به نام خدا

تمرین ۵

درس بازیابی اطلاعات

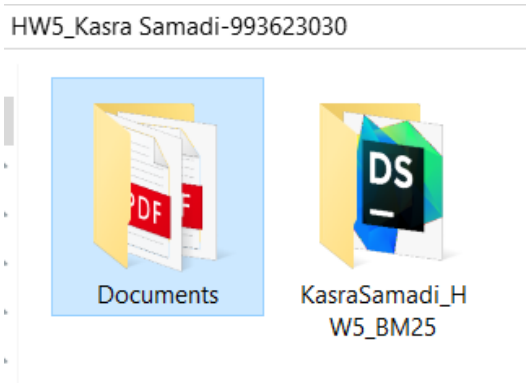
استاد : دکتر رضاپور

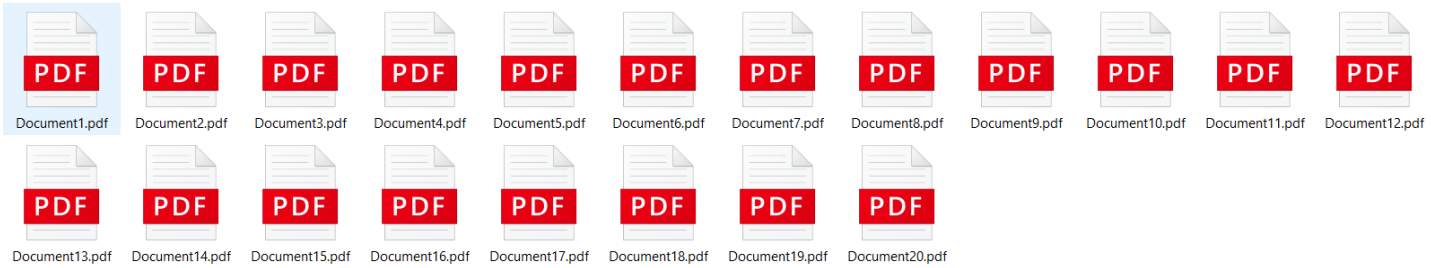
کسری صمدی<۹۹۳۶۲۳۰۳۰>

بهمن ۱۴۰۲

شما می‌بایست با استفاده از روش BM25 score یک سیستم بازیابی روی اسناد جمع آوری شده در تکالیف قبلی توسعه دهید. همچنین در گزارش تکلیف تاثیر ایجاد تغییر در فاکتورهای k و b را روی نتایج نشان داده شده سیستم بررسی کنید. دقت داشته باشید که کدها می بایست با زبان پایتون نوشته شود. برنامه می‌بایست قابل تست باشد وگرنه نمره‌ای به تکلیف تعلق نمی‌گیرد. لذا توضیحات کافی به صورت تصویری در رابطه با برنامه توسعه داده شده در قالب یک فایل ورد پیوست تکلیف باشد.

همانطور که در تمرین قبل خواسته شده بود، ابتدا ۲۰ مقاله با فرمت pdf با نام‌های Document1 تا Document20 دانلود و نام‌گذاری شده و در پوشه‌ای با نام Documents در کنار فایل اجرایی کد پایتون ذخیره شده است.





حال به سراغ بخش کد پایتون می‌رویم :

در این بخش ابتدا کتابخانه‌های موردنیاز را ایمپورت می‌کنیم.

BM25

```
import nltk
import PyPDF2
import string
import math
from collections import defaultdict
from ordered_set import OrderedSet
```

کتابخانه nltk : برای پردازش زبان طبیعی استفاده می‌شود و در این تمرین، برای tokenize کردن جملات موجود در مقالات استفاده شده است.

کتابخانه pyPDF2 : برای استخراج متن از مقالات کاربرد دارد.

کتابخانه string : برای کار با دیتاتایپ استرینگ استفاده می‌شود و در این تمرین، برای حذف علائم نگارشی (Remove punctuation) استفاده شده است.

