Quiz — 6 — The Last Chapter..
MIR — Spring 2026
Instructor: Dr. Amin Eskandari
Head Teaching Assistants: Hamid Namjoo, Amir Hossein Hemati
Teaching Assistants: Ali Mojahed, Ali Nikvan, Amir Javvi

سوال 1

در تیم مهندسی جستجوی گوگل، با حجم عظیمی از صفحات خبری مواجه هستیم که محتوای یکسانی دارند اما درURL ها و قالب‌های متفاوتی منتشر شده‌اند (مثلاً نسخه موبایل و دسکتاپ یا میرورهای سایت‌های خبری). اگر از شما بخواهند الگوریتمی طراحی کنید که این صفحات تقریباً تکراری (Near-Duplicate) را با کمترین پیچیدگی محاسباتی شناسایی کنید، چگونه از مفاهیم Shingling و MinHashاستفاده می‌کنید؟ چرا روش سنتی مقایسه کاراکتر به کاراکتر در مقیاس وب شکست می‌خورد؟

سوال 2

فرض کنید در تیم فیلتر اسپم (Hotmail یا Gmail) کار می‌کنید. یک ایمیل قانونی از بانک (مثل پیام تأیید تراکنش) به دلیل وجود کلماتی شبیه به فیشینگ، توسط یک طبقه‌بند ساده اشتباهاً به عنوان اسپم شناسایی می‌شود .(False Positive) چگونه با استفاده از مفهوم حاشیه نرم (Soft Margin) و «متغیرهای لغزش» در SVM، این مشکل را حل می‌کنید؟ پارامتر C در اینجا دقیقاً چه نقشی در تعادل بین False Positive و False Negative دارد؟

سوال 3

وقتی کاربری عبارت "iPhone 17" را در آمازون جستجو می‌کند، یک کاربر قصد خرید دارد و کاربر دیگر ممکن است فقط دنبال بررسی مشخصات فنی باشد. سیستم تبلیغات آمازون (CPC) چگونه باید با استفاده از دسته‌بندی پرس‌وجوها (Informational, Navigational, Transactional) تصمیم بگیرد که چه نوع تبلیغاتی را نمایش دهد؟ اگر نوع پرس‌وجو را اشتباه تشخیص دهیم، چه پیامدهای اقتصادی برای سیستم دارایی آمازون دارد؟

سوال 4

موتور جستجوی Bing در حال کرال کردن وب است و متوجه می‌شود که بخش قابل توجهی از کرالرهایش در بخش‌هایی از وب گیر کرده‌اند و صفحاتی را ایندکس می‌کنند که هرگز در نتایج جستجو ظاهر نمی‌شوند. با تحلیل ساختار پاپیونی (Bowtie Structure) گراف وب، توضیح دهید این کرالرها کدام بخش از وب (SCC, IN, OUT, Tendrils) در حال گشتن هستند و چرا؟ استراتژی کرالینگ را چگونه بر اساس این ساختار تغییر می‌دهید؟

سوال 5

یوتیوب می‌خواهد کامنت‌های حاوی «سخنان نفرت‌پراکنی» را از «گفتگوهای عادی» جدا کند. مرز بین طعنه زدن و توهین بسیار باریک است و داده‌ها در فضای ویژگی‌های متنی به هیچ وجه خطی جداپذیر نیستند! چگونه با استفاده از SVM غیرخطی و Kernel Trick این مشکل را حل می‌کنید؟ از نظر ریاضی، کرنل چه تغییری در فضای ویژگی ایجاد می‌کند که یک هایپرپلن خطی نتواند آن را انجام دهد؟

سوال 6

شما مدیر ارشد داده در موتور جستجوی DuckDuckGo هستید و می‌خواهید بدانید اندازه ایندکس شما نسبت به گوگل چقدر است (نسبت E2/E1). با توجه به اینکه هیچ موتور جستجویی اجازه دسترسی مستقیم به ایندکسش را نمی‌دهد، چگونه از روش «Capture-Recapture» برای تخمین این نسبت استفاده می‌کنید؟ فرضیات ساده‌کننده این روش چیست و چرا در دنیای واقعی (وب) این فرضیات نقض می‌شوند؟

سوال 7

یک فروشگاه اینترنتی بزرگ مانند دیجی‌کالا میلیون‌ها صفحه پویا (Dynamic) برای فیلتر کردن محصولات دارد (مثلاً گوشی+آبی+بین ۱۰ تا ۲۰ میلیون). موتورهای جستجو این صفحات را به عنوان صفحات بی‌پایان (Spider Traps) می‌شناسند و از ایندکس آن‌ها امتناع می‌کنند (Soft 404). به عنوان مهندس سئو (SEM)، چگونه با استفاده از مفاهیم گراف وب و تفاوت صفحات ایستا و پویا، راهکاری ارائه می‌دهید تا موتور جستجو فقط صفحات ارزشمند را ایندکس کند و در دام کرالر نیفتد؟

سوال 8 - سوال جایزه

در ادبیات پژوهشی Web Search و به خصوص در بحث اسپم و Adversarial IR، یک مسئله باز و بنیادین وجود دارد: "" آیا می‌توان یک مرز تصمیم‌گیری کاملاً مقاوم در برابر دستکاری‌های خصمانه (Adversarial Manipulations) در وب یافت، یا این جنگ بین اسپمرها و موتورهای جستجو یک بازی بدون نقطه تعادل (No Nash Equilibrium) است؟ "" با استناد به مفاهیم کلاس SVM (حداکثر حاشیه) و واقعیت‌های اقتصاد وب (مدل CPC و انگیزه‌های مالی)، استدلال کنید که چرا یافتن یک طبقه‌بند بی‌نقص برای همیشه غیرممکن به نظر می‌رسد.(یعنی واقعا غیرممکنِ؟! چرا؟)

موفق و موید :)