گام اول: دریافت اطلاعات و Preprocess

برای داده‌های ورودی مشتاق بودم تا بتوانم داده‌های توییتر را دریافت کرده و آن‌ها را پردازش کنم؛ پس از ایجاد کد جهت دریافت داده‌های توییتر و استفاده از APIهای داخل دستورکار، ارور گرفتم و متوجه شدم که نیاز به دریافت API از توییتر است. سپس درخواست API کردم. اما توییتر موافقت نکرد و گفت فعلا امکان‌پذیر نیست. سپس به سراغ ساخت بات برای سروش رفتم. پس از ایجاد حساب و ساخت بات برای سروش، توی گوگل سرچ کردم ولی موفق نشدم طریقه عضو کردنِ بات در کانال ها را پیدا کنم و انگلیسی هم نمی‌توانستم سرچ کنم.

سپس به سراغ تلگرام رفتم و چند روزی روی این موضوع سرچ می‌کردم. چندجایی در اینترنت خواندم که امکان عضو کردن بات در کانال‌هایی که خوم ادمین آن‌ها هستم وجود دارد و من نمی‌توانم بات را در کانال دیگران عضو کنم. سپس متوجه شدم که می‌توانم پیام‌های حساب کاربری شخصی خودم را به کمک دریافت api و کتابخانه telethon وارد پایتون کنم. به این شکل امکان دریافت پیام کانال‌های تلگرام وجود داشت. اما توی قسمتی از documentation نوشته بود که اگر سرعت دریافت پیام، علی‌الخصوص برای کشورهای حساس مثل ایران و روسیه، بالا باشد، امکان بسته شدن اکانت شما وجود دارد (لینک مطلب ذکر شده).

به همین دلیل ریسک نکردم و به سراغ توییت‌های سهام‌یاب رفتم

ابتدا کافکا را نصب کرده و به کمک commandهای زیر server و zookeeper را اجرا می‌کنم.

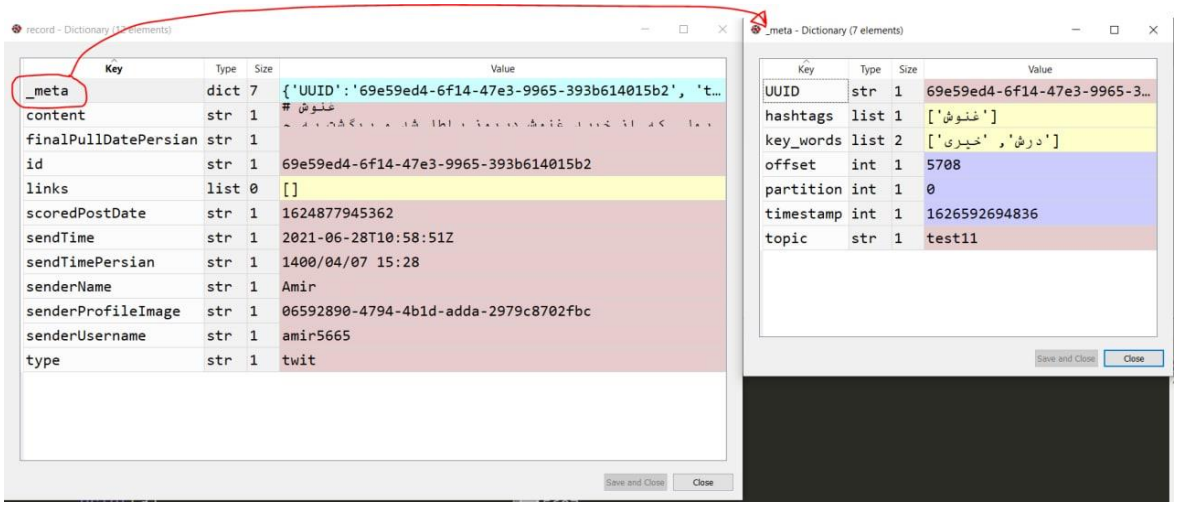
```
### 1st cmd in kafka dir
\ . bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties

### 2nd cmd in kafka dir
\ . bin\windows\kafka-server-start.bat .\config\server.properties
```

حال با فراخوانی مداوم API داده‌ها را از سهام‌یاب دانلود می‌کنیم و سپس آن‌هایی که id تکراری ندارند را به کمک producer وارد topic کافکا به نام raw_data می‌کنیم. در همین حین، با یک consumer داده‌های خام را از تاپیک raw_data دریافت کرده و روی هر توییت پیش پردازش‌های گفته شده را اعمال می‌کنیم.

با دریافت هر توییت، ابتدا یک timestamps و UUID برای آن‌ها تولید کردم. به کمک یک تابع هشتگ‌های موجود در content هر توییت را یافته و آن را به اطلاعات همان توییت اظافه می‌کنیم. همین‌طور لینک‌ها هم استخراج شد.