کتابخانه math : برای انجام عملیات ریاضی استفاده می‌شود.

کتابخانه Defaultdict : نوعی از ساختمان داده دیکشنری است که مزیت‌های بیشتری نسبت به دیکشنری معمولی دارد.

کتابخانه ordered_set : یک داده ساختار ست (مجموعه) است که عناصر با ترتیب در آن ذخیره می‌شوند. همانطور که می‌دانید در مجموعه‌ها (Set)، عناصر تکراری ذخیره نمی‌شوند.

```
documents_folder_name = fr"..\Documents"
documents_name = "Document"
documents_count = 20
all_doc_ids = list()
```

حال در این قسمت، متغیرهایی را تعریف می‌کنیم که به ترتیب شامل نام فولدر ذخیره کننده مقالات (Documents)، نام ثابت برای هر مقاله (Document) و تعداد مقالات (۲۰) است.

متغیر all_doc_ids کل نام‌های داکیومنت‌ها را در خود ذخیره می‌کند.

در زیر سایر متغیرهای مورد نیاز این تمرین به همراه توضیحات کاربرد آن‌ها آورده شده است.

```
'''
dictionary to store the total words in each documents
# keys document names and values are the total words are there in the document
'''
documents_words_num = defaultdict(int)

'''
a dictionary to store the term frequencies (TF)
keys are words and values are defaultdict that keys are document name and
values are the word TF in the document
'''
tf = defaultdict(lambda: defaultdict(float))
```

```
'''
a dictionary to store the Inverse Document Frequency (IDF)
keys are words and values are the word idf in the collection
'''
idf = defaultdict(float)

'''
a dictionary to store the document frequencies (DF)
keys are words and values are the set of documents name that contain the word
'''
df = defaultdict(OrderedSet)

'''
a dictionary to store the TF-IDF scores
keys are words and values are defaultdict that keys are document name and
values are the word TF_IDF in the document
'''
tfidf = defaultdict(lambda: defaultdict(float))
```

```
for i in range(documents_count):
    doc_name = f"{documents_name}{i + 1}"
    doc_address = fr"{documents_folder_name}\{doc_name}.pdf"
    all_doc_ids.append(doc_name)

    number_of_words_in_doc = 0
    with open(doc_address, "rb") as pdf_file:
        read_pdf = PyPDF2.PdfReader(pdf_file)
        number_of_pages = len(read_pdf.pages)
```

در این قسمت، نام‌های مقالات (Document1 تا Document20) همچنین آدرس آن‌ها (..\Documents\Document1..20) به ازای هر مقاله ساخته می‌شود و با استفاده از کتابخانه PyPDF2 مقاله موردنظر خوانده شده و تعداد صفحات آن بدست می‌آید. همچنین نام‌های داکیومنت‌ها در لیست all_dic_ids ذخیره می‌شود. متغیر number_of_words_in_doc برای ذخیره تعداد کل کلمات موجود در هر داکیومنت تعریف شده است.

```
number_of_pages = len(read_pdf.pages)
for page_num in range(number_of_pages):
    page = read_pdf.pages[page_num]
    page_content = page.extract_text()

    lower_text = page_content.lower()
    text_without_punctuation_marks = lower_text.translate(str.maketrans('', '', string.punctuation))

    tokens = nltk.word_tokenize(text_without_punctuation_marks)
```

حال با استفاده از تعداد صفحات بدست آمده برای مقاله موردنظر، متن هر صفحه از آن استخراج می‌شود. سپس متن بدست آمده به حروف کوچک (lowercase) تبدیل شده و علائم نگارشی (punctuation marks) از آن حذف می‌شود.

