



## بازیابی پیشرفته اطلاعات

نیم سال اول ۱۴۰۲-۰۳

مدرس: دکتر حمید بیگی

۸ آبان

زمان: ۲۵ دقیقه

کوییز دوم (۱۰۰ نمره)

سوال ۱: فشرده سازی (۳۰ نمره)

برای سوالات زیر پاسخ کوتاه بنویسید.

- فرض کنید posting list یک کلمه در یک داک، به شکل زیر باشد:

[6, 29, 170, 238]

با استفاده از روش فشرده سازی variable byte این posting list را فشرده کنید.  
پاسخ ابتدا باید فاصله ها را محاسبه کنیم:

6, 29, 170, 238 -&gt; 6, 23, 141, 68

سپس باید هر عدد را به فشرده سازی "variable byte" تبدیل کنیم:

6 -&gt; 10000110 - 23 -&gt; 10010111 - 141 -&gt; 1000000100001101 - 68 -&gt; 11000100

بنابراین پاسخ نهایی این است:

1000011010010111100000010000110111000100

- فرض کنید posting list یک کلمه در یک داک، به شکل زیر و با روش gamma code فشرده شده است:

1100111100011111000001111110000010

با توضیح مراحل میانی، posting list اولیه را به دست آورید.  
پاسخ رمزگشایی را از چپ به راست شروع می کنیم:

110 -&gt; 2 - 01 = 5

1110 -&gt; 3 - 001 = 9

11110 -&gt; 0000 = 16

1111110 -&gt; 6 - 000010 = 66

بنابراین "posting list" نهایی این است:

5 -&gt; 14 -&gt; 30 -&gt; 96

سوال ۲: رتبه بندی سیستم های بازیابی (۲۵ نمره)

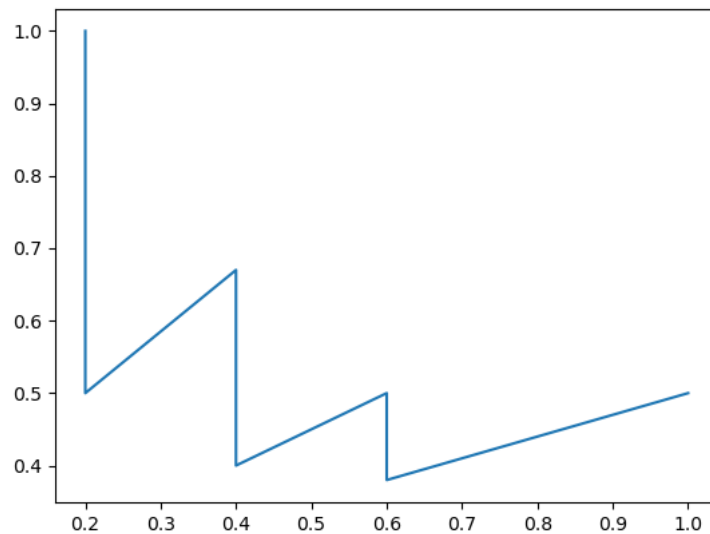
فرض کنید یک سیستم بازیابی رتبه بندی زیر را از چپ به راست ارائه داده است:

R, N, R, N, N, R, N, N, R, R

که R نشان دهنده مرتبط بودن است و N نشان دهنده غیر مرتبط بودن. نمودار P-R curve را برای این سیستم بکشید و همچنین mean average precision را محاسبه کنید.

پاسخ باید P و R را مرحله به مرحله محاسبه کنیم و در هر مرحله یک نتیجه جدید اضافه کنیم.

R: 0.2, 0.2, 0.4, 0.4, 0.4, 0.6, 0.6, 0.6, 0.8, 1  
P: 1, 0.5, 0.67, 0.5, 0.4, 0.5, 0.43, 0.38, 0.44, 0.5  
AP:  $(1 + 0.67 + 0.5 + 0.44 + 0.5) / 5 = 0.62$



### سوال ۳: امتیاز دهی در فضای برداری (۴۵ نمره)

فرض کنید یک سیستم بازیابی اطلاعات با مدل Inc.ltn دارید. همچنین دو document و یک query به شکل زیر به شما داده شده است. با توجه به معیار گفته شده، امتیاز query را برای هر یک از ها document محاسبه کنید. فرض کنید کل corpus تنها شامل این دو document می شود. همچنین مراحل میانی را از جمله به دست آوردن ... tf, idf, بنویسید.

Doc1: hi good course laugh all time

Doc2: bye take life easy it is an easy course

Query: laugh to life easy course

پاسخ

N = 2

Doc1				
normalize	weight	tf-weighted	tf-raw	term
$\frac{1}{2.45} = 0.41$	1	1	1	hi
$\frac{1}{2.45} = 0.41$	1	1	1	good
$\frac{1}{2.45} = 0.41$	1	1	1	course
$\frac{1}{2.45} = 0.41$	1	1	1	laugh
$\frac{1}{2.45} = 0.41$	1	1	1	all
$\frac{1}{2.45} = 0.41$	1	1	1	time
Doc2				
normalize	weight	tf-weighted	tf-raw	term
$\frac{1}{2.9} = 0.34$	1	1	1	bye
$\frac{1}{2.9} = 0.34$	1	1	1	take
$\frac{1}{2.9} = 0.34$	1	1	1	life
$\frac{1.3}{2.9} = 0.45$	1.3	1.3	2	easy
$\frac{1}{2.9} = 0.34$	1	1	1	it
$\frac{1}{2.9} = 0.34$	1	1	1	is
$\frac{1}{2.9} = 0.34$	1	1	1	an
$\frac{1}{2.9} = 0.34$	1	1	1	course

Query					
weight	idf	df	tf-weighted	tf-raw	term
0.3	0.3	1	1	1	laugh
0	0	0	1	1	to
0.3	0.3	1	1	1	life
0.3	0.3	1	1	1	easy
0	0	2	1	1	course

Doc1 score =  $0.3 \times 0.41 + 0 \times 0.41 = 0.123$

Doc2 score =  $0.3 \times 0.34 + 0.3 \times 0.45 + 0 \times 0.34 = 0.237$