سوال ۱

یک مجموعه شامل ۸۰۶۷۹۱ سند داریم. جدول ۱ بیانگر تعداد رخداد عبارات "خودرو"، "صدا"، "بیمه" و "بهترین" در سه سند $Doc1$، $Doc2$ و $Doc3$ است و جدول ۲ تعداد اسنادی را نشان می‌دهد که این کلمات در آنها ظاهر شده است. با توجه به این اطلاعات $tf - idf$ مربوط به هر کلمه در هر سند را با استفاده از رابطه 6.14[1] کتاب IIR محاسبه نمایید

	$Doc1$	$Doc2$	$Doc3$
خودرو	27	4	24
صدا	3	33	0
بیمه	0	33	29
بهترین	14	0	17

جدول ۱

$terms$	$\mathbf{df_t}$
خودرو	18165
صدا	6723
بیمه	19241
بهترین	25235

جدول ۲

پاسخ :

$$\mathrm{idf}_t = \log_{10}\left(N/\mathrm{df}_t\right)$$

$terms$	$\mathbf{df_t}$	$\mathbf{idf_t}$
خودرو	**18165**	**1.65**
صدا	**6723**	**2.08**

[1] $\left(1 + log(tf_{t,d})\right) * idf_t$

بیمه	**19241**	**1.62**
بهترین	**25235**	**1.5**

$$\boldsymbol{w_{t,d} = \left(1 + log\left(tf_{t,d}\right)\right) * \ idf_t}$$

$$\boldsymbol{w_{\text{خودرو},doc1} = \left(1 + log(27)\right) * \ 1.65 = 4.011}$$

و به همین ترتیب خواهیم داشت:

	$Doc1$	$Doc2$	$Doc3$
خودرو	4.011	2.64	3.92
صدا	3.07	5.23	0
بیمه	0	4.07	3.98
بهترین	3.21	0	3.34

سوال ۲

با فرض آنکه ورودی موتور جستجو همواره فقط عبارت‌های تک کلمه‌ای هستند به سوالات زیر پاسخ دهید.

الف) چرا استفاده از یک $global\ champion\ list$ با $r = K$ برای آنکه بتوانیم K سند با بالاترین امتیاز را شناسایی کنیم کافی است؟

ب) با تغییر راهکار بخش الف، راهکاری ساده برای پاسخ دهی به عبارات S کلمه‌ای درصورتی که $S > 1$ باشد ارائه دهید.(راهکار ارائه شده می تواند امن نباشد.)

پاسخ:

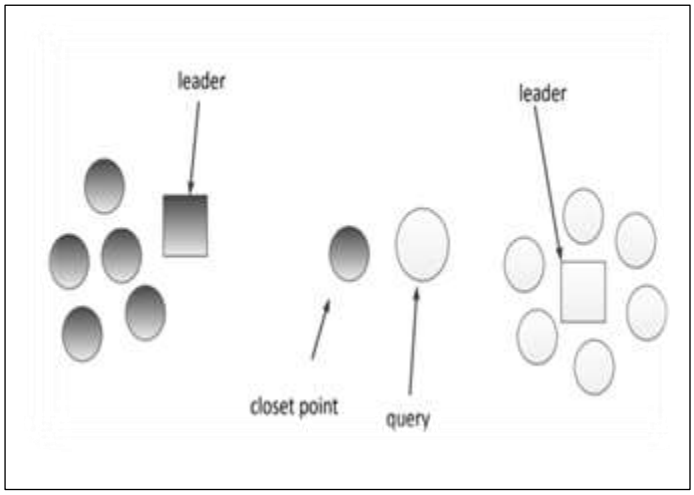
الف) در هنگام محاسبه، به ازای هر کلمه در عبارت جستجو شده، $champion\ list$ آن را آورده و این ها را اجتماع میگیریم و محاسبه تشابه کسینوسی را به این مجموعه محدود میکردیم، حال که عبارت جستجو فقط یک کلمه است، همان $champion\ list$، K تا بهترین جواب است.

ب) برای هر کلمه تعداد $\left\lceil \frac{K}{S} \right\rceil$ از بالاترین امتیاز ها را برمیداریم.

سوال ۳

یکی از روش های بهبود سرعت پاسخ دهی به پرسمان کاربر استفاده از خوشه بندی است. در این روش اسناد در خوشه‌های مختلف قرار می‌گیرند و پرسمان درابتدا تنها با مراکز خوشه‌ها مقایسه می‌شود و سپس k سند مرتبط را در خوشه‌ی نظیر نزدیک‌ترین مرکز جستجو می‌کنیم. با ذکر مثالی $non - safe$ بودن این روش را نشان دهید. می‌توانید از ترسیم خوشه‌ها، مراکز آنها و پرسمان کاربر در یک فضای برداری دو بعدی استفاده کنید.

پاسخ :



در مثالی که در تصویر دیده می شود در هنگام خوشه بندی دو مرکز دسته انتخاب شده(که به صورت مربع نشان داده شده است) و پرسمان درابتدا تنها با آنها مقایسه خواهد شد و پس از آن با اعضای خوشه ای که مرکز آن نزدیک ترین به پرسمان بوده است مقایسه خواهد شد، لذا همان طور که در شکل می بینیم با نزدیک ترین سند نسبت به پرسمان هیچ گاه مقایسه ای صورت نخواهد گرفت و این سند بازیابی نخواهد شد.

سوال ۴

رشته زیر نتیجه جستجوی عبارتی در بین ۱۰۰۰۰ سند را نشان میدهد که R نشان دهنده مرتبط و N نشان دهنده نامرتبط است. ترتیب این پاسخ‌ها از چپ به راست است (سیستم معتقد است چپ ترین پاسخ مرتبط ترین است) با فرض آنکه تنها ۸ سند مرتبط در میان اسناد وجود دارد به سوالات زیر پاسخ دهید.

$$RRNNNNNNRNRNNNRNNNNR$$

الف) دقت ($precision$) این سیستم را محاسبه نمایید.

