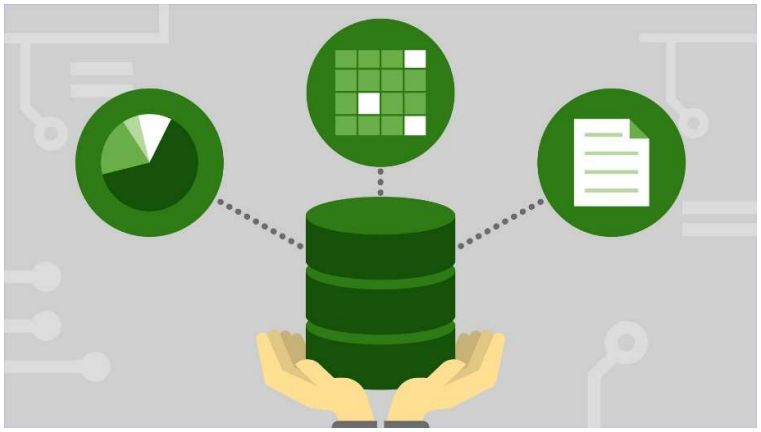
به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر



آزمایشگاه پایگاه‌داده

دستورکار شماره ۷

مهلت تحویل :

۱۴۰۰/۲/۳۱

مجتبی بنائی

دستور کار شماره ۷ - کار با الستیک سرچ

یکی از دیتابیس های معروف در حوزه جستجوی متن با سرعت بالا در ذخیره انواع داده‌های متنی و پاسخگویی به انواع کوئری های کاربر بر روی آنها، الستیک سرچ است که در اکوسیستم استارتاپی ایران هم بسیار پرطرفدار است.

در این دستور کار هم مشابه با دستورکارهای اخیر، هدف اصلی ، آشنایی اولیه با این دیتابیس و نحوه کار با آن است که حداکثر با دوساعت صرف زمان، میتوانید به راحتی آنرا انجام دهید.

کافی است به آدرس زیر مراجعه کرده و تمامی مراحل آنرا انجام دهید :

http://yun.ir/fi9loe

تمام دستورات آنرا از ابتدا تا انتها در محیط کیبانا که محیط گرافیکی کار با الستیک‌سرچ است انجام داده، با گرفتن اسکرین شات از خروجی آنها، گزارش خود را آپلود کنید. داده‌ها را طوری وارد کنید که هر کوئری حداقل دوجواب در خروجی برگرداند.

در انتهای کار، با صدا زدن API زیر در یک برنامه پایتون حداقل هزار توئیت را در الستیک سرچ ذخیره کرده (هشتگ‌ها که همان نمادهای بورسی هستند را جداگانه در یک لیست بریزید و سپس در الاستیک سرچ ذخیره کنید) و یک داشبورد با حداقل دو ویژوالیزیشن ایجاد کنید (مثلا ابرهشتگ‌ها یا تعداد نمادهای پرتکرار)

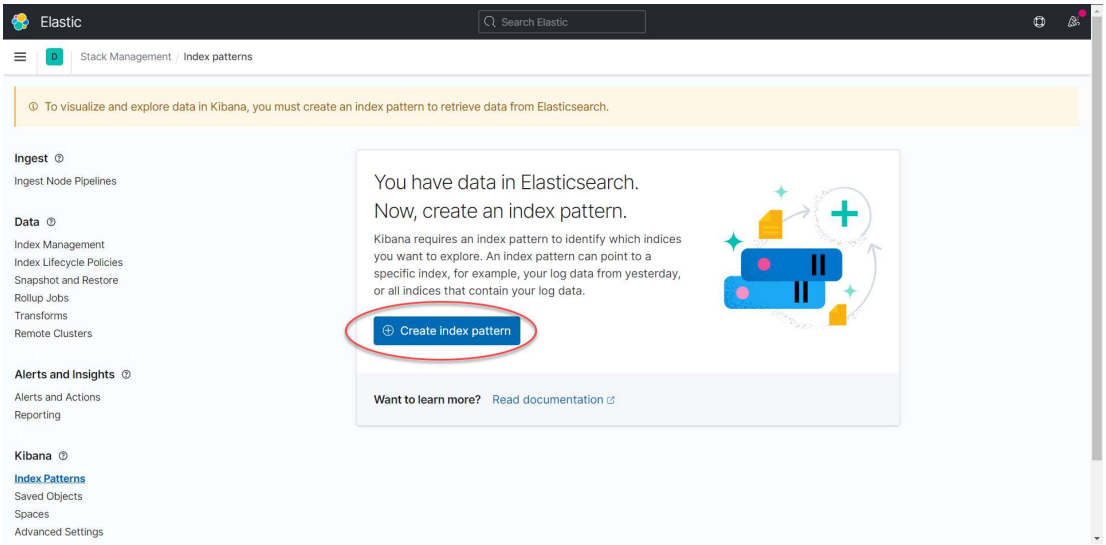
https://www.sahamyab.com/guest/twiter/list?v=0.1

