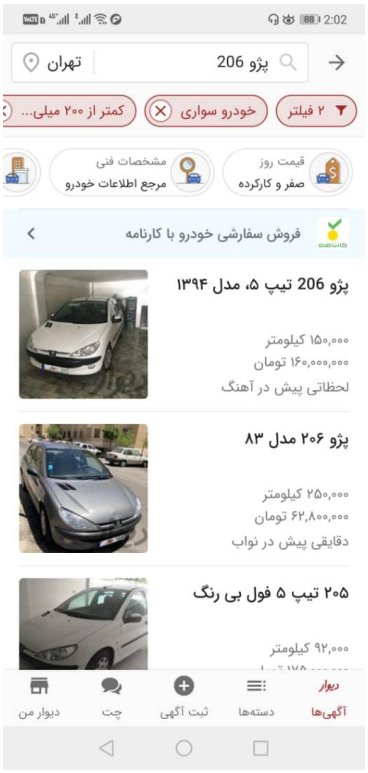
تسک ورودی سامرکمپ تحلیل داده



مقدمه

در فرایند جستجوی آگهی در دیوار همیشه گام اول این است که کاربر یک کوئری (query) برای دیوار می‌فرستد و ما بر اساس آن کوئری لیستی از آگهی‌ها آماده می‌کنیم و به کاربر نمایش می‌دهیم. یک کوئری می‌تواند ترکیبی از جستجوی متنی و چند فیلتر باشد. برای مثال در شکل مقابل نتایج یک کوئری شامل جستجوی متنی عبارت «پژو ۲۰۶» و فیلترهای دسته «خودرو سواری» و قیمت کمتر از ۲۰۰ میلیون تومان در اپلیکیشن اندروید دیوار نمایش داده شده است.

رتبه‌بندی نتایج یک کوئری بر اساس زمان انتشار آگهی‌ها انجام می‌شود. یعنی همواره آخرین آگهی منتشر شده به عنوان نتیجه اول نمایش داده می‌شود و پس از آن بقیه آگهی‌ها به ترتیب زمان انتشارشان در رتبه‌های بعدی قرار می‌گیرند. در نتیجه در پاسخ به یک کوئری فقط باید مشخص کنیم چه آگهی‌هایی قرار است نمایش داده شوند. این یکی از تفاوت‌های سیستم جستجوی آگهی دیوار با موتورهای جستجو مثل گوگل است که در آنها ترتیب نمایش نتایج به روش دیگری تعیین می‌شود.

برای نظارت بر عملکرد سیستم جستجوی آگهی متریک‌های زیادی وجود دارند. در این تسک می‌خواهیم چند تا از این متریک‌ها را بررسی و محاسبه کنیم.

داده

برای این تسک داده رفتار نمونه‌ای از کاربران اپلیکیشن اندروید دیوار در یک روز خاص در اختیار شما قرار می‌گیرد (لینک دانلود). هر سطر این جدول نشان دهنده یک اکشن (action) انجام شده توسط یک کاربر است. ستون‌های این جدول به شرح زیر است:

- action: اکشن انجام شده توسط کاربر، حاوی یکی از مقادیر load_post_page یا click_post
- created_at: زمان وقوع اکشن با فرمت timestamp و با دقت میلی ثانیه
- source_event_id: شناسه یکتای هر کوئری
- device_id: شناسه یکتای هر کاربر
- post_page_offset: شماره لیست لود شده (مربوط به اکشن load_post_page)
- tokens: لیست شناسه آگهی‌های لود شده (مربوط به اکشن load_post_page)

- post_index_in_post_list: رتبه آگهی در لیست (مربوط به اکشن click_post)
- post_token: شناسه یکتای آگهی کلیک شده (مربوط به اکشن click_post)

برای درک بهتر این جدول لازم است اکشن‌های load_post_page و click_post را بیشتر بشناسیم.

همانطور که قبلا بیان شد بعد از ارسال کوئری توسط کاربر، لیستی از آگهی‌ها برای او نمایش داده می‌شود. سیستم جستجوی دیوار آگهی‌ها را به صورت دسته‌های ۲۴تایی برای کاربر لود می‌کند و اگر کاربر به اسکرول کردن در لیست ادامه دهد دسته‌های بعدی ۲۴تایی از آگهی‌ها را مشاهده می‌کند. هر بار لود شدن یک دسته ۲۴تایی از آگهی‌ها به عنوان یک اکشن load_post_page ثبت می‌شود. ستون post_page_offset نشان دهنده این است که این آگهی‌ها چندمین دسته‌ای هستند که برای کوئری فعلی کاربر لود می‌شوند. ستون tokens نیز شامل شناسه آگهی‌هایی است که در این دسته بوده‌اند.

