحله خوشه‌بندی اولیه ندارد و با بهره‌گیری از ساختارهای داده پیشرفته، امکان جستجوی سریع و دقیق در فضای برداری را فراهم می‌کند. این روش برای پروژه‌هایی که نیاز به جستجوی با دقت بالا و زمان کم در دیتاست‌های بزرگ دارند، انتخاب بهتری است.

https://drive.google.com/drive/folders/1-u1s48dxZDJyx_IBwcaketfj3HQ2mRsl?usp=drive_link

```python
import faiss

faiss.normalize_L2(embedding_matrix)  # Normalize for cosine similarity if needed
index = faiss.IndexFlatL2(embedding_matrix.shape[1])  # L2 similarity
index.add(embedding_matrix)  # Add document embeddings to the index

# Save the index
phase_4_dir = "phase_4_result"
os.makedirs(phase_4_dir, exist_ok=True)
faiss.write_index(index, os.path.join(phase_4_dir, "index_flatl2.faiss"))
print("FAISS index saved.")
import time
```

ابتدا بردارهای تعبیه (embedding_matrix) با استفاده از تابع `faiss.normalize_L2` نرمال‌سازی می‌شوند تا طول هر بردار برابر با ۱ شود. این نرمال‌سازی برای استفاده از **شباهت کسینوسی** به جای فاصله اقلیدسی (L2) مورد نیاز است. سپس یک شیء `IndexFlatL2` از FAISS ایجاد می‌شود که برای محاسبه شباهت یا فاصله در فضای برداری طراحی شده است. تعبیه‌های نرمال‌شده به این ایندکس اضافه می‌شوند.

FAISS Index یک ساختار داده‌ای بهینه برای جستجوی سریع در فضای برداری است که توسط کتابخانه **FAISS** طراحی شده است. این ایندکس به‌ویژه برای بازیابی نزدیک‌ترین همسایه‌ها (Nearest Neighbor Search) در مجموعه‌های بزرگ بردارهای تعبیه استفاده می‌شود. به کمک این ایندکس می‌توان در زمان بسیار کوتاه بردارهایی را که بیشترین شباهت (مانند شباهت کسینوسی یا کمترین فاصله اقلیدسی) با یک کوئری دارند، پیدا کرد.

در ادامه، ایندکس ایجادشده در یک دایرکتوری مشخص (`phase_4_result`) ذخیره می‌شود تا در مراحل بعدی برای جستجوی سریع قابل استفاده باشد.

```python
# Query embedding (from Phase 2 query)
query_text = "آخرین نوع ویروس کرونا"
query_embedding = model.encode(query_text).tolist()

# Normalize query embedding
query_vector = np.array(query_embedding, dtype='float32').reshape(1, -1)
faiss.normalize_L2(query_vector)

# Load the FAISS index
index = faiss.read_index(os.path.join(phase_4_dir, "index_flatl2.faiss"))

# Retrieve top d results
d = 5
start_time = time.time()
distances, indices = index.search(query_vector, d)
retrieval_time = time.time() - start_time

# Print results
print(f"Query: {query_text}")
print(f"Retrieval time: {retrieval_time:.4f} seconds")

# Display top d documents
for rank, idx in enumerate(indices[0], start=1):
    doc = documents[idx]
    print(f"Rank: {rank}, Doc ID: {doc['doc_id']}")
    print(f"Content: {doc['content']}\n")
```

ابتدا کوئری متنی ("آخرین نوع ویروس کرونا") به یک بردار تعبیه با استفاده از مدل زبانی تبدیل می‌شود. سپس این بردار با استفاده از `faiss.normalize_L2` نرمال‌سازی می‌شود تا طول آن برابر با ۱ شود، که برای محاسبه شباهت کسینوسی ضروری است.

ایندکس FAISS که قبلاً ایجاد و ذخیره شده بود، بارگذاری می‌شود. سپس جستجو در این ایندکس با استفاده از بردار کوئری انجام شده و d سند برتر (در اینجا ۵ سند) که بیشترین شباهت را با کوئری دارند، بازیابی می‌شوند. در طول جستجو، فاصله بین بردار کوئری و بردارهای ذخیره‌شده در ایندکس محاسبه می‌شود و نتایج بر اساس این فاصله مرتب می‌شوند.