نمونه کد مورد نیاز و نحوه ایجاد یک داشبورد بر اساس داده‌های متنی می تواند به صورت زیر باشد :

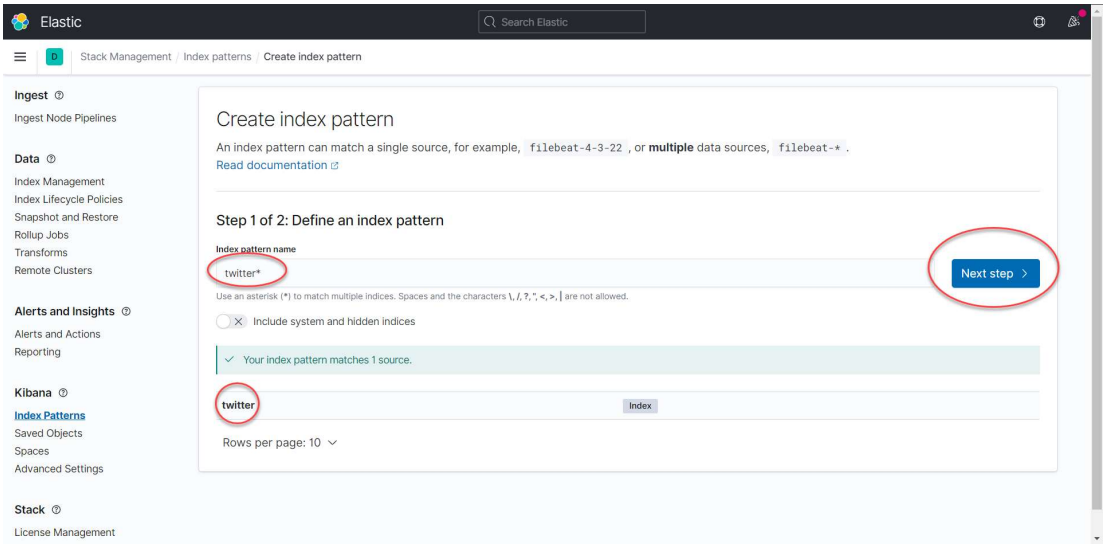
```python
import requests
from elasticsearch import Elasticsearch
import re

url = 'https://www.sahamyab.com/guest/twiter/list?v=0.1'
elasticSearch = Elasticsearch([{'host': 'localhost', 'port': 9200}])
total = 1000
fetched = 0
seenIds = set()
hashtags = list()

while fetched < total:
    response = requests.get(url=url)
    if response.status_code != 200:
        print('HTTP', response.status_code)
        continue
    data = response.json()["items"]  # Check the JSON Response Content documentation below
    for tweet in data:
        if tweet["id"] not in seenIds:
            try:
                tweet["hashtags"] = re.findall(r"#(\w+)", tweet["content"])
                elasticSearch.index(index="twitter", doc_type="twitter", body=tweet)
                seenIds.add(tweet["id"])
                fetched += 1
                print("tweet " + str(tweet["id"]) + "fetched,   total: " + str(fetched))
            except Exception as e:
                print(e)
```

نحوه ایجاد داشبورد هم به صورت زیر است :

