Data cleaning

twitter의 데이터를 받아오지만 아직 정리가 안된 느낌이다.

1. 데이터의 필드에서 text 부분의 mapping 을 keyword 에서 text로 변경 하기
2. 필요한 데이터만 가지고오기

1. Mapping 추가 설정

데이터를 받아오게되면 twitter의 글을 보면 context의 내용이 keyword로 저장되어있다.

```
GET /bts_twitter/_mapping/_doc/field/text
```

```
{
  "bts_twitter" : {
    "mappings" : {
      "_doc" : {
        "text" : {
          "full_name" : "text",
          "mapping" : {
            "text" : {
              "type" : "text",
              "fields" : {
                "keyword" : {
                  "type" : "keyword",
                  "ignore_above" : 256
                }
              }
            }
          }
        }
      }
    }
  }
}
```

keyword type과 text type 의 차이를 알아보자

keyword

https://www.elastic.co/guide/en/elasticsearch/reference/6.7/keyword.html

text

https://www.elastic.co/guide/en/elasticsearch/reference/6.7/text.html

간단한 차이를 말하면 text 로 저장하면 해당 필드에 대한 aggregation 분석을 할수 없고, keyword는 aggregation이 분석이 가능하다.

기존의 맵핑을 조회하면 다음과 같다. 새로운 맵핑을 업데이트해주면 기존의 것의 맵핑 정보는 변하지 않기 때문에, 지우고 새롭게 맵핑 정보를 넣어준다

```
"mappings":{
   "_doc":{
      "properties":{
         "post_date":{
            "type":"date"
         },
         "message":{
            "type":"text",
            "fields":{
               "keyword":{
                  "ignore_above":256,
                  "type":"keyword"
               }
            }
         },
         "user":{
            "type":"text",
            "fields":{
               "keyword":{
                  "ignore_above":256,
                  "type":"keyword"
               }
            }
         }
      }
   }
```

변경한 맵핑 내용은 다음과 같다.

```
PUT bts_twitter
{
  "mappings": {
    "_doc": {
      "properties": {
        "created_at": {
          "type": "date",
          "format": "EEE MMM dd HH:mm:ss Z yyyy"
        },
        "text":{
          "type": "text"
        }
```

```
      }
    }
  }
}
```

그리고 다시 데이터를 넣어주는 방법과 mapping을 업데이트 하는 방법이 있다.

2. 필요한 데이터 가지고오기

이전에 데이터에서 불필요한 정보를 제거한 정보

```
input {
  twitter {
      consumer_key => "abn2I1hzwIVQItsqJfxDbtzTu"
      consumer_secret => "KY7Rvo8ub8PuSnrVd6oXt78mmMbYw3LBelRezZ919iNnlsVtap"
      oauth_token => "1116880164212731904-KgeOGuZFkpeUYANORZ8tmM999nPnOw"
      oauth_token_secret => "ZvZSfNtI8hS1gSm1vlOv9KLBLzqCMm932DkUlswKiRd1E"
      keywords => ["bts"]
      full_tweet => true
  }
}
filter {
  if [lang] != 'ko'{
    drop{}
  }
  mutate{
        remove_field =>
["extended_entities","possibly_sensitive","is_quote_status","entities","retweeted_status","
in_reply_to_user_id_str","in_reply_to_status_id","in_reply_to_screen_name","in_reply_to_scr
een_name",
"coordinates","contributors","in_reply_to_user_id_str","in_reply_to_status_id","id_str","re
tweeted_status","geo","reply_count","user","quote_count","in_reply_to_status_id_str","filte
r_level","possibly_sensitive","retweeted","timestamp_ms","source","truncated","is_quote_sta
tus","favorite_count","entities","place","in_reply_to_user_id"]
  }

  date{
    match => ["created_at", "EEE MMM dd HH:mm:ss Z yyyy"]
    timezone => "UTC"
    locale => "ko"
  }

}
output {
   stdout{}
   elasticsearch{
     hosts => "http://127.0.0.1:9200"
     index => "bts_twitter"
     document_type => "_doc"
   }
}
```

```
~
"bts_twitter.conf" 33L, 1221C
```

필터링 하여 가지고온 데이터의 결과

```
"hits" : [
      {
        "_index" : "bts_twitter",
        "_type" : "_doc",
        "_id" : "aBhDYmoB_XnxK7PEt7X1",
        "_score" : null,
        "_source" : {
          "tags" : [
            "_dateparsefailure"
          ],
          "lang" : "ko",
          "id" : 1122361299868868608,
          "text" : """
RT @morethanever_jk: 박자감이 아주 좋은 토끼일세..🐰 #정국
#BBMAsTopSocial BTS @BTS_twt https://t.co/BTEtaOLJ5V
""",
          "@timestamp" : "2019-04-28T04:46:01.000Z",
          "favorited" : false,
          "@version" : "1",
          "created_at" : "Sun Apr 28 04:46:01 +0000 2019",
          "retweet_count" : 0
        },
        "sort" : [
          1556426761000
        ]
      }
    ]
```

추가로 해야할점 : http 주소등을 다 제거하기

이렇게 되면 text는 이제 analyzer를 통해서 문서 전체 데이터를 검색하고 사용할 수 있게 된다.

