به نام خدا

دانشگاه صنعتی امیرکبیر

دانشکده‌ی مهندسی کامپیوتر

پروژه بازیابی اطلاعات

فاز ۱ زیربخش دوم

اروند درویش

۹۸۳۱۱۳۷

```
posDics = []

for i, doc in enumerate(preprocessedDocs):
    posDics.append(defaultdict(list))
    for j, token in enumerate(doc):
        posDics[i][token].append(j)
```

ما برای هر سند که هرکدام آنها (پس از پردازش اولیه که بصورت چندین ترم درآمده اند) یک المنت از لیست preprocessedDocs هستند، یک دیکشنری ساخته ایم که کلید آن ترم های آن سند است و value آن یک لیست است از شاخص مکانی آن که این پوزیشن ها را با حلقه ای که در تصویر رو به رو مشاهده می کنید، برای هر ترم بدست می آوریم.

در نهایت ما به تعداد اسنادمان، دیکشنری خواهیم داشت که آنها را در لیستی به نام posDics ذخیره کرده ایم و این دیکشنری ها به ترتیب اسناد هستند یعنی posDics[0] دیکشنری (key : ترم ها و value : لیست پوزیشن ها) متعلق به سند id = 0 است.

```
bigDic = {}
docFreq = []
for i, doc in enumerate(preprocessedDocs):
    docFreq.append(dict())
    for term in doc:
        if term in bigDic:
            bigDic[term] += 1
        else:
            bigDic[term] = 1
        if term in docFreq[i]:
            docFreq[i][term] += 1
        else:
            docFreq[i][term] = 1
```

در اینجا برای پیدا کردن تعداد تکرار کلمات در کل اسناد دیکشنری bigDic را ساخته ایم که کلید های آن تمام ترم های کل اسناد است و value های آن تعداد دفعات تکرار این ترم هاست.

برای پیدا کردن تعداد تکرار کلمات در هر سند هم یک لیستی از دیکشنری ها درست کرده ایم که به ترتیب مثلا اولین عضو لیست دیکشنری مربوط به اولین سند ماست (docID = 0) و در این دیکشنری کلید کلمات آن داک و value تعداد دفعات تکرار آن کلمات در آن داک است.

به عنوان مثال برای کلمه "فوتبال" در ۱۰ سند اول اگر برنامه را ران کنیم خروجی ما بصورت زیر خواهد بود :

```
total freq = 8175
document 0 : freq = 2   positions = [4, 10]
document 1 : freq = 2   positions = [26, 33]
document 2 : freq = 1   positions = [34]
document 3 : freq = 7   positions = [17, 38, 62, 74, 95, 116, 142]
document 7 : freq = 1   positions = [8]
document 9 : freq = 1   positions = [34]
```

در صفحه ی بعد مثال جامع تری از کلمه فوتبال میبینید که بین سند های با ۴۹۹۰ تا ۵۰۷۶، تعداد دفعات تکرار و شاخص های مکانی آن را مشاهده می کنید.

```
58              print("document " + str(i) + " : " + "freq = " + str(docFreq[i][word]
59
60      posIndShow('فوتبال')
```

PROBLEMS 17 OUTPUT TERMINAL DEBUG CONSOLE

```
document 4990 : freq = 1   positions = [38]
document 5001 : freq = 1   positions = [4]
document 5005 : freq = 11   positions = [18, 33, 46, 66, 84, 121, 151, 159, 170, 216, 255]
document 5006 : freq = 3   positions = [11, 96, 151]
document 5008 : freq = 5   positions = [60, 186, 307, 330, 468]
document 5009 : freq = 1   positions = [5]
document 5014 : freq = 1   positions = [4]
document 5015 : freq = 1   positions = [162]
document 5016 : freq = 2   positions = [245, 308]
document 5017 : freq = 2   positions = [150, 294]
document 5018 : freq = 1   positions = [8]
document 5020 : freq = 2   positions = [27, 55]
document 5028 : freq = 1   positions = [58]
document 5029 : freq = 11   positions = [29, 49, 55, 68, 84, 132, 145, 152, 156, 230, 258]
document 5030 : freq = 2   positions = [10, 52]
document 5032 : freq = 3   positions = [5, 88, 113]
document 5034 : freq = 1   positions = [47]
document 5040 : freq = 1   positions = [11]
document 5041 : freq = 1   positions = [4]
document 5042 : freq = 2   positions = [43, 203]
document 5043 : freq = 3   positions = [152, 251, 277]
document 5049 : freq = 1   positions = [4]
document 5050 : freq = 1   positions = [4]
document 5051 : freq = 1   positions = [4]
document 5052 : freq = 1   positions = [4]
document 5056 : freq = 1   positions = [5]
document 5057 : freq = 1   positions = [7]
document 5058 : freq = 1   positions = [7]
document 5059 : freq = 1   positions = [4]
document 5062 : freq = 1   positions = [35]
document 5063 : freq = 11   positions = [8, 67, 105, 123, 132, 143, 159, 167, 179, 185, 266]
document 5065 : freq = 1   positions = [5]
document 5067 : freq = 1   positions = [6]
document 5071 : freq = 2   positions = [4, 30]
document 5072 : freq = 3   positions = [5, 85, 128]
document 5074 : freq = 1   positions = [42]
document 5075 : freq = 3   positions = [5, 23, 32]
document 5076 : freq = 1   positions = [356]
```

پایان