پروژه تحلیل داده‌های جستجو ترب

مدرسه تابستانی تحلیل داده OpenAI[1]

مرداد ۱۴۰۱

سلام و خوش آمد به شما از طرف تیم دیتا ترب!

خیلی خوشحال هستیم که در این پروژه قصد داریم داده‌های جستجو (search) ترب را با شما به اشتراک بگذاریم و به کمک شما تحلیل کنیم. این شکل از تحلیل را معمولا با نام EDA یا Exploratory Data Analysis می‌شناسند. داده‌هایی که در اختیار شما قرار گرفته است مربوط به بخش کوچکی از داده‌های واقعی جستجو کاربران در اپلیکیشن ترب در بازه زمانی یک هفته است. به همراه داده‌ها، یک فایل Jupyter Notebook نیز به شما داده شده است که در آن صورت تحلیل‌های خواسته شده قرار گرفته است و شما باید با مطالعه آن به سوالات مورد نظر پاسخ دهید. هدف از انجام این پروژه تمرین مهارت های تحلیل داده، به خصوص مهارت استفاده از کتابخانه Pandas در انجام تحلیل‌ها، و همچنین استخراج اطلاعات جالب و مفید از داده‌های جستجو ترب می باشد. در ادامه، ابتدا معرفی کوتاهی از ترب خواهیم داشت و سپس اطلاعات موجود در داده‌های جستجو را توضیح خواهیم داد. در پایان توضیح مختصری برای هر کدام از تحلیل های خواسته شده ارائه خواهد شد.

> معرفی ترب

ترب یک «موتور جستجو محصولات» است. با استفاده از اپلیکیشن ترب می‌توانید فروشندگان مختلف یک کالا را پیدا کنید، قیمت و شرایط فروش آن ها را با یکدیگر مقایسه کنید و خرید راحت تر و به صرفه تری داشته باشید. در حال حاضر، بیش از ۱۲ میلیون محصول و بیش از ۱۴ هزار فروشگاه اینترنتی در کاتالوگ محصولات ترب موجود است. روزانه صدها هزار جستجو توسط کاربران برای پیدا کردن محصولات مورد نظرشان در ترب انجام میشود. پیشنهاد می کنیم

[1] لینک نسخه آنلاین این داکیومنت (نسخه به روز):
https://docs.google.com/document/d/1r46gWcSXwvxxpS92cOqqA7pLuX1Vq673rokgwtnafb8

برای درک بهتر نحوه کارکرد ترب حتما به سایت ترب سر بزنید و یا اپلیکیشن ترب را روی گوشی خود نصب و استفاده کنید (کافه بازار، مایکت، گوگل پلی).

> داده های پروژه

ابتدا لازم است که دیتا مربوط به پروژه را از منتور خود دریافت کنید و یا با مراجعه به این آدرس آن را در قالب یک فایل zip دانلود کنید.

در این پروژه چهار نوع داده مختلف در قالب چهار فایل مجزا (که در پوشه data قرار دارند) به شما داده شده است. همگی این داده‌ها به صورت یک Pandas DataFrame هستند که به فرمت یک فایل pickle ذخیره شده‌اند. در زیر اطلاعات مربوط به هر فایل و ستون‌های موجود در هر کدام از آنها توضیح داده شده است:

- داده‌های جستجو کاربران (فایل search_logs.pkl): اطلاعات مربوط به جستجوهای انجام شده توسط کاربران در این فایل ذخیره شده است.

نام ستون (column name)	توضیحات (description)
id	هر جستجو توسط یک کاربر یک شناسه یکتا حروفی دارد.
raw_query	عبارت جستجویی که کاربر وارد کرده است (به صورت خام و بدون انجام هیچ گونه پردازش روی آن).
result	نتایج جستجو است و شامل حداکثر ده نتیجه اولی که به کاربر در این جستجو نمایش داده شده، می باشد. هر مقدار این ستون یک لیست پایتون می باشد که شناسه (id) محصولاتی که در جستجو به کاربر نمایش داده شده به ترتیب نمایش آن‌ها، در این لیست ذخیره شده است.
datetime	تاریخ و ساعت انجام جستجو (در تایم زون UTC).
category_id	کاربر می‌تواند جستجو خود را به یک دسته‌بندی مشخص از محصولات محدود کند. برای مثال می‌تواند جستجو عبارت «گوشی» را تنها محدود به محصولات موجود در دسته «گوشی موبایل» بکند. در این حالت شناسه (id) مربوط به آن دسته‌بندی در این ستون قرار می گیرد. در صورتی که دسته‌بندی توسط کاربر انتخاب نشود، مقدار این ستون NaN خواهد بود.
user_id	شناسه یکتا برای کاربری که جستجو را انجام داده است.

- **داده‌های کلیک روی نتایج جستجو (فایل search_click_logs.pkl):** این داده شامل اطلاعات مربوط به کلیک هایی است که کاربر روی نتایج جستجو انجام میدهد.