بعد از آن، با استفاده از کتابخانه nltk، متن موردنظر tokenize (کلمه کلمه) می‌شود و کلمات به دست آمده از متن هر صفحه، در آرایه tokens ذخیره می‌شود.

```
        tokens = nltk.word_tokenize(text_without_punctuation_marks)

        for term in tokens:
            number_of_words_in_doc += 1
            tf[term][doc_name] += 1
            df[term].append(doc_name)

 documents_words_num[doc_name] = number_of_words_in_doc
```

حال به ازای هر کلمه موجود در آرایه tokens ، متغیر number_of_words_in_doc و دیکشنری‌های tf و df مقداردهی می‌شوند. در نهایت تعداد کل کلمات موجود در داکومنت، در دیکشنری documents_word_num ذخیره می‌شود.

تعداد کل کلمات موجود در هر داکیومنت به شرح زیر است:

```
for iden, words in documents_words_num.items():
    print(iden, words)
```

```
Document1 7621
Document2 1394
Document3 7380
Document4 4005
Document5 3434
Document6 4474
Document7 3351
Document8 4009
Document9 7129
Document10 4107
Document11 10184
Document12 7497
Document13 58356
Document14 4468
Document15 7689
Document16 1883
Document17 3012
Document18 2986
Document19 6825
Document20 3643
```

```
# calculate tf and idf for each words
for term, doc_tf in tf.items():
    for doc_id, term_freq in doc_tf.items():
        # calculate tf
        tf[term][doc_id] /= documents_words_num[doc_id]

        # calculate idf
        idf[term] = math.log(documents_count / (len(df[term]) + 1))
        print(term, doc_id, tf[term][doc_id], idf[term])
```

حال به ازای هر ترمی که در دیکشنری tf است، مقادیر tf و idf محاسبه می‌شود و به ترتیب در دیکشنری‌های tf و idf با کلیدهای مربوطه ذخیره می‌شوند.

در تصویر زیر می‌توان قسمتی از مقادیر tf و idf را به ازای هر ترم در داکیومنت‌هایی که شامل ترم موردنظر هستند، مشاهده کرد :

TF IDF

```
efficient Document1 0.0030179766434851 07 0.7985076962177716
efficient Document3 0.00027100271002710027 0.7985076962177716
efficient Document4 0.00024968789013732833 0.7985076962177716
efficient Document5 0.0037856726849155504 0.7985076962177716
efficient Document8 0.0002494387627837366 0.7985076962177716
efficient Document13 0.00022277058057440537 0.7985076962177716
efficient Document14 0.00022381378692927484 0.7985076962177716
efficient Document15 0.00013005592404734037 0.7985076962177716
video Document1 0.012203122949744127 1.3862943611198906
video Document2 0.0014347202295552368 1.3862943611198906
video Document3 0.0008130081300813008 1.3862943611198906
video Document6 0.0049172999552972736 1.3862943611198906
object Document1 0.009841228185277523 1.0498221244986776
object Document3 0.0004065040650406504 1.0498221244986776
object Document6 0.00022351363433169424 1.0498221244986776
object Document8 0.0004988775255674732 1.0498221244986776
object Document9 0.00014027212792818068 1.0498221244986776
object Document11 9.819324430479184e-05 1.0498221244986776
```

```
df["efficient"]
```

```
OrderedSet(['Document1', 'Document3', 'Document4', 'Document5', 'Document8', 'D
ocument13', 'Document14', 'Document15'])
```

در دیکشنری df کلیدها ترم‌ها هستند و مقادیر یک orderedSet است که شامل نام مقالاتی است که ترم مورد نظر در آن مقالات وجود دارد.

```
# Calculate the TF-IDF scores
for term, doc_tf in tf.items():
    for doc_id, term_freq in doc_tf.items():
        tfidf[term][doc_id] = term_freq * idf[term]
```

حال با توجه به مقادیر موجود در دیکشنری‌های tf و idf ، مقادیر tf_idf به ازای در ترم محاسبه شده و در دیکشنری tfidf ذخیره می‌شود.