بعد از لود شدن لیست آگهی‌ها کاربر ممکن است روی بعضی از آنها کلیک کند. هر کلیک کاربر روی یک آگهی به عنوان یک اکشن click_post ثبت می‌شود. ستون post_index_in_post_list نشان دهنده رتبه آگهی در لیست هنگام کلیک گرفتن است. ستون post_token نیز حاوی شناسه آگهی است که روی آن کلیک انجام شده است.

در نهایت تمام اکشن‌هایی که مرتبط با یک کوئری واحد بوده‌اند، توسط ستون source_event_id قابل تشخیص هستند.

مسائل‌ها

سوال ۱) ابتدا کمی در داده گشت و گذار کنید و سعی کنید با جزئیات آن آشنا شوید. همزمان با اپ دیوار هم کار کنید تا با داده ارتباط بیشتری برقرار کنید. در این گشت و گذارها ممکن است به برخی خطاهای جزئی در داده پی ببرید. اگر به چنین مواردی برخوردید به طور خلاصه به چند تا از آنها اشاره کنید. (حداکثر ۵ مورد به صورت بولت‌وار. در صورت نیاز می‌توانید برای هر مورد یک جدول یا نمودار نیز بیاورید)

سوال ۲) دو مورد از متریک‌های مهم سیستم جستجو موارد زیر هستند:

- dark query percent: درصد کوئری‌هایی که برای آنها کمتر از ۱۰ نتیجه نمایش داده‌ایم.
- query bounce rate: درصد کوئری‌هایی که کاربر روی هیچ کدام از نتایج کلیک نکرده.

این دو متریک را با داده‌ای که در اختیار دارید محاسبه کنید.

سوال ۳) برای اینکه بدانیم آگهی‌هایی که به کاربران نمایش می‌دهیم چقدر مطابق خواسته آنهاست، برای هر کوئری متریک‌های زیر تعریف می‌شوند:

- درصد آگهی‌های کلیک شده نسبت به آگهی‌های لود شده
- رتبه اولین کلیک کاربر (برای کوئری‌هایی که حداقل یک کلیک داشته‌اند)

- میانگین فاصله بین رتبه کلیک‌های یک کاربر (مثلا اگر کاربر از بین نتایج کوئری روی آگهی‌های دهم و شانزدهم کلیک کند، این متریک برابر ۲/(۶+۱۰) یعنی ۸ خواهد بود)
- اینکه آیا روی یکی از ۳ نتیجه اول کوئری کلیک شده یا نه

از بین این متریک‌ها یکی را که فکر می‌کنید متریک بهتری است انتخاب کرده و میانگین آن را برای همه کوئری‌ها محاسبه کنید. برای انتخاب خود به طور خلاصه در یک یا دو پاراگراف توضیح دهید.

سوال ۴) چهار متریکی که در سوال قبل بیان شد رابطه بسیار نزدیکی با هم دارند. برای درک بهتر این رابطه مروری روی توزیع برنولی انجام دهید. از این توزیع برای مدلسازی آزمایش‌های دو حالته (مثل پرتاب سکه) استفاده می‌شود. آیا می‌توان با استفاده از این توزیع یک مدل ساده طراحی کرد که رفتار کلیک کاربران روی نتایج کوئری‌ها را مدلسازی کند؟ با استفاده از این مدل اگر مقدار یکی از چهار متریک بیان شده را بدانید آیا می‌توانید تخمینی از سه متریک دیگر محاسبه کنید؟ توضیحات خود را در حداکثر یک صفحه بنویسید.

برای تحویل این تسک تا پایان خرداد فرصت دارید. پاسخ شما می‌تواند در قالب یک فایل متنی به زبان فارسی (مثلا پی‌دی‌اف یا پاورپوینت) به همراه کد پایتون (مثلا نوتبوک ژوپیتر) باشد. همچنین می‌توانید توضیحات متنی را داخل نوتبوک قرار داده و فقط آن را ارسال کنید. در ارزیابی کد موارد زیر بررسی می‌شوند:

- کد خلاصه، تمیز و خوانا باشد.
- در صورت امکان به جای تعریف توابع جدید و یا استفاده از حلقه‌های for روی سطرهای دیتافریم، از توابع و متدهای نامپای و پانداز استفاده شود.