نام ستون (column name)	توضیحات (description)
id	شناسه یکتا برای کلیک.
datetime	تاریخ و ساعت انجام کلیک (در تایم زون UTC).
search_id	شناسه جستجو مربوطه که روی نتایج آن کلیک شده است (متناظر با ستون id در داده های جستجو).
rank	برابر با رتبه محصول کلیک شده است. این رتبه متناظر با اندیس محصولات در ستون result در داده‌های جستجو می باشد (برای مثال رتبه صفر بدین معنی است که محصول موجود در اندیس صفر از لیست result جستجو متناظر، کلیک شده است).

- **داده‌های مشخصات محصولات (فایل products.pkl):** این داده شامل اطلاعات محصولات می‌باشد. دقت کنید که به دلیل محدودیت حجمی، فقط اطلاعات محصولات کلیک شده در این فایل موجود می باشد و شامل همگی محصولات نمی شود.

نام ستون (column name)	توضیحات (description)
id	شناسه یکتا برای محصول.
title	عنوان محصول.
category_id	شناسه دسته‌بندی محصول (متناظر با ستون id در داده‌های دسته‌بندی محصولات).

- **داده‌های مشخصات دسته‌بندی محصولات (فایل categories.pkl):** این داده شامل اطلاعات دسته‌بندی محصولات می باشد.

نام ستون (column name)	توضیحات (description)
id	شناسه یکتا برای دسته‌بندی.

title	عنوان دسته‌بندی.
parent_category_id	اگر این دسته‌بندی زیردسته یک دسته‌بندی دیگر باشد، شناسه آن دسته در این ستون قرار خواهد گرفت (برای مثال دسته «لوازم آشپزخانه» زیردسته «لوازم خانگی» است).

> تسک ها و تحلیل های خواسته شده

در زیر توضیحات مختصری مربوط به هر تسک و تحلیل خواسته شده آورده شده است. در نوتبوک داده شده مواردی که نیاز به پاسخگویی دارند با علامت پرچم مثلثی (🚩) نشان داده شده‌اند. در نوتبوک داده شده توضیحات هر تسک یا تحلیل خواسته شده کامل‌تر است و بنابراین توصیه می‌شود علاوه بر خواندن موارد زیر، توضیحات نوتبوک را برای هر تسک/تحلیل بخوانید. به علاوه، در نوتبوک بعضا نکاتی از جهت راهنمایی به شما ارائه شده است (در قسمت Hint).

- **خواندن و لود داده‌ها به صورت Pandas DataFrame:** در این تسک شما ابتدا باید فایل های داده ای را با استفاده از Pandas لود کنید و هر کدام را در یک متغیر جداگانه ذخیره کنید.
- **بررسی سریع داده های لود شده:** بعد از لود کردن داده‌ها خوب است که چند سطر اول از دیتافریم‌ها را بررسی کنیم تا هم مطمئن بشویم لود داده‌ها به درستی انجام شده است و هم با ساختار داده ها آشنایی مختصری پیدا کنیم.
- **اعتبارسنجی شناسه‌های یکتا:** قبل از شروع تحلیل می‌خواهیم مطمئن شویم که داده‌ها با توصیف داده شده از آن ها و انتظارات ما تطابق دارند. یکی از این موارد بررسی یکتا بودن شناسه‌ها (ستون id) در چهار دیتافریم داده شده می باشد.
- **اعتبارسنجی رتبه کلیک‌های جستجو:** یکی دیگر از موارد اعتبارسنجی می تواند بررسی مقادیر رتبه کلیک‌های جستجو (ستون rank) باشد. می خواهیم مطمئن شویم که مقدار NaN در آن ها وجود ندارد (زیرا همه کلیک‌ها باید رتبه داشته باشند) و همچنین مقادیر آن ها بین صفر تا ۹ است (زیرا برای هر جستجو حداکثر ده محصول اول از نتایج جستجو داده شده است).
- **به دست آوردن تعداد جستجوها و کلیک های جستجو:** یکی از ساده ترین آمارهایی که از داده ها می توانیم به دست آوریم، تعداد جستجوها و کلیک‌های جستجو است.
- **تعداد عبارت‌های جستجو یکتا:** در این تحلیل می خواهیم بدانیم چه تعداد عبارت جستجو یکتا وجود دارد (به کلمه «یکتا» یا همان «unique» دقت کنید).
- **تعداد کاربران یکتا:** تعداد کاربران یکتایی که جستجو انجام داده اند، چه اندازه است؟