جهت محاسبه کلمات کلیدی، از معیار tf/idf استفاده شده است. ابتدا لیستی از ایست‌واژه‌ها یافتم و با حذف کردن آن‌ها از متن توییت، tf/idf را محاسبه کردم. در tf/idf فراوانی هر کلمه را در هر توییت می‌یابیم و همین‌طور فراوانی همان کلمه را در همه‌ی توییت‌ها بررسی می‌کنیم. کلمه‌ای از tf/idf بالا برخوردار است (کلمه کلیدی است) کهدر یک توییت وجود دشته باشد اما در بقیه توییت‌ها کمتر تکرار شده باشد. سپس برای هر توییت علاوه بر کلمات خاصی که صورت سوال ذکر کرده است، دو کلمه‌ای که tf/idf بزرگتری دارند را به عنوان کلمات کلیدی در نظر گرفتم.

هر توییت ما یک دیکشنری است. در عکس زیر یک key, value توییت را نمایش داده‌ام و در سمت راست تصویر اطلاعاتی که به عنوان متادیتا ذخدی خواهد شد، قرار دارد



در عکس بالا می‌بینیم که partition برابر صفر و offset برابر ۵۷۰۸ ایت. یعنی ایت توییت در تاپیک test11 و در پارتیشن اول ان و همین‌طور در خانه‌ی ۵۷۰۸ پارتیشن اول قرار دارد.

در نهایت به کمک یک producer جدید داده های پیش‌پردازش شده را در تاپیک proccesed_data قرار می‌دهیم.

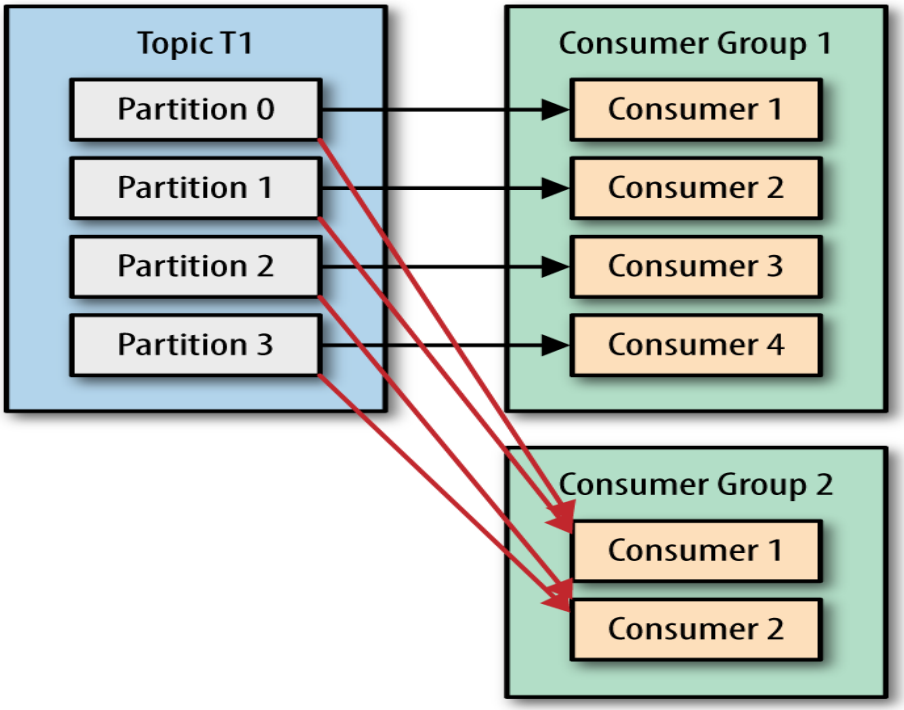
معماری کافکا:

کافکا شامل جمعی از record، topic، consumer، producer، broker، log و cluster است و هر record دارای key (اختیاری)، value و timestamp است.

