



بازیابی پیشرفته اطلاعات

نیمسال اول ۱۴۰۰

مدرس: دکتر لشکری

شماره درس: ۴۰۳۲۴

میان ترم

زمان: ۹۰ دقیقه

لطفا نام و نام خانوادگی و شماره دانشجویی را بروی تمام برگه های جواب بنویسید.

پرسش اول ----- ۵ نمره

لطفا جملات نادرست را تصحیح کنید.

- الف) ما به راحتی می توانیم تعداد عبارات منحصر به فرد در یک سند خاص را از یک نمایه معکوس بدست آوریم.
- ب) هدف مدل های بازیابی که آموخته ایم بهبود برخی معیارهای ارزیابی IR خاص، مانند NDCG و MAP است.
- ج) Stemming به بهبود recall مدل بازیابی بولین کمک می کند.
- د) مدل فضای برداری معادل مدل Bag-of-Word است.
- ه) $p@2=0.2$ به این معنی است که شما به طور متوسط ۲۰٪ شانس پیدا کردن یک سند مرتبط در موقعیت ۲ را در تمام پرس و جوها خواهید داشت.
- و) شباهت کسینوس در مدل های فضای برداری ترجیح داده می شود زیرا با توجه به طول سند نرمال شده است.
- ز) همیشه بین recall و precision مصالحه (trade-off) وجود دارد.

پرسش دوم ----- ۵ نمره

تفاوت بین بازیابی دودویی (Boolean retrieval) و بازیابی رتبه بندی (Ranked retrieval) شده چیست؟ در بیشتر موتورهای جستجو چه روش بازیابی به کار برده می شود؟ چرا؟

پرسش سوم ----- ۲ نمره

در مجموعه معینی از اسناد (documents) فارسی، فراوانی رایج ترین کلمه ۱۲۷۰۸۷۳ است. بر این اساس فرکانس تخمینی دومین کلمه و سومین کلمه پرتکرار در این مجموعه چقدر است و چرا؟

پرسش چهارم ----- ۳ نمره

آیا idf می تواند در رتبه بندی برای پرس و جوی یک ترم (one-term queries) تأثیر بگذارد؟ لطفا با مثال توضیح دهید؟



بازیابی پیشرفته اطلاعات

نیمسال اول ۱۴۰۰

مدرس: دکتر لشکری

شماره درس: ۴۰۳۲۴

میان ترم

زمان: ۹۰ دقیقه

پرسش پنجم-----۵ نمره

چگونه می توانیم شناسه اسناد (doc-Id) که به روش gap encoding در رشته زیر به روش گاما posting فشرده سازی شده اند را بازیابی کرد؟ (لطفا تمام مراحل را توضیح دهید)

۱۱۱۰۰۰۱۱۱۰۱۰۱۰۱۱۱۱۱۰۱۱۰۱۱۱۱۰۱۱

پرسش ششم-----۸ نمره

الگوریتم پردازش پرس و جو به روش Document-At-A-Time و Term-At-A-Time توضیح دهید و آنها را با یکدیگر مقایسه کرد و مزایا و معایب آنها را بیان کنید .


پرسش هفتم-----۱۰ نمره

یک سامانه بازیابی در پاسخ به دو پرس و جو مستندات زیر را بازگردانده است. ترتیب مستندات بازگردانده شده از چپ به راست هست. مستندات مرتبط (relevant) با رنگ سیاه مشخص شده اند. مقدار $R@4$, $p@2$, MRR برای این سامانه محاسبه کنید. لطفا مراحل محاسبه با جزییات توضیح دهید.

 = relevant documents for query 1

Ranking #1



 = relevant documents for query 2

Ranking #2





بازیابی پیشرفته اطلاعات

نیمسال اول ۱۴۰۰

مدرس: دکتر لشکری

شماره درس: ۴۰۳۲۴

میان ترم

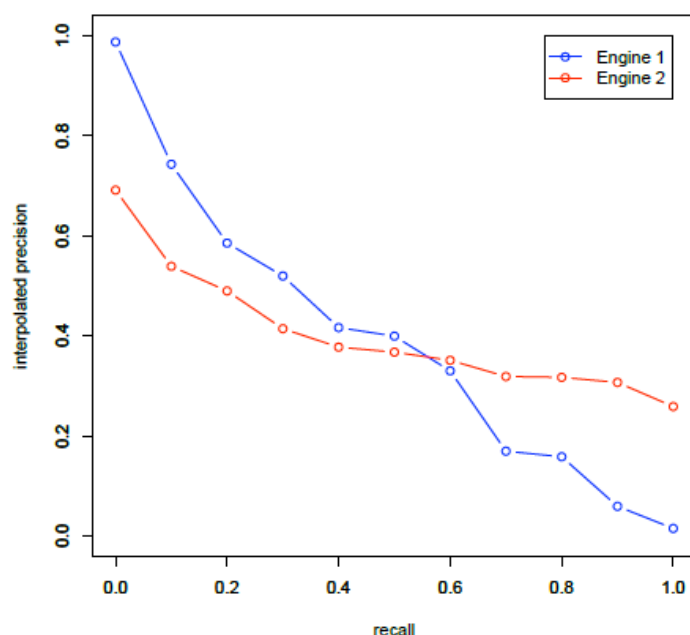
زمان: ۹۰ دقیقه

پرسش هشتم ----- ۷ نمره

- الف) دلایل استفاده از فشرده سازی دیکشنری را توضیح دهید؟
ب) بهترین روش فشرده سازی دیکشنری را به صورت مختصر توضیح دهید.
ج) برای روش فشرده سازی از نوع lossy یا lossless یک مثال بزنید.

پرسش نهم ----- ۵ نمره

شکل زیر منحنی‌های فراخوان شده precision and recall را برای دو موتور جستجو که مقالات تحقیقاتی را فهرست‌بندی می‌کنند، نشان می‌دهد. هیچ تفاوتی بین موتورها وجود ندارد مگر در نحوه امتیازدهی آنها. تصور کنید دانشمندی هستید که به دنبال همه کارهای منتشر شده در مورد یک موضوع است. شما نمی‌خواهید هیچ نقل قولی را از دست بدهید. کدام موتور را ترجیح می‌دهید و چرا؟





بازیابی پیشرفته اطلاعات

نیمسال اول ۱۴۰۰

مدرس: دکتر لشکری

شماره درس: ۴۰۳۲۴

میان ترم

زمان: ۹۰ دقیقه

پرسش تشویقی-----۱۰ نمره

آزمایشگاه دی جی کالا در حال انجام یک پروژه تحقیقاتی مهم برای بهبود اثربخشی جستجوی محصول digikala.com است. چندین تیم ادعا می کنند که الگوریتم های آنها بهترین هستند و باید به کار گرفته شوند. به عنوان مدیر این پروژه، معیار قضاوت شما چه خواهد بود؟ بر این اساس چگونه باید یک تصمیم منطقی بگیرید؟

موفق باشید.