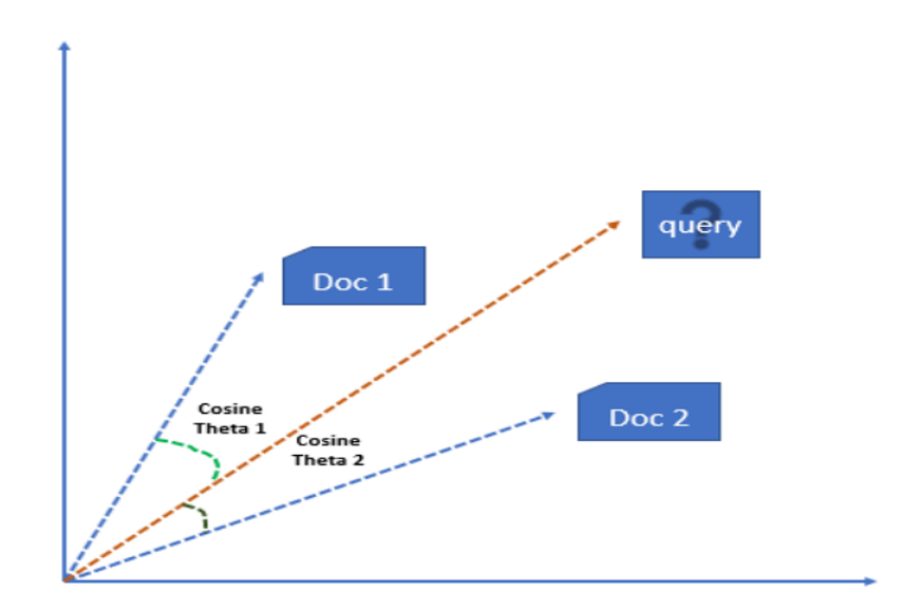
TF-IDF cosine Ranging

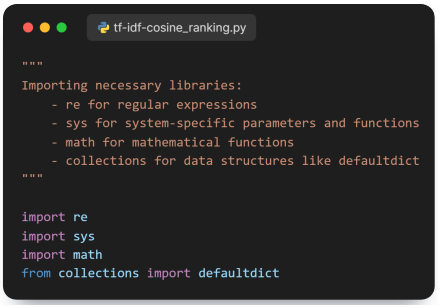


سیدمحمدحسین هاشمی ۴۰۲۲۳۶۳۱۴۳

فروردین ۱۴۰۳

فهرست مطالب

۱ کتابخانه‌های مورد نیاز



کتابخانه‌های مورد نیاز که در پروژه لود شده‌اند.

۲ کلاس Inverted Index

تمام کدهای مربوطه در این کلاس نوشته می‌شود

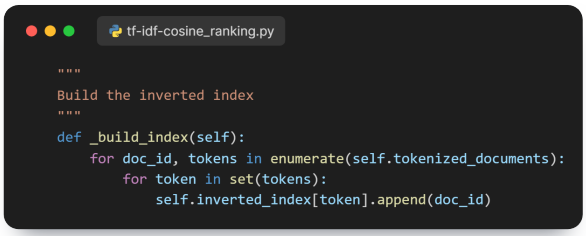
۳ متد –init–

در اینجا متغییرهای موردنیاز ایجاد و متدهای build_index، tf_df و compute_tfidf به ترتیب برای ساخت ایندکس معکوس، محاسبه tf و idf استفاده می‌شود.

۴ متد _preprocess

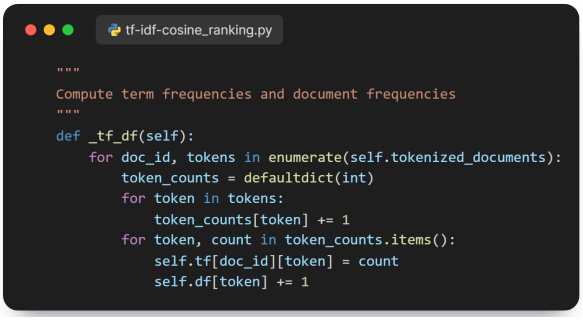
در این متد جداسازی کلمات از متن انجام می‌شود.