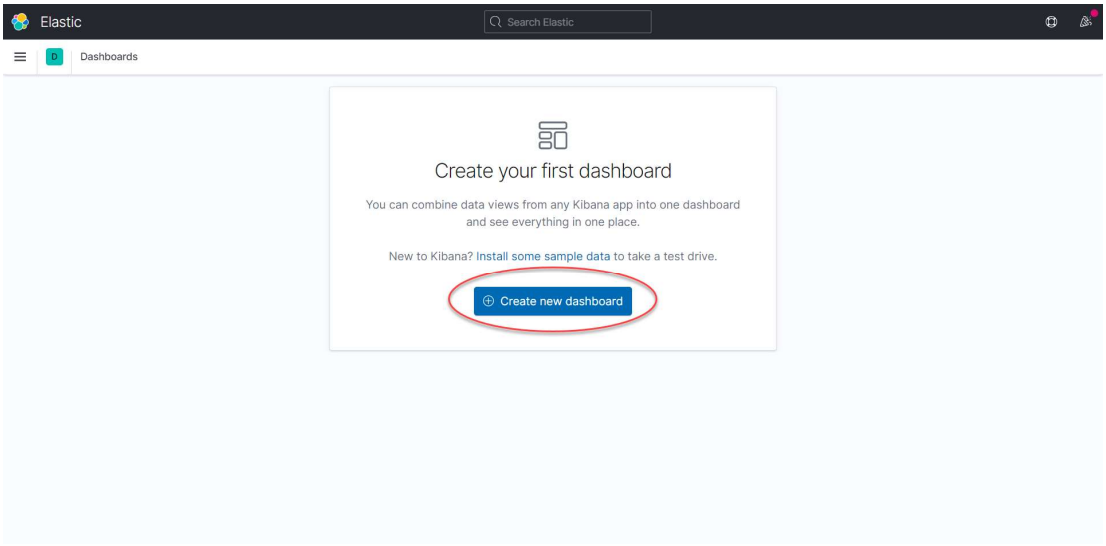
یک index pattern می سازیم.



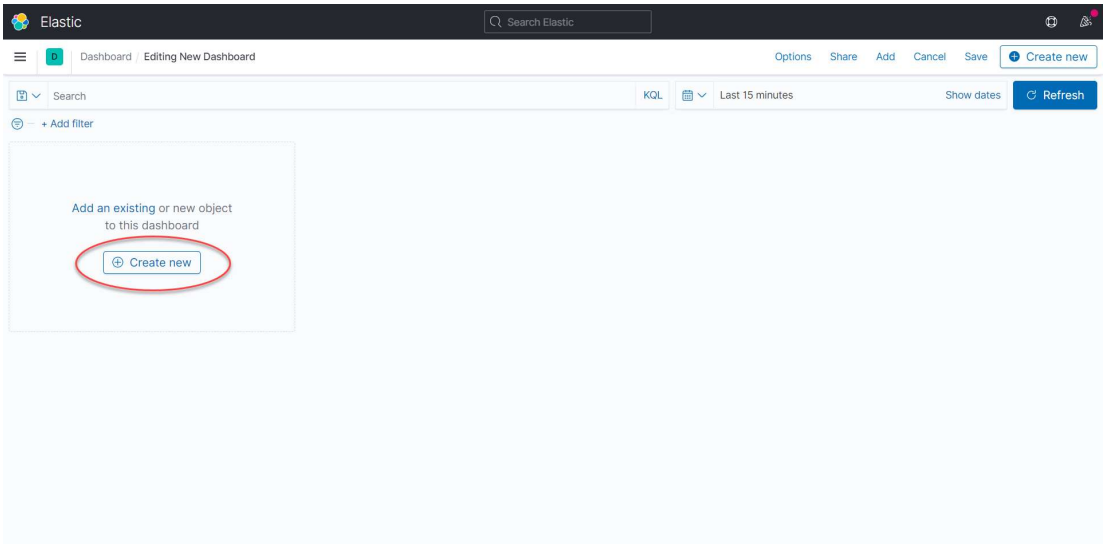
ایندکس با نام twitter را انتخاب کرده و next را می زنیم.



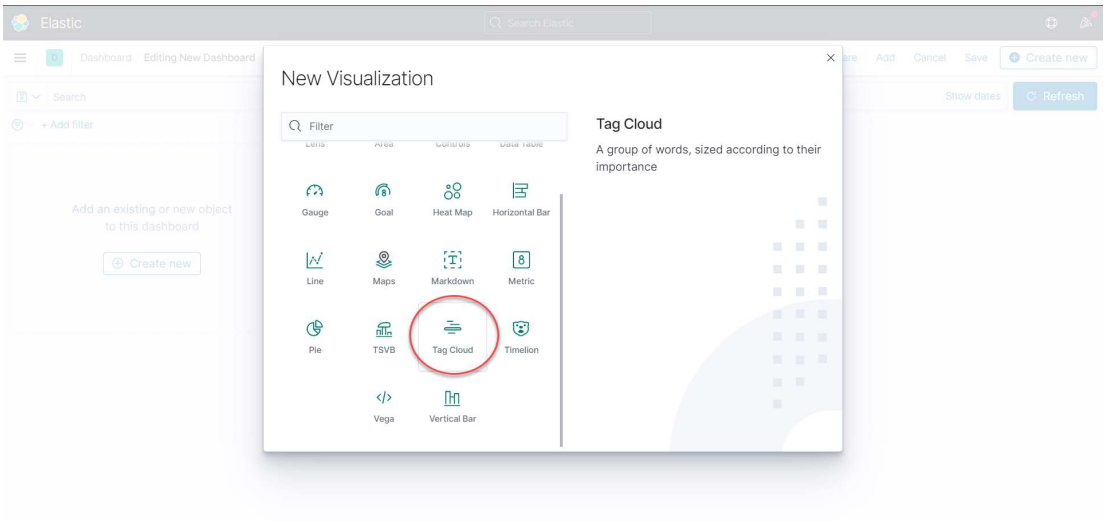
سپس یک داشبورد می سازیم.