در نهایت، شناسه و محتوای d سند برتر به همراه زمان کل پردازش (از جستجو تا بازیابی) نمایش داده می‌شود.

الف) کتابخانه و مدل مورد استفاده برای جست‌وجوی فضای برداری مورد استفاده خود را ذکر کنید.

کتابخانه مورد استفاده: FAISS (Facebook AI Similarity Search)
مدل مورد استفاده: مدل زبانی **ParsBERT** از کتابخانه **SentenceTransformers** (با نام کامل "HooshvareLab/bert-fa-zwnj-base").

ب) موارد (الف)، (ب)، (ج)، (د) فاز ۲ را با نتیجه جست‌وجوی فضای برداری انجام دهید.

نتیجه بخش (الف) با جست‌وجوی فضای برداری (نیازی به گزارش امتیاز نیست):

Query: ویروس کرونا
Retrieval time: 0.0114 seconds
Rank: 1, Doc ID: 3224
Content: بیماری کروناویروس ۲۰۱۹ (به انگلیسی: Coronavirus disease 2019) یا کووید-۱۹ (انگلیسی: COVID-19) که به آن بیماری تنفسی حاد ان‌کاو-۲۰۱۹ یا به‌شکل عمومی به آن کرونا نیز می‌گویند

Rank: 2, Doc ID: 6523
Content: دنیاگیری کووید-۱۹ یک دنیاگیری در جریان از کووید-۱۹، ناشی از ویروس کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2) است. این بیماری برای نخستین بار در دسامبر ۲۰۱۹ در شهر ووهان استان هوبئی، کشور چین شناسایی شد. در ۱۱ مارس، سازمان جهانی بهداشت شیوع بیماری را دنیاگیری اعلام کرد.

Rank: 3, Doc ID: 6517
Content: کروناویروس‌ها (نام علمی: Coronaviruses) خانواده بزرگی از ویروس‌ها و عضو خانوادهٔ ویروسی کروناویریده هستند که از ویروس سرماخوردگی معمولی تا عامل بیماری‌های شدیدتری همچون سارس، مرس و کووید ۱۹ را شامل می‌شود. کروناویروس‌ها در دههٔ ۱۹۶۰ کشف شدند و مطالعه بر روی آن‌ها به‌طور مداوم تا اواسط دهه ۱۹۸۰ ادامه داشت. این ویروس‌ها به‌طور طبیعی در پستانداران و پرندگان شیوع پیدا می‌کنند، با این حال تاکنون هفت کروناویروس منتقل شده به انسان، کشف شده‌است. آخرین نوع آن‌ها، کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2)، در دسامبر ۲۰۱۹ در شهر ووهان چین با همه‌گیری در انسان شیوع پیدا کرد. این کروناویروس پس از مدت کوتاهی تمام جهان را درگیر کرد

Rank: 4, Doc ID: 55
Content: از نوامبر ۲۰۰۲ تا ژوئیهٔ ۲۰۰۳ شیوع سارس در استان‌های جنوبی چین باعث بروز ۸٬۰۹۸ مورد بیماری احتمالی شد و در پایان مرگ ۷۷۴ تن در ۱۷ کشور گزارش شد. آخرین نوع ویروس کرونا، کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2)، در دسامبر ۲۰۱۹ در شهر ووهان چین با همه‌گیری در انسان توسط خوردن خفاش شیوع پیدا کرد. کرونا ویروس پس از مدت کوتاهی موجب دنیاگیری کروناویروس در جهان شد. ضربات کرونا به امور حوزه دیپلماسی و تصور جهانیان از چینی‌ها شد. به دلیل عدم شفافیت‌ها، در نهایت آثار منفی کرونا بر موقعیت جهانی چین تا جایی پیش رفت که کشورهای گوناگونی چین را به مخفی‌کاری دربارهٔ انتشار کرونا و همدستی با سازمان بهداشت جهانی برای مخفی‌کاری خطاب کردند. به دلیل این، هم سازمان بهداشت جهانی کمک سالانه ۴۰۰ میلیون دلار واشینگتن را از دست داد.