قسمتی از مقادیر ذخیره شده در دیکشنری tfidf در تصویر زیر آمده است:

```
for term, doc_id in tfidf.items():
    print(term+ " => ",tfidf[term])
```

```
efficient =>  defaultdict(<class 'float'>, {'Document1': 0.002409877576828336,
'Document3': 0.00021639774965251263, 'Document4': 0.00019937770192703411, 'Doc
ument5': 0.003022888774266462, 'Document8': 0.00019917877181785276, 'Document1
3': 0.0001778840230795639, 'Document14': 0.00017871703138267047, 'Document15':
0.00010385065629051524})
video =>  defaultdict(<class 'float'>, {'Document1': 0.01691712053328301, 'Doc
ument2': 0.00198894456401706, 'Document3': 0.0011270685862763339, 'Document6':
0.0068168251999637})
object =>  defaultdict(<class 'float'>, {'Document1': 0.010331539081144314, 'D
ocument3': 0.00042675696117832425, 'Document6': 0.00023464955844851982, 'Docum
ent8': 0.0005237326637558881, 'Document9': 0.00014726078334951294, 'Document11
': 0.00010308544034747424})
segmentation =>  defaultdict(<class 'float'>, {'Document1': 0.011848571096240023, 'Document4': 0.00030061742929486543, 'Document5': 0.0007012072244181341, '
Document6': 0.004843878068365411, 'Document7': 0.009700765657057677})
```

برای محاسبه نتیجه و بازیابی نام داکیومنت‌ها با روش BM25 score ، تابع MB25 را با ورودی‌های query ، k، b و avgdl مطابق تصویر زیر تعریف می‌شود:

```
def BM25(query, k = 2, b = 1, avgdl = 0):

    query_terms = query.lower().split() # Tokenize the query

    '''
    a dictionary to store the document score
    keys are document names and values are the document score
    '''
    doc_scores = defaultdict(float)

    '''
    a dictionary to store TF for each term are there in query
    keys are terms in query and values are term TF
    '''
    tf_query_term = defaultdict(float)

    '''
    a dictionary to store count duplicate terms are there in query
    keys are terms in query and values are counts
    '''
    term_count = defaultdict(float)
```

```
    '''
    a dictionary to store TF_IDF for each term are there in query
    keys are terms in query and values are term TF_IDF
    '''
    tfidf_query_term = defaultdict(float)
```

حال مطابق تصویر زیر tf_idf برای هر ترم موجود در query محاسبه می‌شود و در دیکشنری tfidf_query_term ذخیره می‌شود :

```
# calculate tf_idf for query tokens
for term in query_terms:
    term_count[term] += 1
for term, count in term_count.items():
    tf_query_term[term] = count / len(query_terms)
for term, tf_value in tf_query_term.items():
    idf_value = idf[term]
    tfidf_query_term[term] = tf_value * idf_value
```

در نهایت به ازای هر داکیومنت و به ازای ترم‌های موجود در query ، امتیاز کوئری با روش BM25 برای هر داکیومنت محاسبه شده و در دیکشنری doc_scores به صورت نام داکیومنت و امتیاز محاسبه شده آن داکیومنت، ذخیره می‌شود. داکیومنت‌ها با توجه به امتیازهایشان، مرتب شده و نتیجه ریترن می‌شود. کد این عملیات در تصویر زیر آمده است:

```
    # calculate score of each document for user query
    query_terms = set(query_terms) # ignore same and duplicate words
    for document_name in all_doc_ids:
        document_score = 0
        for term in query_terms:
            if term in tf:
                term_idf = idf[term]
                for doc_name, termferq in tf[term].items():
                    if doc_name == document_name:
                        docLength = documents_words_num[document_name]
                        tf_component = ((termferq * (k + 1)) \
                                        / (termferq + k * (1 - b + b * (docLength / avgdl))))
                        document_score += term_idf * tf_component
                        break

            doc_scores[document_name] = document_score

    # Sort the documents by their relevance scores
    sorted_docs = sorted(doc_scores.items(), key=lambda x: x[1], reverse=True)
    return sorted_docs
```