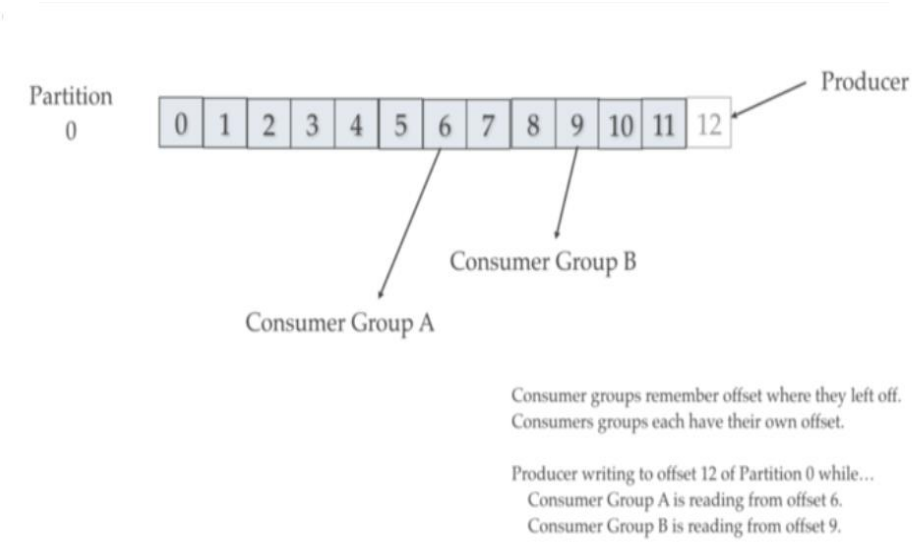
record هایی که وارد کافکا می‌شوند immutable هستند؛ یعنی producer به تنهایی فقط می‌تواند پیام‌ها را به تاپیک topic اظافه کند و امکان تغیر در recordهای موجد را ندارد.

هر تاپیک یک log دارد که محل ذخیره‌ی topic روی دیسک است. هر topic متشکل از تعدادی partition است. هر پارتیشن متشکل از تعدادی offset است (عکس زیر درک بهتری ایجاد می‌کند). هر record وارد topic مورد نظر و سپس وارد partition مورد نظر می‌شود و در آنجا به ترتیب درون offsetها ذخیره می‌شود. برای آنکه چندین consumer بتوانند همزمان record دریافت کنند، نیاز به وجود partition داریم. در عکس زیر

نحوه‌ی اتصال consumerها به partition را می‌بینیم. با ایجاد consumer با group id متفاوت، می‌توان همزمان از یک تاپیک پیام دریافت کرد.



هر پارتیشن بدین شکل است:

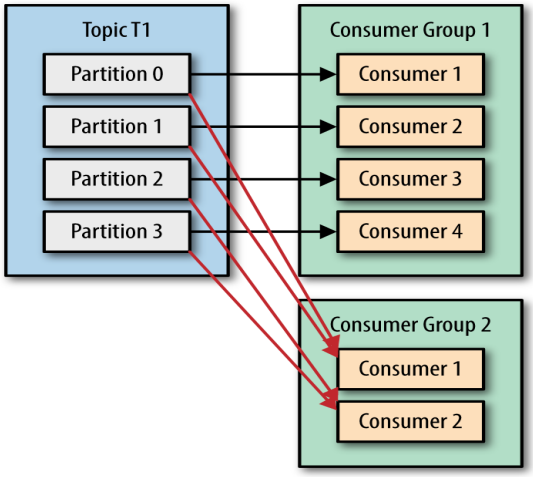


Producer برای تولید جریان داده و ارسال آن به topic استفاده می‌سود. Consumer پیام‌ها را از partition کافکا می‌خواند (هر topic شامل تعدادی partition است).

Cluster کافکا متشکل از تعدادی broker است و هر broker شامل تعدادی topic است.

گام دوم : persistence

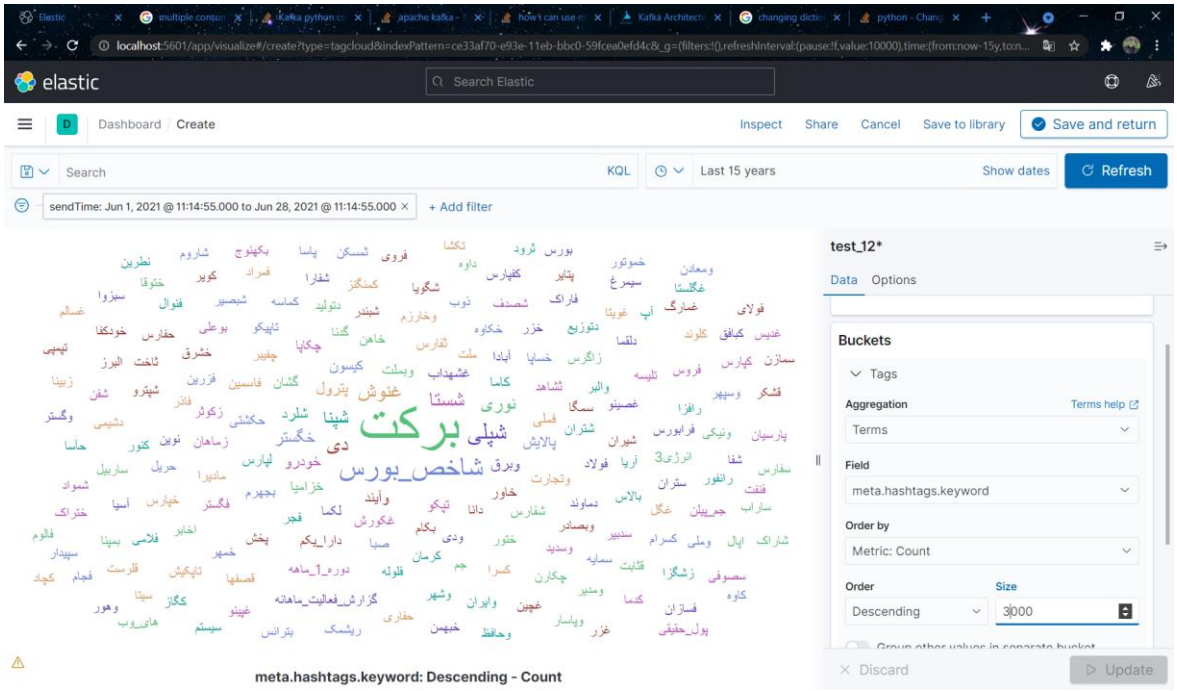
در این قسمت داده های پیش پردازش شده‌ی مرحله قبل را به کمک consumer با group id معین دریافت می‌کنیم (چون در گام های بعدی از تاپیک processed_data داده دریافت خواهیم کرد و در واقع داریم به صورت همزمان از partitionهای واحد توسط چند consumer پیام دریافت می‌کنیم، تعیین group id حیاتی است. این موضوع به عکس زیر اشاره دارد:



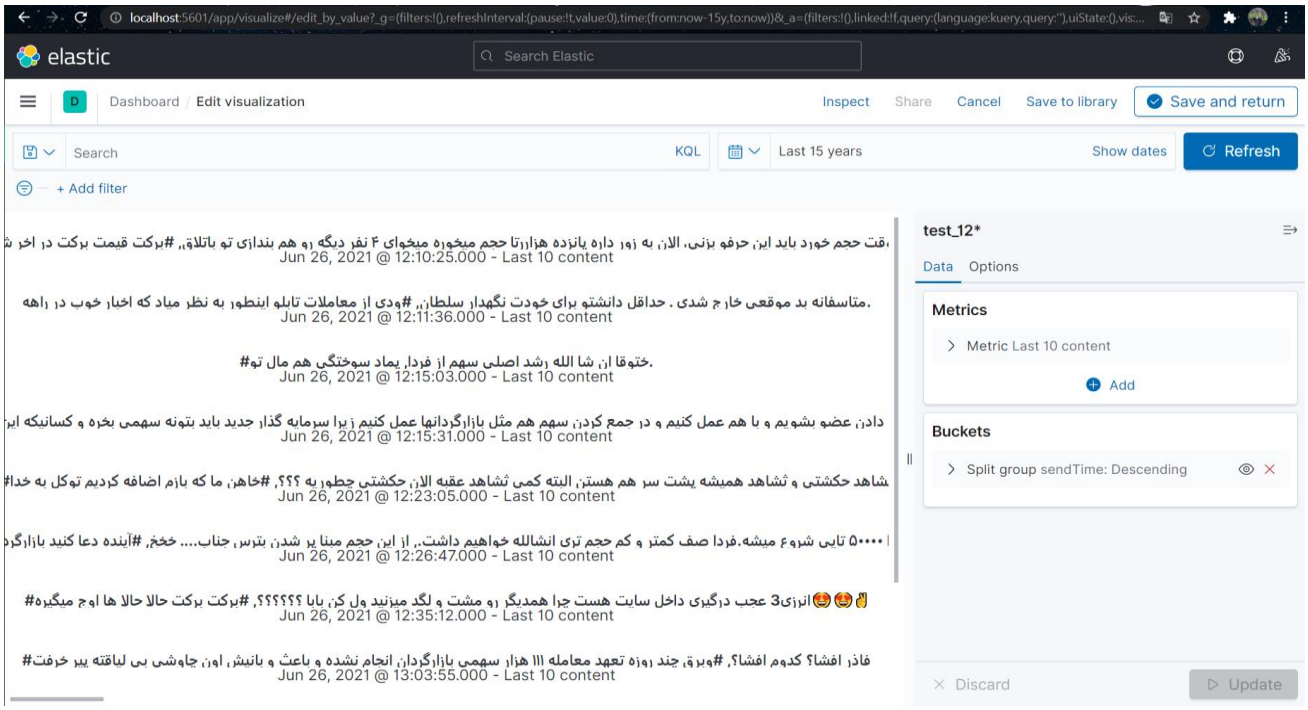
پس از دریافت داده از از کافکا، داده‌ها را وارد index الاستیک سرچ می‌کنیم. سپس به کمک کیبانا، داشبوردهای خواسته شده را نمایش می‌دهیم.