سپس برای create new را می زنیم تا twitter را اضافه کنیم.

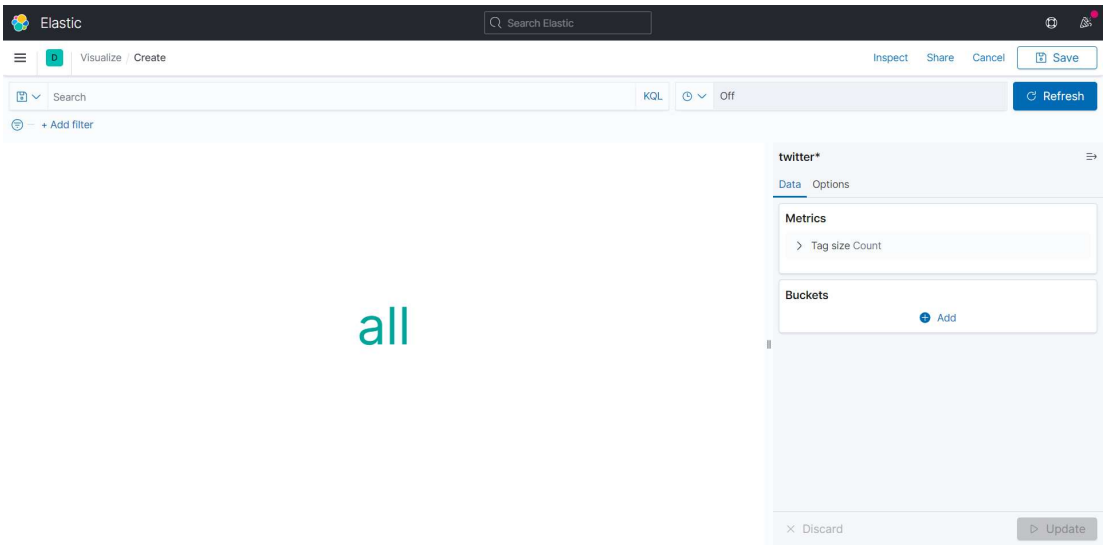


برای visualization اول tag cloud را انتخاب می کنیم که بسته به تعداد تکرار کلمه، سایز آن ست می کند.

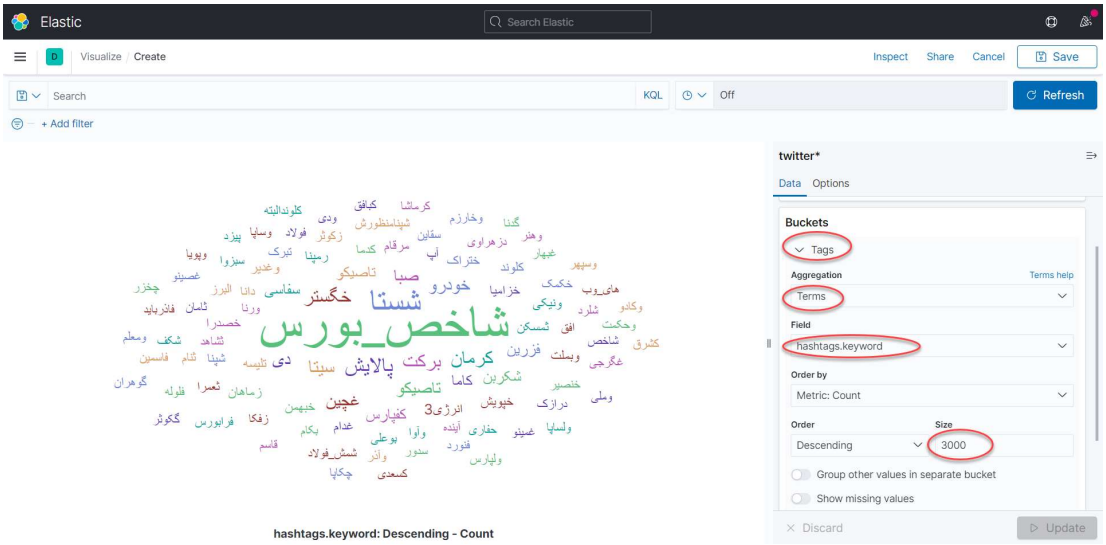


سپس twitter را انتخاب می کنیم.

