گزارش مرحلهٔ دوم فاز اول پروژهٔ درس بازیابی اطلاعات – ساخت شاخص مکانی

نگار موقتیان، ۹۸۳۱۰۶۲

نحوه ساخت شاخص مکانی را با ذکر نمونه خروجی نمایش دهید.

برای ساخت شاخص مکانی در ابتدا چند کلاس تعریف شده‌اند تا بتوانند اطلاعات مربوط به قسمت‌های مختلف شاخص را نگهداری کنند.
ساختمان دادهٔ کلی استفاده شده dictionary است که کلمات (term) مختلف را به یک شیء از کلاس Index نگاشت می‌کند (مزیت استفاده از dictionary در سرعت بالای دسترسی به value مربوط به هر یک از term هاست). کلاس Index خود تعداد اسنادی که کلمهٔ فوق در آن‌ها تکرار شده (doc. frequency) و همچنین یک لیست از اشیاء کلاس Postings را نگهداری می‌کند (postings list). کلاس Postings نیز یک شماره سند دارد که نشان می‌دهد کلمهٔ فوق در آن سند چند بار (term frequency) و در چه مکان‌هایی تکرار شده.

اولین قدم برای ساخت شاخص مکانی مرتب‌سازی توکن‌های بدست آمده از مرحلهٔ قبل بر اساس حروف الفباست تا بتوانیم کلمات یکسان را یکی کرده و شاخص مکانی هر یک را بسازیم.
پس از آن به ترتیب بر روی تمام توکن‌ها پیمایش کرده و مانند زیر عمل می‌کنیم:

۱. یک Index جدید بساز
۲. تا وقتی که ترم توکن‌ها یکسان است:
۳. یک Postings جدید بساز
۴. تا وقتی که شمارهٔ اسناد یکسان است:
۵. توکن را به Postings اضافه کن
۶. Postings را به Index اضافه کن
۷. ترم را به همراه Index به دیکشنری شاخص مکانی اضافه کن

با این کار همان ساختاری میان کلاس‌ها که توضیح داده شد رعایت خواهد شد. لیست مکان کلمات یکسان در اسناد یکسان درون یک Postings مشخص قرار خواهد گرفت و لیست اسناد با کلمات یکسان نیز درون یک Index مشخص قرار خواهد گرفت. در نهایت نیز به تدریج با افزودن کلمات متفاوت دیکشنری ساخته خواهد شد.

در نهایت خروجی برنامه به ازای توکن‌های مرتب‌شدهٔ زیر:

```
sorted tokens:  [0 - 6 : AFC, 3 - 146 : یتآ, 3 - 44 : مارآ, 3 - 31 : وزرآ, 2 - 50 : دازآ, 0 - 5 : ایهآ, 0 - 21 : ایهآ, 0 - 35 : ایهآ, 0 - 57 : ایهآ, 3
 - 49 : یایهآ, 3 - 131 : یایهآ, 3 - 28 : هدامآ, 3 - 46 : هدامآ, 3 - 67 : هدامآ, 3 - 145 : هدامآ, 3 - 18 : هدامآ, 4 - 40 : یگدامآ, 0 - 65 : نهآ, 3
 - 36 : قالخا, 3 - 144 : ودرا, 3 - 102 : دنمشزرا, 0 - 24 : سامسا, 2 - 8 : لالقتحسا, 2 - 30 : لالقتحسا, 2 - 24 : دنخسا, 1 - 20 : عالطا, 1 - 17 : راهظا, 1
 - 27 : راهظا, 3 - 128 : مازعا, 3 - 79 : افرعا, 3 - 98 : افرعا, 0 - 23 : مالعا, 3 - 50 : مالعا, 3 - 10 : الکیپکی, 3 - 23 : الکیپکی, 3 - 42 : الکیپکی, 3 - 69
 : الکیپکی, 3 - 90 : الکیپکی, 3 - 109 : الکیپکی, 3 - 142 : الکیپکی, 3 - 148 : الکیپکی]
```

مانند زیر می‌باشد:

```
positional index:  {'AFC': #1 -> [0 : #1 -> [6]], 'یتآ': #1 -> [3 : #1 -> [146]], 'مارآ': #1 -> [3 : #1 -> [44]], 'وزرآ': #1 -> [3 : #1 -> [31]], 'ازآ
د': #1 -> [2 : #1 -> [50]], 'ایهآ': #1 -> [0 : #4 -> [5, 21, 35, 57]], 'یایهآ': #1 -> [3 : #2 -> [49, 131]], 'هدامآ': #1 -> [3 : #4 -> [28, 46, 67, 1
45]], 'یگدامآ': #1 -> [3 : #1 -> [18]], 'نهآ': #1 -> [4 : #1 -> [40]], 'روآ#درو': #1 -> [0 : #1 -> [65]], 'قالخا': #1 -> [3 : #1 -> [36]], 'ودرا': #1
 -> [3 : #1 -> [144]], 'دنمشزرا': #1 -> [3 : #1 -> [102]], 'سامسا': #1 -> [0 : #1 -> [24]], 'لالقتحسا': #1 -> [2 : #2 -> [8, 30]], 'دنخسا': #1 -> [2 : #
1 -> [24]], 'عالطا': #1 -> [1 : #1 -> [20]], 'راهظا': #1 -> [1 : #2 -> [17, 27]], 'مازعا': #1 -> [3 : #1 -> [128]], 'افرعا': #1 -> [3 : #2 -> [79, 98]]
, 'مالعا': #2 -> [0 : #1 -> [23], 3 : #1 -> [50]], 'الکیپکی': #1 -> [3 : #8 -> [10, 23, 42, 69, 90, 109, 142, 148]]}
```

* در مثال بالا تنها بخشی از توکن‌ها از اسناد ۱ تا ۵ آورده شده‌اند.