ابر کلمات یک کانال یا خبرگزاری خاص در یک بازه زمانی

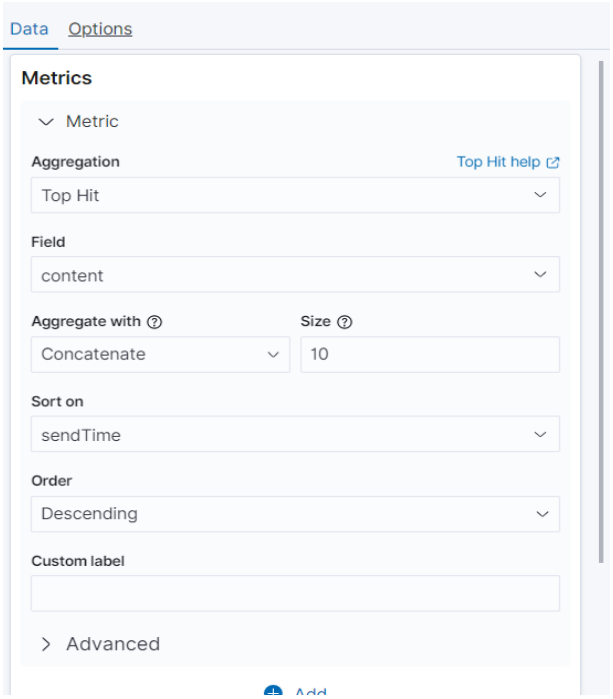
از آنجایی که ما از داده‌های سهام‌یاب استفاده می‌کنیم، من ابر کلمات هشتگ‌ها در یک بازه‌ی خاص را رسم کرده‌ام