ب) مقدار F_1 برای این پاسخ چند است؟

ج) فرض کنید این ۲۰ سند، کل پاسخ سیستم باشند، مقدار MAP را برای این پرسمان محاسبه کنید.

د) کمترین مقدار ممکن MAP برای این سیستم باتوجه به شرایط زیر چقدر است؟ چرا؟

*۲۰سند اول بازیابی شده سیستم به صورت رشته ذکر شده باشد.

*محدودیتی بر روی تعداد سند بازیابی شده وجود ندارد.

پاسخ :

الف) $precision = \frac{6}{20} = 0.3$

ب) $recall = \frac{6}{8} \rightarrow F_1 = \frac{2PR}{P+R} = \frac{3}{7} = 0.43$

ج) $MAP = \frac{1}{6} * \left(1 + 1 + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20}\right) = 0.555$

د) $MAP_{smallest} = \frac{1}{8}\left(1 + 1 + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{7}{9999} + \frac{8}{10000}\right) = 0.417$

سوال ۵

یک شرکت قصد دارد گزارش‌های اقتصادی در زمینه کامپیوتر را بازیابی کند. ۳ گزارش که تعداد تکرار کلمات آنها در جدول ۳ بیان شده است را در اختیار داریم. با استفاده از روش $nnc.nnc$ این گزارش‌ها را براساس میزان ارتباط شان مرتب نمایید.

	تهران	کامپیوتر	ایران	اقتصادی
گزارش ۱	۱۲	۰	۳	**۲۴**
گزارش ۲	۱۰	۵	۲۰	**۱۰**
گزارش ۳	۰	۱۲	۹	**۸**
پرسمان	۱	۱	۱	**۱**

جدول۳

پاسخ :

$$\cos(\text{Report A}, \text{Query}) = \frac{12+3+24}{\sqrt{4}\sqrt{12^2+3^2+24^2}} = \frac{39}{54} = 0.72$$

$$\cos(\text{Report B}, \text{Query}) = \frac{10+5+20+10}{\sqrt{4}\sqrt{10^2+5^2+20^2+10^2}} = \frac{45}{50} = 0.90$$

$$\cos(\text{Report C}, \text{Query}) = \frac{12+9+8}{\sqrt{4}\sqrt{12^2+9^2+8^2}} = \frac{29}{34} = 0.85$$

ترتیب بر اساس ارتباط: گزارش ۲، گزارش ۳، گزارش ۱.

سوال ۶

در پاسخ یک پرسمان تعدادی سند به ترتیب به عنوان پاسخ آمده است. از یک متخصص درخواست شده است تا بین نتایج قضاوت کند. این قضاوت به صورت اعداد صحیح ۰ تا ۳ بیان می شود که ۰ به معنای کاملا غیرمرتبط و ۳ به معنای کاملا مرتبط است و ۱و۲ نیز دو سطح میانی ارتباط را نشان می‌دهند.

نتیجه موتور جستجو و قضاوت متخصص به شرح زیر می باشد:

$$D_1, D_2, D_3, D_4, D_5, D_6$$

3,2,3,0,1,2

الف) باتوجه به این اطلاعات عملکرد موتورجستجو را بر اساس معیارهای DCG و CG ارزیابی کنید.

ب) حال فرض کنید جواب موتور جستجو به صورت زیر باشد. مجددا DCG و CG را محاسبه نمایید و نتیجه این قسمت را با نتیجه قسمت الف مقایسه نمایید.

$$D_1, D_2, D_4, D_3, D_5, D_6$$

برای محاسبات خود می توانید از جدول ۴ استفاده نمایید.

i	$log_2(i+1)$
۱	1
۲	1.585
۳	2
۴	2.322
۵	2.585
۶	2.807

جدول۴

پاسخ :

الف)

$$\mathrm{CG}_6 = \sum_{i=1}^{6} rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11$$

$$\mathrm{DCG}_6 = \sum_{i=1}^{6} \frac{rel_i}{\log_2(i+1)} = 3 + 1.262 + 1.5 + 0 + 0.387 + 0.712 = 6.861$$

ب)

$$DCG_6 = 3 + 1.262 + 0 + 1.29 + 0.387 + 0.712 = 6.840$$

مقدار CG بدون تغییر باقی می‌ماند درنتیجه اگر ترتیب اسناد بازیابی شده اهمیت دارد معیار DCG ارزیابی بهتری از موتورجستجو را ارائه می دهد.

سوال ۷

یک شرکت تبلیغاتی با پایگاه داده مشتریان ۲۰۰۰ نفری میخواهد از یک روش جدید برای ارسال پیام تبلیغاتی استفاده کند. روش قبلی آن ها ارسال پیام زیر بود:

"مهلت خرید محصول به زودی تمام میشود، برای خرید به لینک $A1$ مراجعه کنید"

با استفاده از روش $A/B\ testing$ راهکاری برای این شرکت ارائه دهید تا بتواند مقایسه‌ای میان پیام قبلی و پیام جدید انجام دهد و از این به بعد از پیام بهتر استفاده کند.

پاسخ :

این شرکت باید پیام جدید مثل "فقط یک روز دیگر باقی است، به لینک $B1$ مراجعه کنید" را برای چند درصد از کاربران خود مثلاً ۵٪ امتحان کند. سپس میزان مراجعه به لینک های $A1$ و $B1$ را به نسبت تعداد پیام های ارسال شده بسنجند و اگر نتایج مناسب بود میتوانند پیام دوم را جایگزین پیام اول کند.