- **پرتکرارترین عبارت‌های جستجو:** هدف محاسبه محبوب‌ترین (پرتکرارترین) عبارت‌های جستجو به همراه تعداد دفعاتی که جستجو شده اند می باشد.
- **محاسبه توزیع طول عبارت های جستجو:** هدف محاسبه توزیع تعداد کلمات عبارت‌های جستجو است. می خواهیم بدانیم چه تعداد عبارت جستجو با طول های مختلف وجود دارند. به صورت شهودی انتظار داریم اغلب عبارت‌های جستجو حداکثر ۳ کلمه داشته باشند. در اینجا می توانید فرض کنید کلمات یک عبارت جستجو با یک یا تعدادی کاراکتر white space از هم جدا شده اند.
- **پرتکرارترین کلمات استفاده در عبارت‌های جستجو:** مشابه با تحلیل «پرتکرارترین عبارت‌های جستجو»، اما اینبار می خواهیم بدانیم کدام کلمات (و نه عبارت) از همه بیشتر در عبارت های جستجو تکرار می شوند.
- **محلی سازی تاریخ و ساعت داده‌های جستجو و کلیک جستجو:** همانطور که بالاتر در توضیحات داده‌ها اشاره شد، تاریخ و ساعت داده‌ها در تایم زون UTC ذخیره شده است. ولی برای سادگی درک داده ها و تحلیل ها بهتر است از تایم زون ایران (تهران) در تحلیل های خود استفاده کنیم. بنابراین ابتدا می خواهیم مقادیر ستون‌های مربوطه (ستون datetime) در هر دو دیتافریم را به تایم زون تهران (Asia/Tehran) تبدیل کنیم.
- **تعداد جستجوها و کلیک‌های جستجو در هر روز هفته:** بعد از تبدیل تایم زون، می‌خواهیم تعداد جستجوها و کلیک‌های جستجو را در هر روز هفته (شنبه، یکشنبه، دوشنبه و ...) حساب کنیم و به صورت یک bar plot در کنار هم نمایش بدهیم.
- **تعداد جستجوها و کلیک‌های جستجو بر حسب ساعت روز:** مشابه با مورد قبلی، اما اینبار می خواهیم بدانیم ترافیک جستجو و کلیک جستجو در هر ساعت از شبانه‌روز به چه شکل است و آن ها در یک bar plot در کنار هم نمایش دهیم.
- **پرتکرارترین عبارت‌های جستجو بر حسب روز هفته:** می خواهیم بدانیم در هر روز هفته چه عبارت‌های جستجویی بیشتر از همه تکرار شده اند.
- **تعداد جستجوهای کاربران به صورت گروه بندی شده:** هدف پیدا کردن تعداد کاربرانی است که تعداد جستجوهای آن ها مقدار مشخصی دارد (حداکثر ۳ جستجو، بین ۴ تا ۹ جستجو، و حداقل ۱۰ جستجو).
- **تعداد جستجوهای با انتخاب دسته‌بندی و بدون دسته‌بندی:** می خواهیم محاسبه کنیم تعداد جستجوهایی که کاربر آن را محدود به یک دسته‌بندی کرده است در مقایسه با تعداد جستجوهای بدون انتخاب دسته بندی، چگونه است.

- **پرتکرارترین عبارت‌های جستجو در بین جستجوهای محدود شده به دسته‌بندی:** برای جستجوهایی که کاربر آن را محدود به یک دسته‌بندی از محصولات کرده است، پرتکرارترین عبارت‌های جستجو چه هستند.
- **توزیع تعداد کلیک‌های جستجو بر اساس رتبه محصول کلیک شده:** هدف محاسبه توزیع تعداد کلیک‌های جستجو بر اساس رتبه محصول کلیک شده و رسم نمودار آن به صورت bar plot می‌باشد.
- **توزیع تعداد جستجوها بر حسب تعداد نتایج جستجو:** می خواهیم بدانیم چه تعداد از جستجوها کمتر از ۱۰ نتیجه دارند. در اینجا توزیع کامل تعداد جستجوها بر حسب تعداد نتایج به تفکیک (صفر نتیجه، یک نتیجه، دو نتیجه و ...) مطلوب است.
- **دسته‌بندی هایی که بیشترین جستجو در آنها انجام شده است:** هدف پیدا کردن دسته‌بندی های محبوب در جستجو در بین کاربران است. به عبارت دیگر می‌خواهیم بدانیم کدام دسته ها بیشتر از همه در جستجوها انتخاب شده اند.
- **توزیع تعداد کلیک ها در هر جستجو:** می خواهیم تعداد جستجوهایی که بدون کلیک هستند یا بالعکس کلیک زیادی دارند را محاسبه کنیم. بنابراین بهتر است توزیع تعداد کلیک‌ها در هر جستجو را محاسبه کنیم (یعنی چند جستجو بدون کلیک، با یک کلیک، با دو کلیک و ... داریم؟).
- **محصولات با بیشترین تعداد کلیک:** به عبارتی هدف پیداکردن محبوب‌ترین محصولات است. محبوبیت یک محصول را به طور ساده می توانیم بر اساس تعداد کلیک‌های روی آن محصول در نتایج جستجو به دست آوریم.
- **دسته‌بندی‌های با بیشترین تعداد کلیک:** هدف پیدا کردن محبوب‌ترین دسته‌بندی‌ها است. به عبارتی، به دنبال دسته‌بندی‌هایی هستیم که جمع تعداد کلیک‌های محصولات موجود در آنها بیشترین باشد.

موفق باشید و خوش بگذره!

راستی! شرکت ترب در موقعیت‌های شغلی مختلف نیرو استخدام می کند؛ اگر مایل به همکاری بودید، حتما سری به صفحه موقعیت‌های شغلی ما بزنید و به دوستانتون هم خبر بدید!