به نام خدا

نمونه میان‌ترم درس بازیابی هوشمند اطلاعات

---

۱. الف) فرض کنید یک سیستم بازیابی در پاسخ به یک پرس‌وجو لیست مرتب شده‌ای از ۱۰ سند را برمی‌گرداند که وضعیت مرتبط بودن آنها در جدول مقابل نشان داده شده است. "+" در ستون relevance به معنی مرتبط بودن و "-" به معنی نامرتبط بودن سند به پرس‌وجو است. فرض کنید در کل 5 سند مرتبط به پرس‌وجو در مجموعه وجود دارد. Precision، recall و Mean Average Precision (MAP) این سیستم برای این پرس‌وجو را محاسبه کنید.

رتبه	relevance
1	+
2	-
3	-
4	+
5	+
6	-
7	-
8	-
9	-
10	-

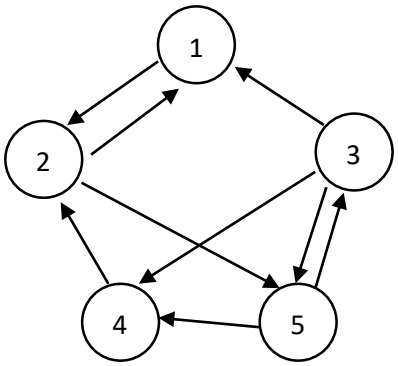
ب) به طور خلاصه توضیح دهید چرا MAP معیار بهتری از precision در k سند (prec@k) برای مقایسه دو سیستم از نظر دقت رتبه بندی است؟

۲. الگوریتم PageRank برای رتبه‌دهی صفحات را می توان به این صورت توصیف کرد:

$$\vec{R} = (1-\alpha)M\vec{R} + \alpha\vec{P}$$

که R مقادیر اهمیت صفحات است و M بر اساس گراف وب ساخته می شود.

الف) برای گراف نمونه شکل، ماتریس M و بردارP را بسازید.



ب) می خواهیم PageRankهای حساس به موضوع (topic-sensitive) را برای موضوع sport بسازیم. فرض کنید صفحات ۱ و ۲ در رابطه با sport هستند. کدام مولفه در محاسبه PageRank بخش (الف) تغییر می کند؟ قسمت تغییر داده شده را بنویسید.

۳. الف) مزایای تحلیل محدودیت‌ها در چارچوب Axiomatic کدامند؟

ب) با استفاده از چارچوب Axiomatic می‌خواهیم یک تابع بازیابی جدید با توجه به روش پایه Dirichlet Prior ایجاد کنیم. روش Dirichlet Prior مرتب‌سازی اسناد را به این صورت انجام می‌دهد:

$$s(Q,D) = \left[\sum_{w\in Q, w\in D} c(w,Q)\ln\left(1+\frac{c(w,D)}{\mu p(w|C)}\right)\right] + |Q|\ln\frac{\mu}{\mu+|D|}$$

برای این تابع بازیابی دو تابع primitive weighting function و query growth function را بدست آورید.

۴. مدل بازیابی احتمالی سندها را به این ترتیب مرتب می‌کند:

$$\log O(R=1|Q,D) \overset{Rank}{\approx} \sum_{i=1,d_i=q_i=1}^{k} \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

که در آن pi = P(Ai = 1|Q, R = 1) احتمال این است که ترم Ai در یک سند مرتبط ظاهر شده باشد و qi = P(Ai = 1|Q, R = 0) احتمالی است که ترم Ai در یک سند نامرتبط ظاهر شده باشد.

(الف) یکی از چالش‌های اصلی در مدل بازیابی احتمالی تخمین پارامترهاست(pi, qi). توضیح دهید که در حالتی که Relevance Judgment نداریم چگونه این پارامترها تخمین زده می‌شوند. فرمول محاسبه log O(R=1|Q,D) را بدست آورید.

(ب) فرمولی که در بخش (الف) بدست آورده‌اید سندها را بر چه اساسی مرتب می‌کند؟ کدام کلمات بیشترین تاثیر را در مرتب سازی خواهند داشت؟ به طور دقیق و با فرمول بحث کنید.

۵. کد گاما و کد دلتای معادل عدد ۳۰ بنویسید.

۶. الف) درخواست Q و سند D را در نظر بگیرید. فرمول امتیازدهی D بر اساس Q با استفاده از روش بازیابی query likelihood را بدست آورید. فرض کنید:

- برای هموارسازی از روش هموارسازی Dirichlet Prior با پارامتر μ استفاده می‌کنیم و
- مدل زبانی مجموعه (collection language model)، p(w|C) است.

ب) توضیح دهید که فرمولی که در بخش (الف) بدست آوردید چگونه ابتکاراتی مانند وزن‌دهی TF-IDF و هموارسازی طول سند را پیاده‌سازی می‌کند.