Rank: 5, Doc ID: 6528
Content: در پی دنیاگیری ۲۰–۲۰۱۹ کروناویروس در نقاط مختلف جهان، دنیاگیری کروناویروس در ایران رسماً در تاریخ ۲۹ بهمن ۱۳۹۸ تأیید شد. روز چهارشنبه ۱۴ اسفند حسن روحانی، رئیس‌جمهور ایران گفت، شیوع ویروس جدید کرونا تقریباً همه استان‌های کشور را دربرگرفته است. به گفته روابط عمومی وزارت بهداشت، درمان و آموزش پزشکی ایران تا ظهر یکشنبه ۱۰ شهریور ۱۳۹۹، شمار ۳۷۵۲۱۲ بیمار مبتلا به کووید-۱۹ در کشور شناسایی شده که از این شمار ۲۱۵۷۱ نفر جان باخته‌اند و همچنین ۳۲۳۲۳۳ نفر از مبتلایان به این ویروس تاکنون بهبود یافته‌اند.این آمار تا تاریخ ۱۰ شهریور تنها بر مبنای موارد مثبت تست کرونا طبق استاندارد سازمان جهانی بهداشت و از این تاریخ برمبنای عوارض بالینی سی‌تی اسکن ریه در کنار تست بوده‌است. تخمین‌هایی که توسط پژوهشگران و صاحب‌نظران زده شده، از جمله پژوهش‌هایی در دانشگاه تورنتو، میزان ابتلا و مرگ و میر را بسیار بیشتر از آمار رسمی دانسته‌اند.

نتیجه بخش (ب) با جست‌وجوی فضای برداری (نیازی به گزارش امتیاز نیست):

Query: ویروس کرونا
Retrieval time: 0.0147 seconds
Rank: 1, Doc ID: 3224
Rank: 2, Doc ID: 6523
Rank: 3, Doc ID: 6517
Rank: 4, Doc ID: 55
Rank: 5, Doc ID: 6528
Rank: 6, Doc ID: 6521
Rank: 7, Doc ID: 3232
Rank: 8, Doc ID: 6519
Rank: 9, Doc ID: 3231
Rank: 10, Doc ID: 6529
Rank: 11, Doc ID: 303
Rank: 12, Doc ID: 3233
Rank: 13, Doc ID: 3234
Rank: 14, Doc ID: 7910
Rank: 15, Doc ID: 6518
Rank: 16, Doc ID: 3225
Rank: 17, Doc ID: 4423
Rank: 18, Doc ID: 4260

Rank: 19, Doc ID: 5082
Rank: 20, Doc ID: 6526

نتیجه بخش (ج) با جست‌وجوی فضای برداری:
زمان ب - Retrieval time: 0.0147 seconds
زمان الف - Retrieval time: 0.0114 seconds

نتیجه بخش (د) با جست‌وجوی فضای برداری:
Query: بهترین بازیکن فوتبال جهان
Retrieval time: 0.0112 seconds
Rank: 1, Doc ID: 2366
Content: علی دایی (زادهٔ ۱۴ بهمن ۱۳۴۸) بازیکن فوتبال بازنشستهٔ تیم ملی ایران و باشگاه پرسپولیس است که علاوه بر بازی در لیگ برتر ایران، سابقه حضور در لیگ‌های ستارگان قطر، امارات و بوندسلیگا آلمان را هم در کارنامه دارد. او اکنون مربی فوتبال، تجارت‌پیشه، کارآفرین و بنیانگذار و مدیرعامل شرکت پوشاک ورزشی دایی است. دایی که بهترین گلزن تاریخ تیم ملی ایران به‌شمار می‌رود؛ در نظرسنجی‌های سایت ای‌اف‌سی و برنامه نود با کسب اکثریت آرا به ترتیب به عنوان بهترین مهاجم تاریخ جام ملت‌های آسیا و نیز بهترین مهاجم بعد از انقلاب در ایران انتخاب شد. همچنین نام وی از سوی ای‌اس‌پی‌ان در لیست ده بازیکن برتر تاریخ فوتبال آسیا قرار گرفته‌است. او با زدن ۱۰۹ گل در ۱۴۹ بازی ملی، رکورد بیشترین گلِ زده در بازی‌های ملی فوتبال مردان جهان و با زدن ۱۴ گل، رکورد بیشترین گلزنی را در ادوار جام ملت‌های آسیا در اختیار دارد.

