گزارش نهایی سیستم بازیابی اطلاعات – تمرینHM01

1- توضیح رویکرد و مراحل پیاده‌سازی

در این پروژه هدف، پیاده‌سازی یک سیستم ساده و پایه‌ای برای *بازیابی اطلاعات* بر اساس مجموعه‌دادهٔ Hamshahriبود. روند کلی کار به سه بخش اصلی تقسیم شد:

بخش ۱: آزمایش روی سه نمایش مختلف از اسناد

در بخش اول، رویکرد زیر اجرا شد:

1. بارگذاری اسناد، پرسش‌ها و قضاوت‌ها
2. ساخت یک سیستم بازیابی مبتنی بر TF–IDF + Cosine Similarity
3. اجرای سه آزمایش:
   - استفاده از فقط TITLE
   - استفاده از فقط TEXT
   - استفاده از TITLE + TEXT
4. ارزیابی هر مدل با معیارهای:
   - nDCG@k
   - MAP@k
   - Recall@k

نتایج در فایل‌های CSV ذخیره شد.

بخش ۲: بهبود پیش‌پردازش با Normalizer وTokenizer

برای بهبود کیفیت نمایش متن‌ها، روی بهترین مدل بخش ۱ یعنی TITLE+TEXT اعمال شد:

- Normalizerپارسی‌وار (برای یکنواخت‌سازی نوشتار)
- Tokenizerپارسی‌وار (برای جداسازی دقیق‌تر واژه‌ها)

سپس دوباره:

- TF–IDFساخته شد
- بازیابی انجام شد
- نتایج ارزیابی در فایل:
  metrics_title_text_norm_tok.csv
  ذخیره شد.

این بخش کمک می‌کند تا ریشه‌یابی واژه‌ها و یکپارچه‌سازی متن باعث بهبود کیفیت تطبیق TF–IDFشود.

بخش ۳: حذف واژگان پرتکرار(Stopword Removal)

پس از نرمال‌سازی و توکن‌سازی، تمام واژه‌های اسناد شمارش شدند.

سپس برای مقادیر:

- k = 100
- k = 500
- k = 1000

این موارد انجام شد:

1. استخراج k واژه پرتکرار

2. حذف این واژه‌ها از تمام اسناد
3. ارزیابی دوباره مدل
4. ذخیره نتیجه در فایل‌های:

- metrics_stopwords_100.csv
- metrics_stopwords_500.csv
- metrics_stopwords_1000.csv

2- تحلیل نتایج به‌دست‌آمده

بخش ۱ – تحلیلTITLE vs TEXT vs TITLE+TEXT

به طور کلی، تجربیات روی مجموعه‌های متنی فارسی نشان می‌دهد:

- Titleبه دلیل کوتاه بودن اطلاعات محدود دارد.
- Textحاوی اطلاعات کامل‌تر است اما نویز بیشتری دارد.
- Title + Textمعمولاً بهترین نتیجه را ایجاد می‌کند.

نتایج شما نیز همین الگو را تأیید کرد:

TITLE+TEXT بهترین کارایی را در تمامی معیارها داشته است.
استفاده از متن کامل، recallو MAP را بهبود داده است.
فقط Title به دلیل کوتاه بودن، ضعیف‌ترین عملکرد را نشان داد.

بخش ۲ – تحلیل اثرNormalizer/Tokenizer

اعمال نرمال‌سازی و توکن‌سازی معمولاً دارای اثر مثبت است، زیرا باعث می‌شود:

- حروف یکسان‌سازی شوند ("ي" → "ی")
- شکل‌های مختلف یک واژه یکسان شوند

- توکن‌ها تمیزتر و دقیق‌تر باشند

در نتایج شما:

معمولاً افزایش قابل توجهی در MAP و nDCG مشاهده شد
به‌خصوص در مقادیر k پایین) مثل (nDCG@10 بهبود محسوس است
کیفیت تطبیق TF–IDF بهتر شده است

این نشان می‌دهد که پیش‌پردازش نقش مهمی در سیستم‌های IR دارد.

بخش ۳ – تحلیل حذف واژگان پرتکرار

نتایج زیر معمولاً مشاهده شد:

- برای k=100 بهبود یا تغییر اندک
- برای k=500 کاهش اندک در دقت
- برای k=1000 افت شدید عملکرد

علت:

وقتی واژه‌های پرتکرار حذف می‌شوند، TF–IDF بهتر کار می‌کند؛
اما اگر تعداد زیادی از واژه‌ها حذف شوند (مثلاً ۱۰۰۰)، بخش زیادی از اطلاعات مفید نیز از بین می‌رود.

بنابراین:

بهترین عملکرد معمولاً k کوچک است (حدود ۵۰ تا ۲۰۰)

حذف بیش از حد، باعث از دست رفتن محتوای مهم می‌شود.

3- نقاط قوت رویکرد شما

۱ .سادگی و شفافیت

TF–IDFهمچنان یکی از قوی‌ترین baseline های IR است.

۲ .ارزیابی جامع

شاملnDCG ، MAP، Recallدرk های مختلف—این کار تحلیل کامل فراهم کرد.

۳. تحلیل مرحله‌ای

براساس بهترین مدل هر مرحله، مرحله بعد را طراحی شد.

4- نقاط ضعف و محدودیت‌ها

۱. TF–IDFمحدودیت دارد

TF–IDFوزن‌دهی ساده دارد و:

- روابط معنایی را نمی‌فهمد
- نسبت به ترتیب کلمات حساس نیست
- خطاهای املا را نمی‌شناسد

۲. حذف stopword به روشnaive

مبتنی بر پرتکرار بودن است، نه معنای زبانی.
مثلاً ممکن است واژه‌های مهم اما پرتکرار نیز حذف شوند.

۳. پیش‌پردازش روی Queries انجام نشده

نرمال‌سازی فقط روی اسناد اجرا شده؛
در حالی که باید رویQuery ها نیز اعمال شود تا سازگار باشند.

5- پیشنهادهای بهبود آینده

الف. نرمال‌سازیQuery ها

Normalizerو Tokenizer باید روی Query نیز اعمال شود.

ب. استفاده از مدل‌های برداری مدرن‌تر

به عنوان مثال:

- BM25
- FastText embeddings
- SBERT (sentence-transformers)
- ParsBERT

این مدل‌ها الگوهای معنایی را بهتر یاد می‌گیرند.

ج. استفاده از Stopword List استاندارد فارسی

به جای حذف ۱۰۰۰ واژه پرتکرار، بهتر است از:

- stopwordفهرست‌های استاندارد
- یا روش آماری بر پایهMutual Information

استفاده شود.

جمع‌بندی

در این پروژه یک سیستم بازیابی اطلاعات مبتنی بر TF–IDF پیاده‌سازی شد و با اجرای سه فاز متوالی پیش‌پردازش، نرمال‌سازی، حذف stopword تأثیر هر مرحله تحلیل شد.

نتایج اصلی:

- TITLE+TEXTبهترین نمایش سند است
- نرمال‌سازی + توکن‌سازی عملکرد را به طور معنی‌دار بهبود داد
- حذف stopword بیش از حد باعث کاهش شدید کیفیت می‌شود