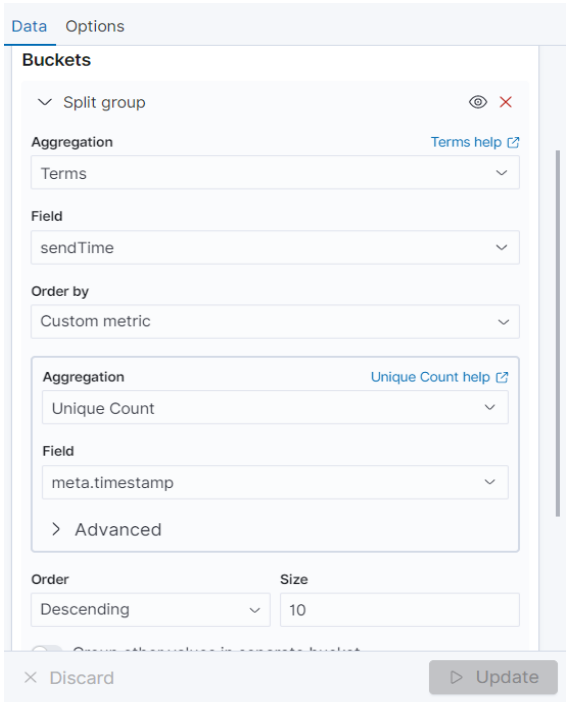


متن ده پست اخیری که دریافت شده است

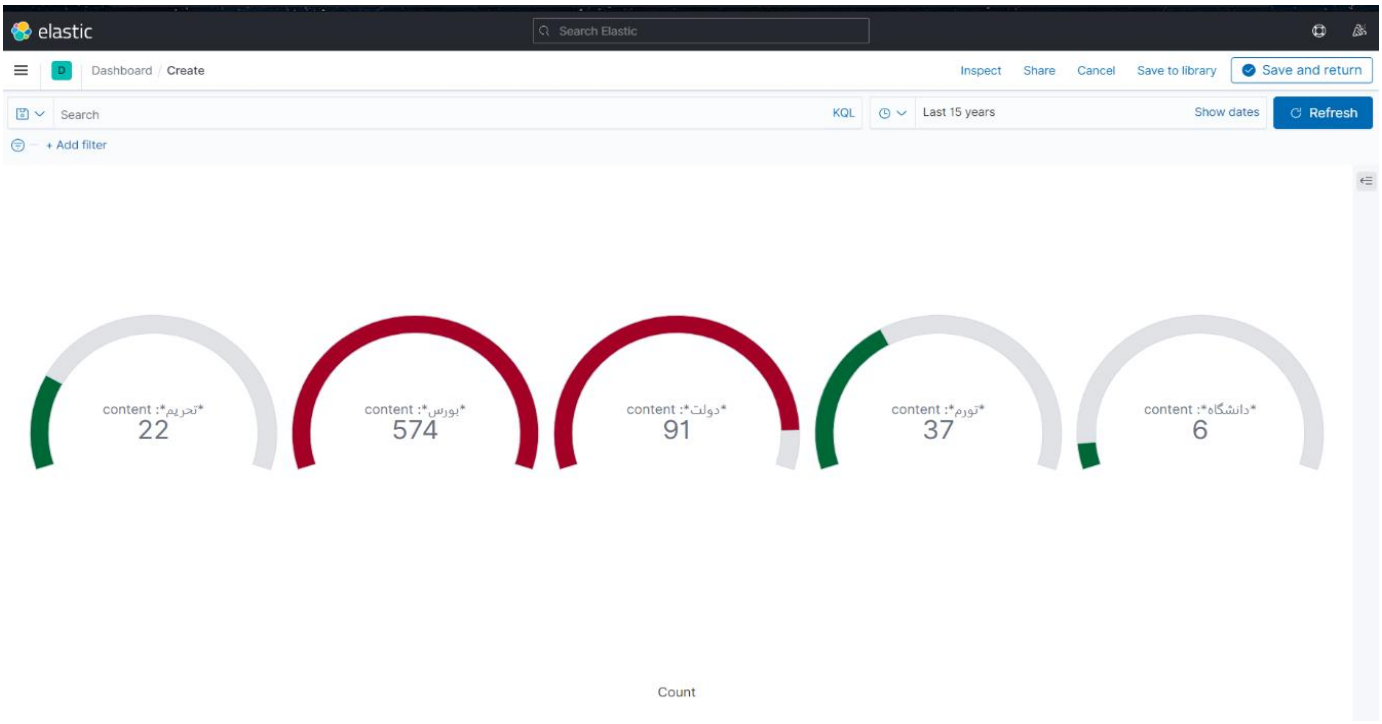


پارامترهای تعیین شده جهت نمایش داشبورد در عکس‌های زیر آورده شده است (یعنی metrics و buckets).

