گزارش فاز سوم پروژهی بازیایی اطلاعات

۱. مقایسهی نتایج بازیابی هنگام استفاده از خوشهبندی

خوشهبندی را با هایپرپارامترهای زیر انجام دادهایم.

```
num_clusters = 15  # K (num seeds)
max_iterations = 50
min_cluster_changed_ratio = 0.01
epochs = 5  # Number of tries for finding the optimal clustering (based on rss)

clusters = None
b = 3  # We search `b` nearest clusters for query
```

به كمك خوشهبندى، براى بازيابى اسناد مرتبط به يك پرسمان، نيازى نيست شباهت پرسمان را با تمامى اسناد محاسبه كنيم، و مىتوانيم تنها اسنادى كه (مراكز) خوشههايشان نزديك به پرسمان است را در نظر بگيريم. مطابق انتظار، با انجام اين كار سرعت بازبابى اطلاعات بالا رفت.

این افزایش سرعت به بهای کاهش اندکی از پوشش است، چرا که بخشی از اسناد مرتبط به پرسمان، ممکن است در خوشههای مورد جستجو قرار نگیرند. در این جا، چون ما نتایج جستجو را (در ادامهی فازهای قبلی) به ۵ نتیجه محدود کردهایم، تفاوت معناداری در نتایج احساس نشده، و هر دو روش اسناد مرتبط و قابل قبولی را برمی گردانند.

پرسمان ۱. دانشگاه امیرکبیر

- با خوشەبندى

- بدون خوشەبندى

```
venv) — ~/PycharmProjects/InfoRet-Phase1
  └─(01:53:57 on master •)─> python main.py
                                                                                                                                                                                                     —(Tue, Jan25)—
 INFO:io_handler:Positional Indexing loaded from ./dataset/preprocessed/IR3_50k_news_pos_idx.pkl
                                                                                               tokens
                     ...ایسنا, ،, پس, استعفا, عل, نکو, رئیس, شاخه, آک_
      ...ایسنا, ،, امیر, محسن, ب, اینکه, لیگ, طور, هفت]
      ...آرر فرهادر گفر گور ایسنار ،ر درباره ر آخرین, و آ
...ایسنار ،ر فدراسیون, بین, العلل, شنار قانون, ا] ... 4
...[بسنار ،ر فدراسیون, جهان, صعود, ورزش] ... 5
 [5 rows x 6 columns]
BaseQueryEngine Initializing finished
 Num CPU cores: 4
 WordEmbeddingQueryEngine Initializing finished
Enter your query: دانشگاه امیرکبیر
Docs found (ranked):
Doc#42633 Title: به گزارش حوزه دانشگاهی خبرگزاری فارس، حجت الاسلام ... Similarity: 0.8041917085647583

URL: https://www.farsnews.ir/news/13991121000805/ ... به گزارش حوزه دانشگاه ما با صنایح دفاعی و ونظامی و ارتباط بیشتری و بگیرند ... Similarity: 0.7094548940658569

URL: https://www.farsnews.ir/news/13990728000040/ ... به گزارش خبرنگار تشکیل مای دانشگاهی خبرگزاری فارس ... Similarity: 0.6000096201896667

URL: https://www.farsnews.ir/news/13990924000535/ ... به گزارش خبرنگار تشکیل مای دانشگاهی خبرگزاری فارس ... Similarity: 0.6000096201896667

URL: https://www.farsnews.ir/news/13990924000535/ ... به گزارش خبرنگار تشکیل مای دانشگاهی خبرگزاری فارس ... Similarity: 0.8083666781908
Enter your query:
```

پرسمان ۲. واکسن آسترازنکا

- با خوشەبندى

پرسمان ۳. ژیمناستیک

- با خوشەبندى

بدون خوشەبندى

```
Enter your query: وبعناستيك (بعناستيك found (ranked):

Doc#36118 Title: به گزارش خبرگزاری فارس، طبق اعلام روابط عمومی فدرا ( Similarity: 0.8471663594245911 

URL: https://www.farsnews.ir/news/13991202000288/ ربیا به از به المعدود المعدود
```

۲. مقایسهی نتایج بازیایی هنگام استفاده از دستهبندی

- **دستهی ورزشی: «سپاهان»** دسته نندی

```
Enter your query: cat:sports سباهان المدور ورزشی خبرنگار و مس رفسنجان | Similarity: 0.6632120609283447

Doc#33721 Title: بالمدور المدور المدو
```

بدون دستهبندی

با جستجوی کلمهی سپاهان بدون دستهبندی، بیشتر نتایج سیاسی مرتبط به سپاه، یا اقتصادی مرتبط به فولاد مبارکهی سپاهان بازیابی میشوند. اما با مشخص کردن دستهبندی ورزشی، نتایج محدود به باشگاه ورزشی سپاهان میشود.

