

## تمرین پیش پردازش و تبدیل داده

دوره کارآموزی علوم داده

**Becomexpert**

تابستان 1403

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 آگهی ها 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. رتبه بندی نتایج یک کوئری بر اساس زمان انتشار آگهی ها 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 رتبه های بعدی قرار میگیرند. در نتیجه در پاسخ به یک کوئری فقط باید مشخص کنیم چه آگهی هایی قرار است نمایش داده شوند. این یکی از تفاوت های 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 متریک ها را بررسی و محاسبه کنیم.

برای این تسک داده رفتار نمونه ای از کاربران اپلیکیشن اندروید دیوار در یک روز خاص در اختیار شما قرار میگیرد. 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 آگهی های 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 `page_post_load` و `post_click` 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 آگهی ها را به صورت دسته های ۲۴ تایی برای کاربر لود میکند و اگر کاربر به اسکرول کردن در لیست ادامه دهد دسته های بعدی ۲۴ تایی از آگهیها را مشاهده میکند. هر بار لود شدن یک دسته ۲۴ تایی از آگهیها به عنوان یک اکشن `page_post_load` ثبت میشود. ستون `offset_page_post` نشان دهنده این است که این آگهیها چندمین دسته‌های هستند که برای کوئری فعلی کاربر لود میشوند. ستون `tokens` نیز شامل شناسه آگهی هایی است که در این دسته بوده‌اند. بعد از لود شدن لیست آگهی ها کاربر ممکن است روی بعضی از آنها کلیک کند. هر کلیک کاربر روی یک آگهی به عنوان یک اکشن `post_click` ثبت میشود. ستون `list_post_in_index_post` نشان دهنده رتبه آگهی در لیست هنگام کلیک گرفتن است. ستون `token_post` نیز حاوی شناسه آگهی است که روی آن کلیک انجام شده است. در نهایت تمام اکشن هایی که مرتبط با یک کوئری واحد بوده‌اند، توسط ستون `id_event_source` قابل تشخیص هستند.

## مساله ها

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 بولت وار. 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:

- `percent query dark`: درصد کوئری هایی که برای آنها کمتر از ۱۰ نتیجه نمایش داده ایم.

- `rate bounce query`: درصد کوئری هایی 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 آگهی هایی که به کاربران نمایش میدهم چقدر مطابق خواسته آنهاست، برای هر کوئری متریکهای زیر تعریف میشوند:

- درصد آگهی های کلیک شده نسبت به آگهی های لود شده

- رتبه اولین کلیک کاربر (برای کوئری هایی که حداقل یک کلیک داشته اند)

- میانگین فاصله بین رتبه کلیک های یک کاربر (مثلا اگر کاربر از بین نتایج کوئری روی آگهی های دهم و شانزدهم کلیک کند، این متریک برابر  $\frac{10+6}{2}$  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 کوئری ها 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 حالتی (مثل پرتاب سکه) استفاده میشود. آیا میتوان با استفاده از این توزیع یک مدل ساده طراحی کرد که رفتار کلیک کاربران روی نتایج کوئری ها 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.

در صورت امکان به جای تعریف توابع جدید و یا استفاده از حلقه های for روی سطرهای دیتافریم، از توابع و متدهای نامپای و پانداز استفاده شود. تمامی نتایج و تفسیر خودتون رو تو یک گزارش بیارید.