Rank: 2, Doc ID: 0
Content: باشگاه فوتبال آرسنال (به انگلیسی: Arsenal Football Club) یک باشگاه فوتبال انگلیسی در شمال شهر لندن است که موفق به کسب ۱۳ عنوان قهرمانی در لیگ دسته اول و لیگ برتر انگلستان، ۱۴ قهرمانی در جام حذفی فوتبال انگلستان ، ۱۶ قهرمانی در جام خیریه انگلستان و دو قهرمانی در جام اتحادیه فوتبال انگلستان شده‌است. آن‌ها رکورددار طولانی‌ترین مدت صدرنشینی بدون وقفه در لیگ فوتبال انگلیس، بیشترین بازی بدون باختِ پیاپی (۴۹ بازی) و همچنین قهرمانی بدون شکست در یک فصل (۰۴–۲۰۰۳) می‌باشند و توانستند اولین و تنها تیمی در تاریخ لیگ برتر باشند که جام طلایی را بدست می‌آورند.

Rank: 3, Doc ID: 4081
Content: منچستر یونایتد با داشتن بیشترین هوادار، باشگاه‌های هواداری و بیشترین میانگین تماشاگر برای هر بازی خانگی، پرطرفدارترین تیم جهان به حساب می‌آید. این باشگاه بالغ بر ۲۰۰ باشگاه هواداری رسمی را در ۲۴ کشور دنیا اداره می‌کند. باشگاه منچستر یونایتد به خاطر تورهای تابستانی خود و سفر به نقاط مختلف جهان در

تعطیلات، محبوبیت ویژه‌ای در نزد مردم جهان دارد. نتایج یک نظرسنجی نیز در سال ۲۰۱۲ نشان می‌دهد که منچستر یونایتد با ۶۵۹ میلیون هوادار در سرتاسر دنیا، پرطرفدارترین تیم فوتبال دنیاست. نتایج یک تحقیق در سال ۲۰۱۴ نشان داد که هواداران یونایتد، پرسروصداترین هواداران در لیگ برتر انگلستان هستند.

Rank: 4, Doc ID: 3152

Content: باشگاه فوتبال استون ویلا (به انگلیسی: Aston Villa F.C.) یک باشگاه حرفه‌ای فوتبال در لیگ برتر فوتبال انگلستان است که در شهر بیرمنگام در کشور انگلستان قرار دارد. این باشگاه در سال ۱۳۶۸ تأسیس شد و ورزشگاه خانگی آن‌ها از سال ۱۸۹۷ ورزشگاه ویلا پارک است. این باشگاه یکی از اعضای مؤسس لیگ فوتبال در سال ۱۸۸۸ و لیگ برتر فوتبال انگلستان در سال ۱۹۹۲ است. استون ویلا یکی از پنج باشگاه انگلیسی است که موفق به قهرمانی در ۸ جام باشگاه‌های اروپا شده‌است؛ آن‌ها در فصل ۸۲–۱۹۸۱ فاتح این رقابت‌ها شدند. آن‌ها هم‌چنین موفق شده‌اند هفت بار فاتح سطح اول لیگ فوتبال انگلستان، هفت بار فاتح جام حذفی فوتبال انگلستان، پنج بار فاتح جام اتحادیه باشگاه‌های انگلستان و یک بار قهرمان سوپرجام اروپا شوند.

Rank: 5, Doc ID: 3415

Content: لیگ برتر با پخش شدن در بیش از ۲۱۲ سرزمین جهان و در ۶۴۳ میلیون خانه با بینندگانی که حداکثر به تعداد ۴/۷ میلیارد نفر می‌رسند، پربیننده‌ترین لیگ ورزشی در دنیاست. در فصل ۱۵–۲۰۱۴، هر بازی در لیگ برتر به صورت میانگین ۳۶,۰۰۰ تماشاگر داشت که پس از بوندسلیگا با ۴۳,۵۰۰ نفر، پرتماشاگرترین لیگ حرفه‌ای در دنیا بود. در بیشتر بازی‌های لیگ برتر، استادیوم‌ها به صورت تقریباً کامل پر می‌شوند. لیگ برتر در جدول ضریب یوفا که بر پایهٔ عملکرد باشگاه‌ها در رقابت‌های اروپایی در پنج سال گذشته تنظیم می‌شود، در جایگاه سوم قرار دارد..