۵ متد _build_index



در این متد ایندکس معکوس ساخته‌شده و ذخیره می‌شود.

۶ متد _tf_df

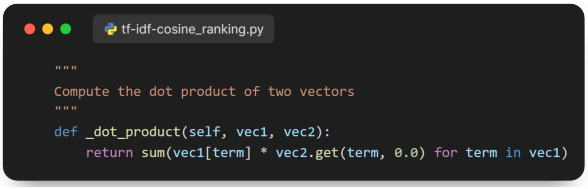


در این متد tf و df محاسبه و ذخیره می‌شوند.

۷ متد __compute__tfdf

در این متد امتیاز tf idf برای هر داکیومنت محاسبه و ذخیره می‌شود.

۸ متد __magnitude

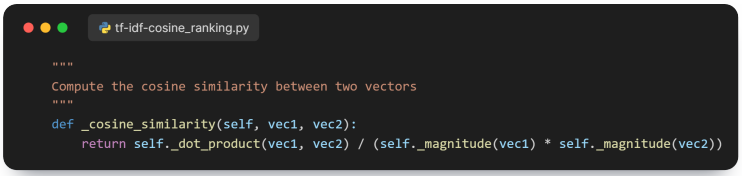


در این متد ضرب داخلی دو بردار محاسبه می‌شود.

۹ متد __magnitude

در این متد اندازه بردار محاسبه می‌شود.

۱۰ متد __cosine__similarity



در اینجا امتیاز شباهت دو بردار محاسبه می‌شود و به طور کلی فرمول آن به شرح زیر است:

$$cosine\_similarity = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

۱۱ متد search

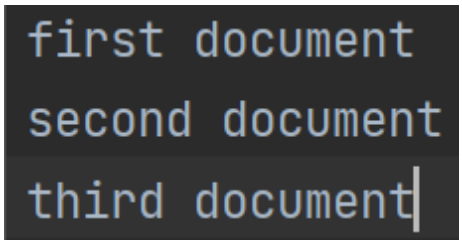
در این متد جست‌وجو براساس امتیازاتی که در متد __cosine__similarity محاسبه شد انجام و ۱۰ نتیجه اول به ترتیب تشابه برگشت‌داده می‌شود.

۱۲ استفاده

tf-idf-cosine_ranking.py

```python
if __name__ == '__main__':
    # Check command-line arguments
    if len(sys.argv) != 2:
        print("Usage: python BM25_ranking.py [file_name]")
        sys.exit(1)

    # Read documents from file
    filename = sys.argv[1]
    documents = []
    with open(filename, encoding="utf8") as file:
        for line in file:
            documents.append(line.strip())

    # Create an InvertedIndex object
    ii = InvertedIndex(documents)

    # Get query from user
    query = input('Search: ')

    # Search for documents matching the query
    results = ii.search(query)
    print("Results for query '%s':" % query)
    for doc_id, score in results[0: 11]:
        print("Doc %d: score: %.5f" % (doc_id, score))
```

برای استفاده از کلاس گفته شده این کد نوشته شده که بعد از فراخوانی کلاس ایجاد شده و پس از خواندن داکیومنت‌ها و ایجاد Inverted Index، بردار tf-idf ایجاد و در نهایت با دریافت ورودی با استفاده از فرمول تشابه cosine نتایج دریافت و چاپ می‌شود.
برای مثال برای ورودی مانند تصویر زیر:



خروجی مانند تصویر زیر تولید می‌شود.

```
Search: third
Results for query 'third':
Doc 2: score: 1.00000
Doc 0: score: 0.00000
Doc 1: score: 0.00000
```