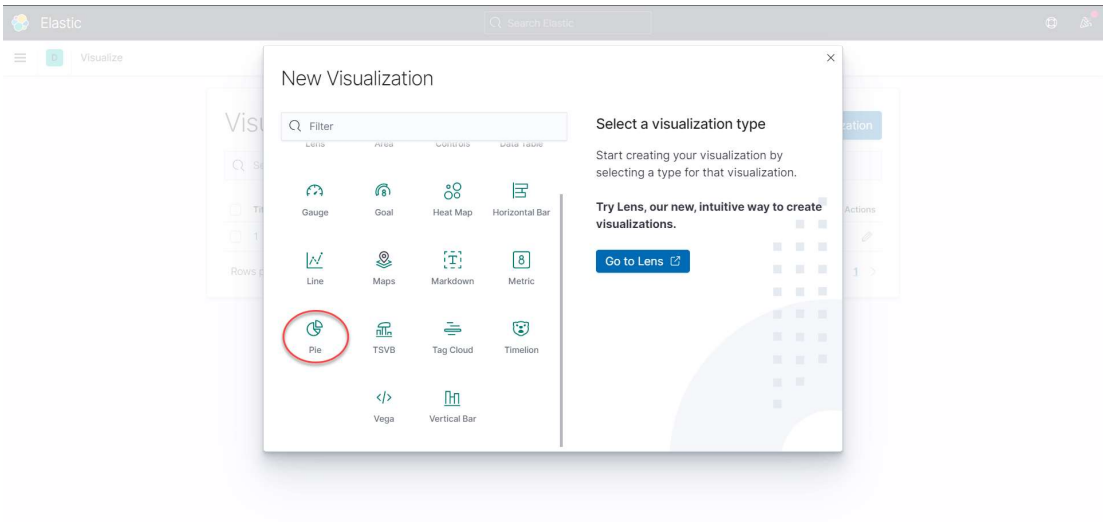
در ابتدا تنها all نشان داده می شود.



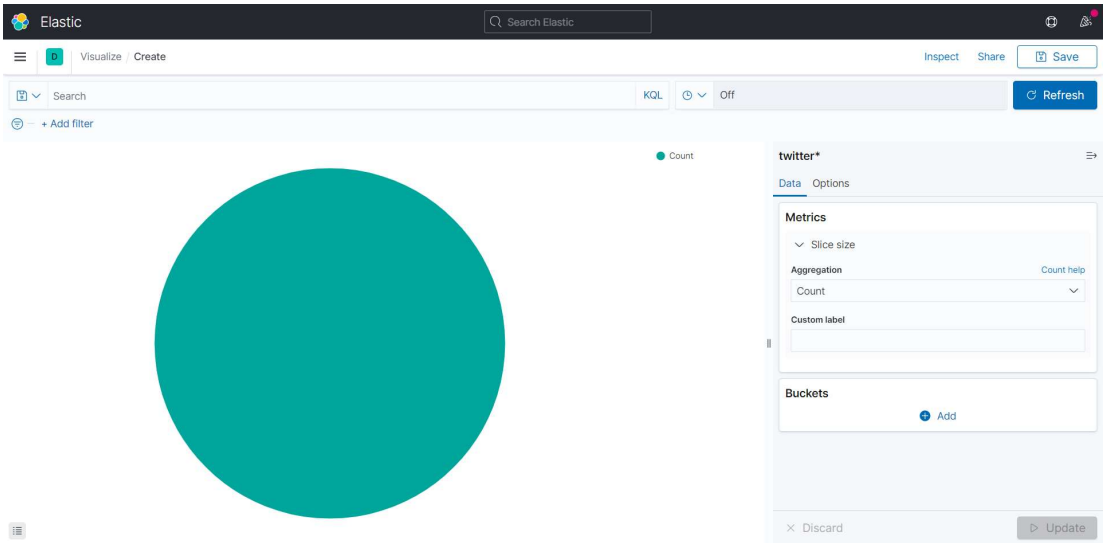
تنظیمات رو در سمت راست تغییر می دهیم تا برای hashtag ها این visualization انجام شود که خروجی به صورت زیر می شود.



برای visualization دوم، Pie را انتخاب می کنیم که سایز slice دایره بر اساس تعداد تکرار کلمه می شود.



خروجی اولیه به این صورت می شود.



در تنظیمات bucket مانند قبل hashtags رو انتخاب می کنیم و خروجی نهایی به این صورت می شود.