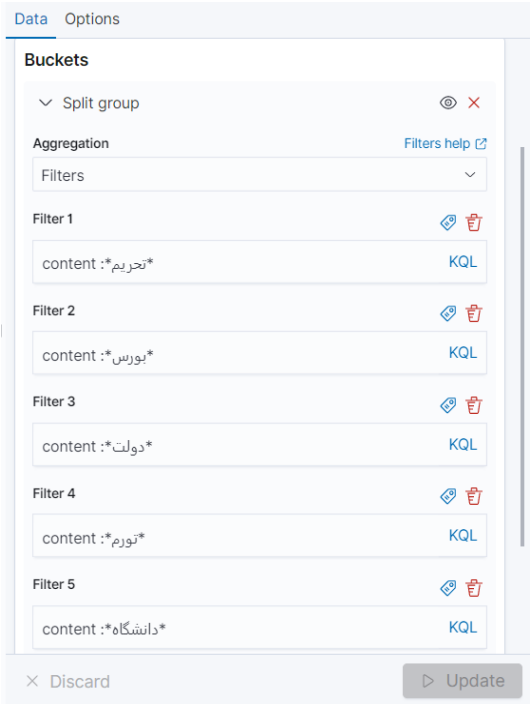




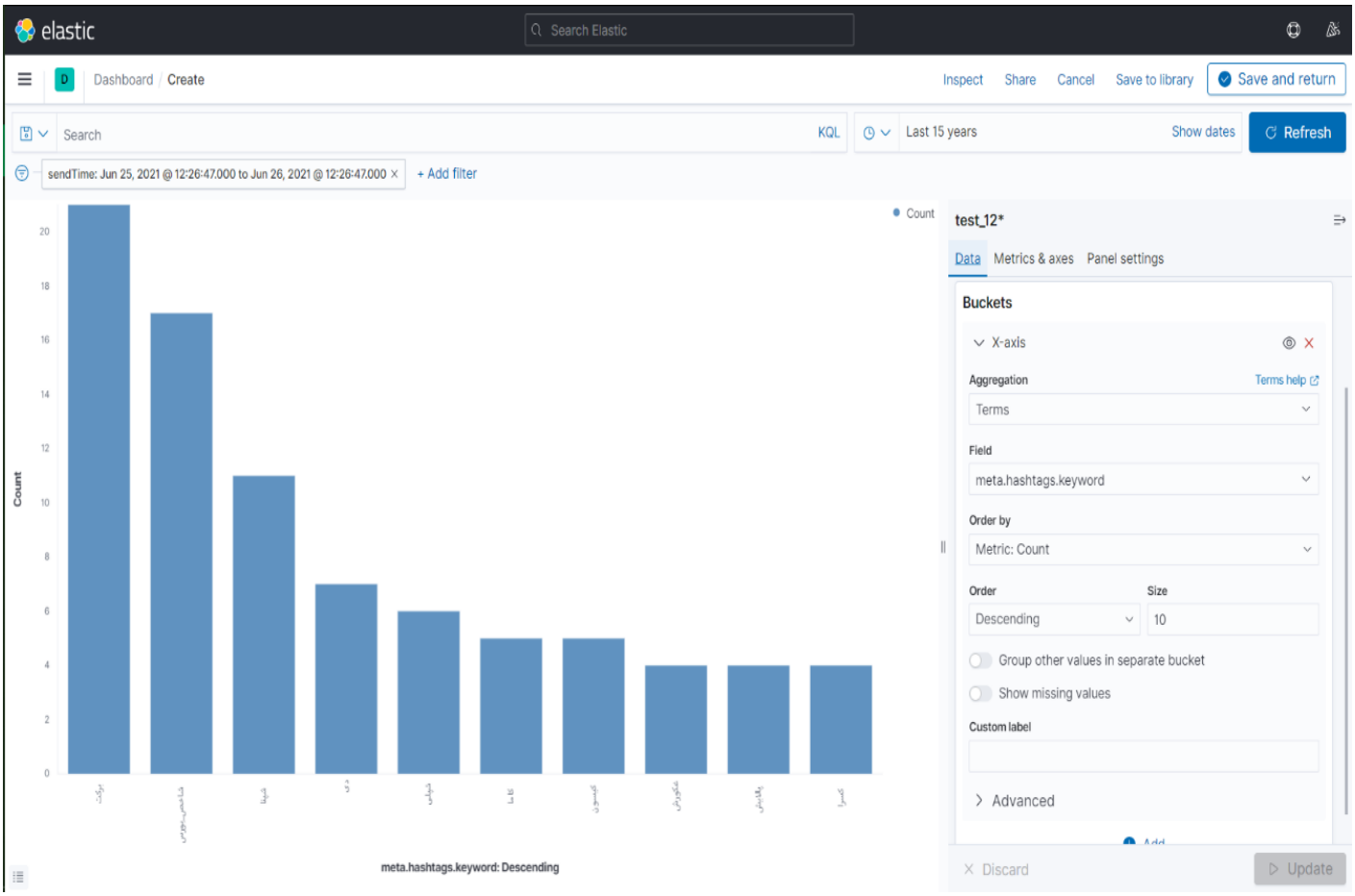
تعداد پست‌های ارسال‌شده به ازای چند تا از کلمات کلیدی خاص
که در مرحله قبل مشخص شده است در یک بازه زمانی خاص



