دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش زیربخش دوم فاز اول پروژه بازیابی اطلاعات

نگارش
مهدیه سادات بنیس

استاد درس
دکتر احمد نیک آبادی

نیم سال دوم ۱۴۰۱

۳.۱.ساخت شاخص مکانی

با استفاده از اسناد پیش‌پردازش‌شده در زیر فاز قبل، یک شاخص مکانی می سازیم. در شاخص مکانی که ساختیم موارد زیر را برای هر کلمه در دیکشنری مشخص کردیم:

- تعداد تکرار آن کلمه در کل اسناد
- جایگاه کلمه در هر سند
- تعداد تکرار کلمه در هر سند

ساختمان داده استفاده شده برای پیاده سازی شاخص مکانی dictionary می باشد.

dictionary ها برای ذخیره مقادیر داده در جفت های key:value استفاده می شوند.dictionary کالکشنی است که ترتیب دارد، قابل تغییر است و اجازه تکرار ندارد.

پیاده سازی:

ورودی تابع ساخت شاخص مکانی کل اسناد پیش پردازش شده است.

روی اسناد پیمایش کرده و برای هر سند روی توکن های آن(که در قسمت content هستند) عملیات زیر را انجام میدهیم:

۱- اگر توکن در دیکشنری ما بود:

۱-۱- اگر سند در سند هایی که توکن در آن ها بوده از قبل در دیکشنری باشد تنها لازم است به جایگاه فعلی آن را به جایگاه توکن در سند اضافه کنیم و همچنین به تعداد تکرار کلمه در سند یکی اضافه کنیم.

۱-۲- اگر سند در سند هایی که توکن در آن ها هنوز اضافه نشده باشد اولین جایگاه را اضافه میکنیم و تعداد تکرار کلمه در سند را یک قرار میدهیم.

تعداد تکرار کلمه در کل اسناد یکی اضافه میکنیم.

۲- اگر توکن در دیکشنری از قبل نباشد، تعداد تکرار کلمه در کل اسناد را یک قرار داده و در قسمت سند ها اطلاعات(جایگاه در سند و تعداد تکرار = ۱) اولین سند که توکن در آن قرار دارد را در دیکشنری ایجاد میکنیم.

```
def Postings_List(Docs):
    my_dict = {}
    for index in Docs:
        for position, token in enumerate(Docs[index]['content']):

            if token in my_dict:   ۱
                |
                if index in my_dict[token]['docs']:    ۱-۱
                    my_dict[token]['docs'][index]['positions'].append(position)
                    my_dict[token]['docs'][index]['number_of_token'] += 1
                else:
                    my_dict[token]['docs'][index] = {  ۱-۲
                                                     'positions': [position],
                                                     'number_of_token': 1
                                                     }
                my_dict[token]['frequency'] += 1
            else:
                my_dict[token] = {    ۲
                 'frequency': 1,
                 'docs': {
                       index: {
                           'positions': [position],
                           'number_of_token': 1
                           }
                   }
                }
    return my_dict
```

```
dic = Postings_List(pre_processed_docs)
```

نمونه خروجی برای کلمه “فارس”

```
dic['فارس']
```

```
{'frequency': 13859,
 'docs': {'0': {'positions': [2], 'number_of_token': 1},
  '1': {'positions': [2], 'number_of_token': 1},
  '2': {'positions': [2], 'number_of_token': 1},
  '3': {'positions': [2], 'number_of_token': 1},
  '4': {'positions': [4], 'number_of_token': 1},
  '5': {'positions': [2], 'number_of_token': 1},
  '6': {'positions': [2, 19, 57], 'number_of_token': 3},
  '7': {'positions': [4], 'number_of_token': 1},
  '8': {'positions': [4, 98, 246], 'number_of_token': 3},
  '9': {'positions': [4], 'number_of_token': 1},
  '10': {'positions': [2], 'number_of_token': 1},
  '11': {'positions': [4], 'number_of_token': 1},
  '12': {'positions': [2], 'number_of_token': 1},
  '13': {'positions': [4], 'number_of_token': 1},
  '14': {'positions': [2], 'number_of_token': 1},
  '15': {'positions': [4], 'number_of_token': 1},
  '16': {'positions': [2], 'number_of_token': 1},
  '17': {'positions': [2], 'number_of_token': 1},
```