https://www.elastic.co/guide/en/elasticsearch/reference/current/search-analyzer.html

```
GET bts_twitter/_doc/_search
{
  "query":{
    "match":{
      "text":{
        "query": "정국"
      }
    }
  }
}
```

결과


```
"hits" : {
    "total" : 840,
    "max_score" : 3.0186539,
    "hits" : [
      {
        "_index" : "bts_twitter",
        "_type" : "_doc",
        "_id" : "Fhg_YmoB_XnxK7PE2KiQ",
        "_score" : 3.0186539,
        "_source" : {
          "tags" : [
            "_dateparsefailure"
          ],
          "lang" : "ko",
          "id" : 1122360234461122561,
          "text" : """
RT @IDAREU_JK: 190426 미화당 팬사인회 - ♡♡♡ BTS 정국 JUNGKOOK Focus.

4K 🎥 https://t.co/OHg83W34aB (Full ver.)

#정국 #JUNGKOOK #BTS @BTS_twt
#B…
""",
          "@timestamp" : "2019-04-28T04:41:46.000Z",
          "favorited" : false,
          "@version" : "1",
          "created_at" : "Sun Apr 28 04:41:46 +0000 2019",
          "retweet_count" : 0
        }
      },
      {
        "_index" : "bts_twitter",
        "_type" : "_doc",
        "_id" : "FRhAYmoB_XnxK7PEHqkg",
        "_score" : 3.0186539,
        "_source" : {
          "tags" : [
            "_dateparsefailure"
          ],
          "lang" : "ko",
          "id" : 1122360309782462464,
          "text" : """
RT @IDAREU_JK: 190426 미화당 팬사인회 - ♡♡♡ BTS 정국 JUNGKOOK Focus.

4K 🎥 https://t.co/OHg83W34aB (Full ver.)

#정국 #JUNGKOOK #BTS @BTS_twt
#B…
""",
          "@timestamp" : "2019-04-28T04:42:04.000Z",
          "favorited" : false,
```

```
          "@version" : "1",
          "created_at" : "Sun Apr 28 04:42:04 +0000 2019",
          "retweet_count" : 0
        }
      },
      {
        "_index" : "bts_twitter",
        "_type" : "_doc",
        "_id" : "RRhFYmoB_XnxK7PEbbtA",
        "_score" : 3.0186539,
        "_source" : {
          "tags" : [
            "_dateparsefailure"
          ],
          "lang" : "ko",
          "id" : 1122361769349959681,
          "text" : """
RT @IDAREU_JK: 190426 미화당 팬사인회 - ♡♡♡ BTS 정국 JUNGKOOK Focus.

4K 🎥 https://t.co/OHg83W34aB (Full ver.)

#정국 #JUNGKOOK #BTS @BTS_twt
#B…
""",
          "@timestamp" : "2019-04-28T04:47:52.000Z",
          "favorited" : false,
          "@version" : "1",
          "created_at" : "Sun Apr 28 04:47:52 +0000 2019",
          "retweet_count" : 0
        }
      },
      {
        "_index" : "bts_twitter",
        "_type" : "_doc",
        "_id" : "GxhHYmoB_XnxK7PEdMLg",
        "_score" : 3.0186539,
        "_source" : {
          "tags" : [
            "_dateparsefailure"
          ],
          "lang" : "ko",
          "id" : 1122362326282190849,
          "text" : """
RT @IDAREU_JK: 190426 미화당 팬사인회 - ♡♡♡ BTS 정국 JUNGKOOK Focus.

4K 🎥 https://t.co/OHg83W34aB (Full ver.)

#정국 #JUNGKOOK #BTS @BTS_twt
#B…
""",
          "@timestamp" : "2019-04-28T04:50:05.000Z",
          "favorited" : false,
          "@version" : "1",
```

```
          "created_at" : "Sun Apr 28 04:50:05 +0000 2019",
          "retweet_count" : 0
        }
      },
      {
        "_index" : "bts_twitter",
        "_type" : "_doc",
        "_id" : "gBhHYmoB_XnxK7PEScHj",
        "_score" : 2.9561784,
        "_source" : {
          "tags" : [
            "_dateparsefailure"
          ],
          "lang" : "ko",
          "id" : 1122362280593645568,
          "text" : """
中 웨이보 '방탄소년단 정국' 직캠 3시간만에 105만 돌파! "대륙 뒤흔들다"
#정국 #전정국 #jungkook #BTS @BTS_twt https://t.co/7sLJsJtnHG
""",
          "@timestamp" : "2019-04-28T04:49:54.000Z",
          "favorited" : false,
          "@version" : "1",
          "created_at" : "Sun Apr 28 04:49:54 +0000 2019",
          "retweet_count" : 0
        }
      }
    ]
```

다음에는 nori를 적용해보자 http://를 제거를 해야하지만 먼저 하고싶어서 일단 시작

번외

retweet이 0 이 아닌것을 찾았는데 전부 ㅠㅠ 없다라고 나오네

```
GET bts_twitter/_search
{
  "query": {
    "bool": {
      "must_not": [
        {
          "match": {
            "retweet_count": 0
          }
        }
      ]
    }
  }
}
```