از bucket زیر استفاده شده است

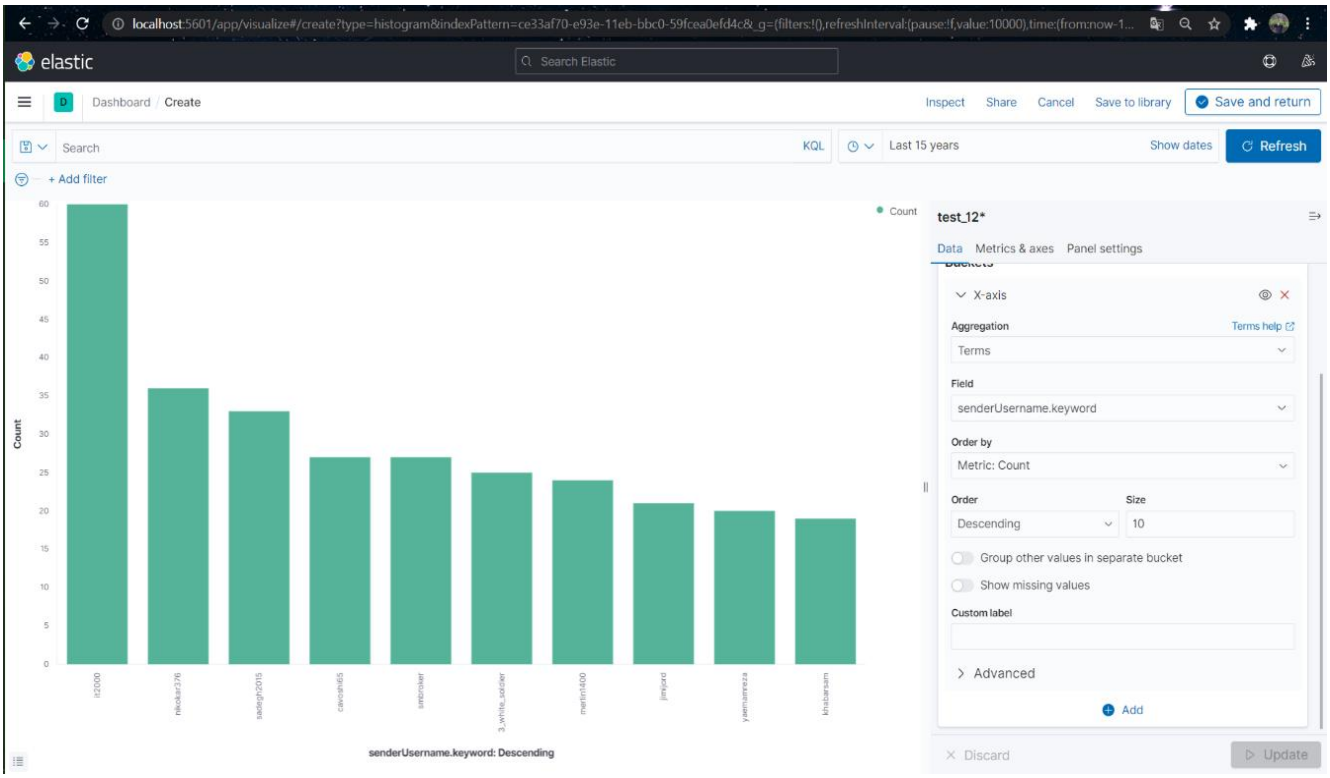


ده هشتگ بیشتر استفاده شده در پست‌های تمام کانال‌ها در یک بازه‌ی زمانی با تعداد تکرار هر هشتگ (یک نمودار ستونی) مثلا هشتگ‌های بیشتر استفاده شده در یک روز اخیر

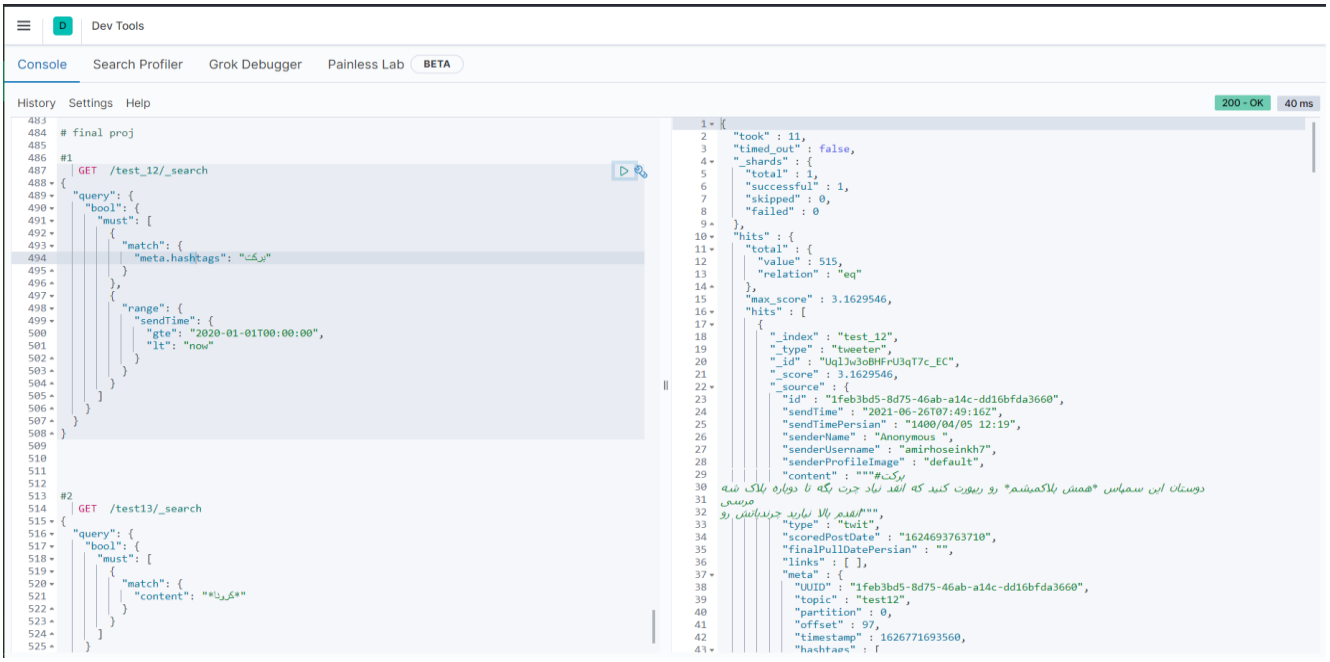


یک نمودار به انتخاب خودتان

در این قسمت تعداد توییت‌های یک کاربر را در یک نمودار ستونی نمایش داده ام



اگر به دنبال تمام پست‌های جاوی یک کلمه خاص از یک خبرگزاری یا کانال خاص در یک بازه مشخص هستیم، چه دستوری باید بنویسیم (و یا یک هشتگ خاص یا یک کاربر خاص در توییت‌ها)



تعداد توییت‌های ارسالی به ازای یک کلمه خاص را به ازای یک هشتگ خاص در توییت‌ها در یک بازه زمانی