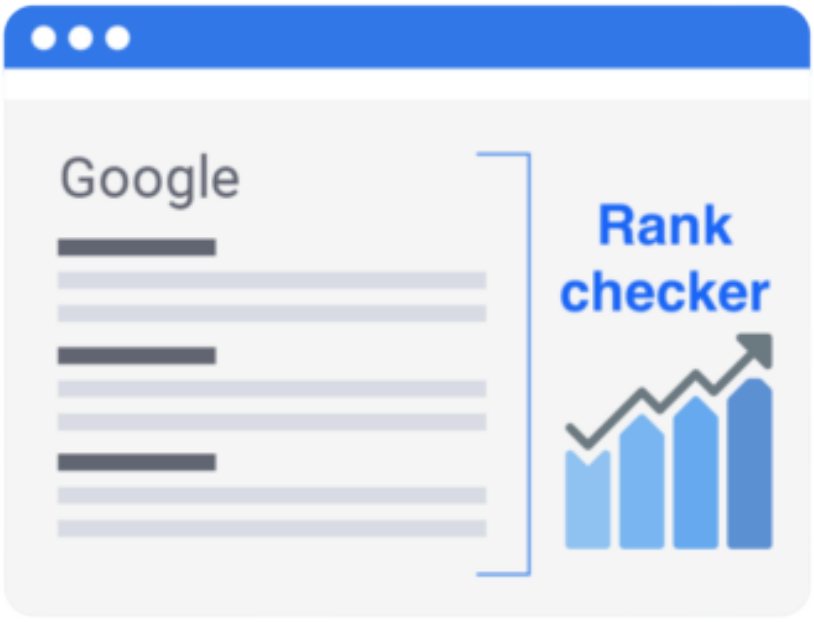
TF-IDF ranking



سیدمحمدحسین هاشمی ۴۰۲۲۳۶۳۱۴۳

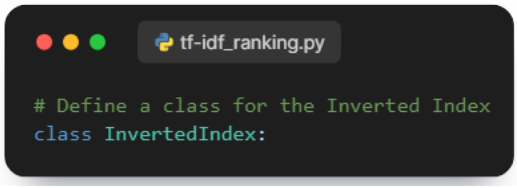
فروردین ۱۴۰۳

فهرست مطالب

۱ کتابخانه‌های مورد نیاز

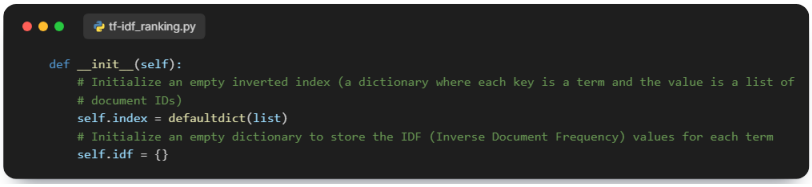
کتابخانه‌های مورد نیاز که در پروژه لود شده‌اند.

۲ کلاس Inverted Index



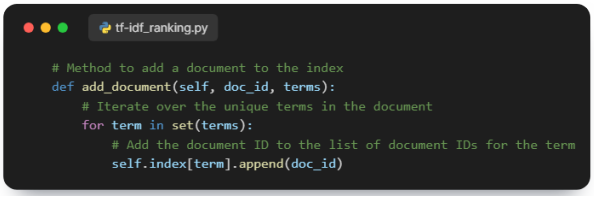
تمام کدهای مربوطه در این کلاس نوشته می‌شود

۳ متد –init–



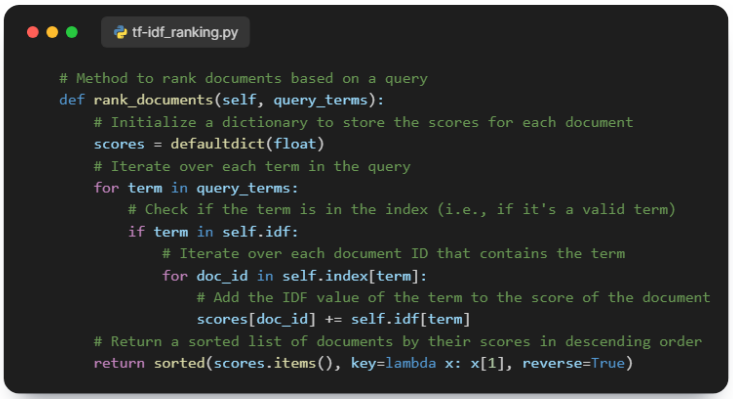
در اینجا یک کالکشن خالی برای ذخیره Inverted Index و همچنین یک دیکشنری برای ذخیره idf هر داکیومنت ایجاد می‌شود.

۴ متد add_document



در این متد عملیات اشتراک بین posting listها انجام می‌شود و نتیجه جست‌وجو برگشت داده می‌شود.

۵ متد rank_document



در این تابع براساس ورودی (query_terms) با فرمول tf-idf امتیازدهی انجام می‌شود و داکیومنت‌ها به‌ترتیب امتیاز برگشت‌داده می‌شوند.

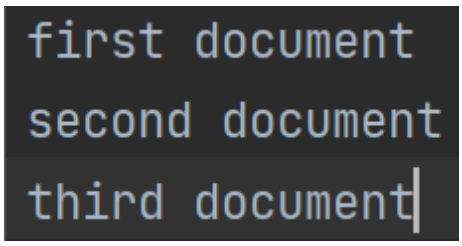
۶ استفاده

tf-idf_ranking.py

```python
# Usage
if __name__ == '__main__':
    # Check if the correct number of command-line arguments are provided
    if len(sys.argv) != 2:
        print("Usage: python tf-idf_ranking.py [file_name]")
        sys.exit(1)

    # Get the filename from the command-line argument
    filename = sys.argv[1]

    # Create an instance of the InvertedIndex class
    index = InvertedIndex()

    # Add documents from the file to the index
    with open(filename, encoding="utf8") as file:
        record_id = 0
        for line in file:
            record_id += 1
            # Split the line into terms using a regular expression and convert to lowercase
            terms = re.split('[^a-zA-z]', line.lower())
            # Add the document to the index
            index.add_document(record_id, terms)

    # Calculate the IDF values for each term
    index.calculate_idf()

    # Prompt the user to enter a query
    query = re.split('[^a-zA-z]', input('Search: ').lower())

    # Rank documents for the query
    ranked_docs = index.rank_documents(query)

    # Print the search results
    print(f"Search results for query '{' '.join(query)}':")
    for doc_id, score in ranked_docs[0: 11]:
        print(f"Document {doc_id}: TF-IDF score = {score:.4f}")
```

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 در آن با استفاده از رنکینگ tf-idf جست‌وجو انجام می‌شود و ۱۰ نتیجه اول به همراه امتیاز خروجی داده می‌شود.

برای مثال برای ورودی مانند تصویر زیر:



خروجی مانند تصویر زیر تولید می‌شود.

```
Search: third
Search results for query 'third':
Document 3: TF-IDF score = 1.6094
```