دستهی فرهنگی: «تهاجم»
 با دستهبندی

```
Enter your query: cat:culture المائة المائة
```

دستەىندى

```
Enter your query: تهاجه المحدد المدالة المدال
```

با جستجوی واژهی تهاجم، بیشتر اخبار سیاسی، مرتبط با جنگها و تهاجمها بازیابی میشوند. اما وقتی دستهبندی را به «فرهنگی» محدود میکنیم، اسناد به اخبار تهاجم فرهنگی نزدیکتر میشوند. برای مثال در حالت دستهبندی شده، نتیجهی اول، خبر در مورد کتابیست که دربارهی موضوع تهاجم فرهنگی نوشته شده است.


```
Enter your query: cat:politics المناس (ranked):

Doc#41968 Title: مناس (بعدل المناس خبرگزاری فارس، جمعی از ه Similarity: 0.771375834941864

URL: https://www.farsnews.ir/news/13991205000436/ المناس مجان المناس معال المناس المناس
```

```
Enter your query: استقلال المراققة المراقة المراققة المر
```

کلمهی استقلال هم به طور پیشفرض منجر به اخبار مرتبط با باشگاه ورزشی استقلال میشود. حال برای یافتن اسناد مرتبط با استقلال سیاسی، اسناد همچنان مرتبط به باشگاه استقلال هستند، اما بیشتر اخبار جنبههای سیاسی دارند، اعم از تجمع مقابل مجلس، نامهی نمایندههای مجلس شورای اسلامی و پیگیریهای وزارتخانه.

دستهی سلامت

با دستهبندی

```
Enter your query: cat:health ورزش Docs found (ranked):

Doc#6973 Title: معايش مغنه م Similarity: 0.7310285568237305

URL: https://www.isna.ir/news/98091107750/٩٠ ... كنتر رشيد حيدرى مقدم دوشنبه ١١ آذر در معايش مغنه م Similarity: 0.7310285568237305

URL: https://www.isna.ir/news/98091107750/٩٠ ... كنتر رشيد حيدرى مقدم دوشنبه ١١ آذر در معايش مغنه م URL: https://www.isna.ir/news/98091107750/٩٠ ... كنتر رشيد حيدرى مقدم دوشنبه ١١ آذر در معايش مغنه م Similarity: 0.475998193025589

URL: https://www.isna.ir/news/98082516132... ... كنترل -ديابت-چيست/Similarity: 0.475998193025589

URL: https://www.isna.ir/news/98082516132... .: محسن صحر انورد بيا اشاره به دلايل بروز ديابت انورد ... Similarity: 0.475998193025589

URL: https://www.isna.ir/news/98082516132... كنترل -ديابت-چيست/Similarity: 0.45250198245048523

Doc#6907 Title: كليد كنترل -ديابت-چيست/Similarity: 0.45250198245048523

URL: https://www.isna.ir/news/98080402002/مدينا المناه به جيست/Similarity: 0.45250198245048523
```

دستەىندى

```
Enter your query: ورزش ایران» با قدمت چهار س Similarity: 0.8640766739845276

URL: https://www.isna.ir/news/99043022528/ مدرای ایسنا، برنامه «ورزش ایران» با قدمت چهار س Similarity: 0.8640766739845276

URL: https://www.isna.ir/news/99043022528/ مدراش ایسنا، برنامه «ورزش ایران» با قدمت چهار س Similarity: 0.8640766739845276

URL: https://www.isna.ir/news/99043022528/ مدراش ایران» با قدمت چهار س من Similarity: 0.8451473712921143

URL: https://www.farsnews.ir/news/139912190003564/ معالی ورزش مسلمانی و ایران س من Similarity: 0.8365150690078735

URL: https://www.farsnews.ir/news/13991109000350/ مدرا ورزش ایران سامن و ایران و ایران سامن و ایران و
```

واژهی ورزش بدون دستهبندی منجر به اخبار ورزشی می شود. اما انتخاب دستهبندی سلامت، اخبار بازیابی شده، همه در رابطه با اثرات ورزش بر روی سلامتی هستند.

دستهی اقتصادی: «بحران»

با دستهبندی

بدون دستهبندی

```
Enter your query: بعران المدورة والمدورة والمدو
```

جستجوی واژهی بحرانی، بدون دستهبندی نتایجی از دستههای سیاسی به همراه خواهد داشت. اما با محدود کردن دسته به دستهی اقتصادی، اخبار بیشتر به اخبار اقتصادی محدود میشوند.