حال مطابق کد زیر، avgdl که میانگین طول سند در مجموعه اسناد هست را محاسبه می‌کنیم. این مقدار برابر است با مجموع کل توکن‌ها (کلمات) در داکیومنت‌ها تقسیم بر تعداد داکیومنت‌های موجود (۲۰)

```
num_docs_words = sum(words_count for words_count in documents_words_num.values())
avgdl = num_docs_words / len(documents_words_num)
```

حال تابع BM25 را به ازای کوئری‌های مختلف و مقادیر k = 2 (مقیاس فراوانی کلمات) و b=1 (تاثیر نرمال‌سازی طول سند را کنترل می‌کند) فراخوانی می‌کنیم و نتایج به ازای بعضی از کوئری‌های ورودی به شرح زیر است:

کوئری ورودی : the

```
n = 20 # return Top n of documents
query = "the"
# query = "kasra"
# query = "ySF"

results = BM25(query, k = 2, b = 1, avgdl = avgdl)
print(f"Top {n} most relevant documents:")
for doc_id, score in results[:n]:
    if score !=0:
        print(f"{doc_id}: {score}")
```

```
Top 20 most relevant documents:
Document13: -0.0005304837759401666
Document3: -0.00319674750070999
Document1: -0.0034180209973787294
Document15: -0.004528239741392427
Document9: -0.004601412397754279
Document19: -0.004747037829458164
Document12: -0.004839631164540966
Document11: -0.004993929648208769
Document6: -0.005662954975007843
Document7: -0.0056718089850938845
Document5: -0.0064193829970557925
Document10: -0.00679324694333479
Document14: -0.008028476229743915
Document4: -0.00819548300199195
Document8: -0.009475698397222137
Document17: -0.010725873080656397
Document20: -0.011212529598737375
Document18: -0.011757172182246969
Document2: -0.017323289807246997
Document16: -0.01865424547524403
```

کوئری ورودی : kasra

```
n = 20 # return Top n of documents
# query = "the"
query = "kasra"
# query = "γSF"

results = BM25(query, k = 2, b = 1, avgdl = avgdl)
print(f"Top {n} most relevant documents:")
for doc_id, score in results[:n]:
    if score !=0:
        print(f"{doc_id}: {score}")
```

```
Top 20 most relevant documents:
```

داکیومنتی شامل کلمه kasra وجود ندارد.

کوئری ورودی : γSF

```
n = 20 # return Top n of documents
# query = "the"
# query = "kasra"
query = "γSF"

results = BM25(query, k = 2, b = 1, avgdl = avgdl)
print(f"Top {n} most relevant documents:")
for doc_id, score in results[:n]:
    if score !=0:
        print(f"{doc_id}: {score}")
```

```
Top 20 most relevant documents:
Document17: 0.020386368097990614
```

در ۲۰ داکیومنت، کلمه γSF فقط در Document17 وجود دارد.

$$score(q,d) = \sum_{i=1}^{|q|} idf(q_i) \cdot \underbrace{\frac{tf(q_i,d) \cdot (k_1+1)}{tf(q_i,d) + k_1(1-b+b \cdot \frac{|d|}{avgdl})}}_{TF25}$$

با تغییر مقدار k و b برای کوئری‌های قبل، نتایج ریترن شده مرتب‌شده داکیومنت‌ها، تقریبا مشابه نتایج قبل است. برای مثال برای کوئری γSF و مقدار k = 5 و b = 0.5، همان داکیومنت Document17 به عنوان خروجی نمایش داده می‌شود که نتیجه‌ای درست است.

```
n = 20 # return Top n of documents
# query = "the"
# query = "kasra"
query = "γSF"

results = BM25(query, k = 5, b = 0.5, avgdl = avgdl)
print(f"Top {n} most relevant documents:")
for doc_id, score in results[:n]:
    if score !=0:
        print(f"{doc_id}: {score}")
```

```
Top 20 most relevant documents:
Document17: 0.009216382418017977
```