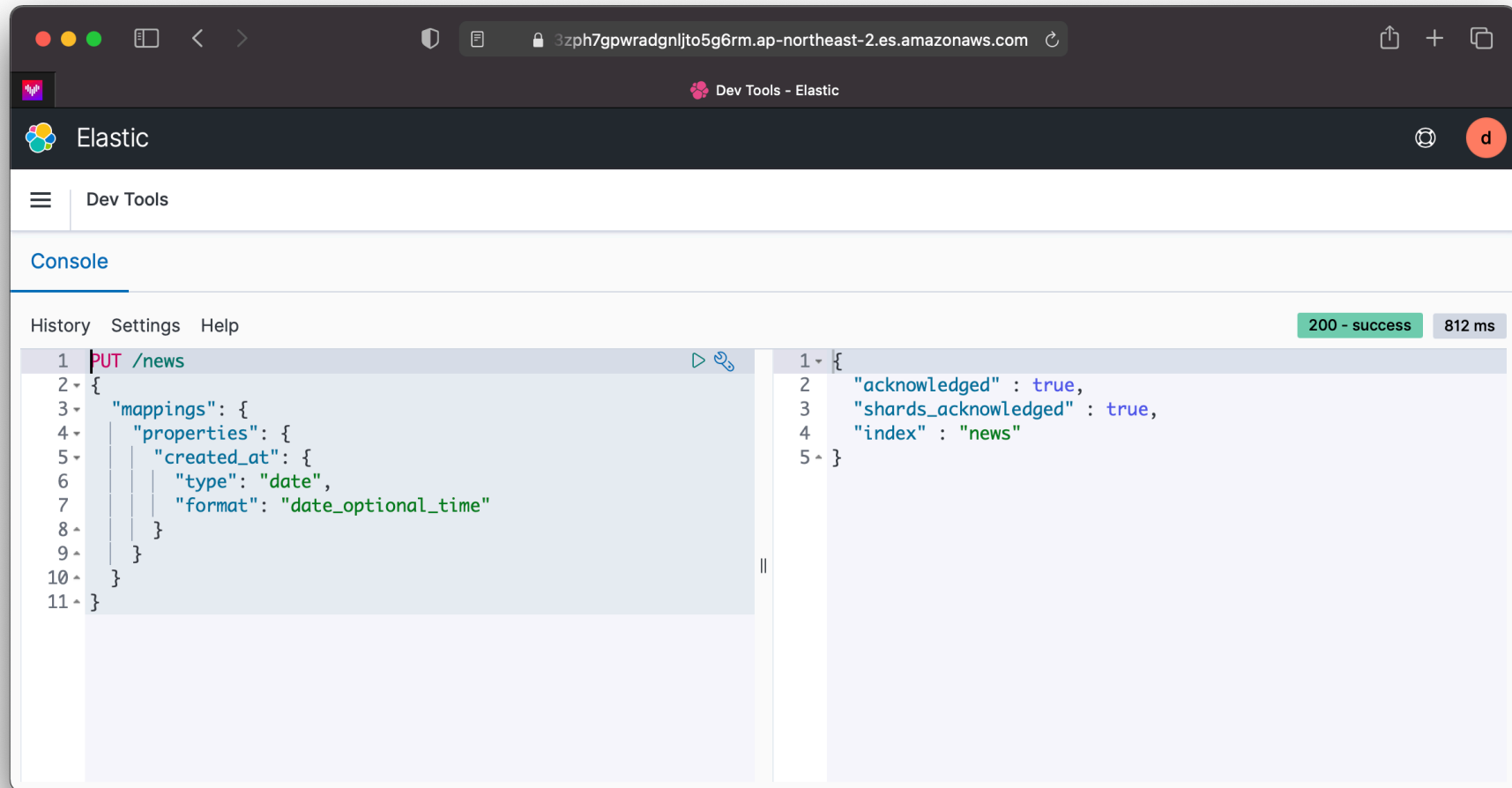


뉴스 기사 수집하기

BAF675 금융 빅데이터 분석

이재훈, Week 5

ElasticSearch: 뉴스 본문 저장할 인덱스 타입 선언



뉴스 본문 크롤러

- news-contents-crawler.py 파일 참조
 - SQS 에서 메시지 꺼내오기
 - 해당 뉴스 본문 크롤링
 - Parsing: 언론사, 뉴스 등록 날짜, 원본 url, 뉴스 본문, 기자 이름 및 이메일 등
 - 꺼내온 데이터 Elasticsearch 서버에 저장

Regular Expression (Regex)

- Reference: <https://docs.python.org/3/library/re.html>
- Sandbox: <https://regex101.com/>
- Usage
 - `re.search(pattern, string)`
 - `re.findall(pattern, string)`
 - `re.sub(pattern, replacement, string)`

Regular Expression (Regex)

- Special characters
 - . (dot) : matches any character
 - \w : word character
 - \s : empty space
 - + : resulting RE to match 1 or more repetitions
 - * : resulting RE to match 0 or more repetitions
 - ^ (caret) : beginning of string
 - \$ (dollar) : end of string

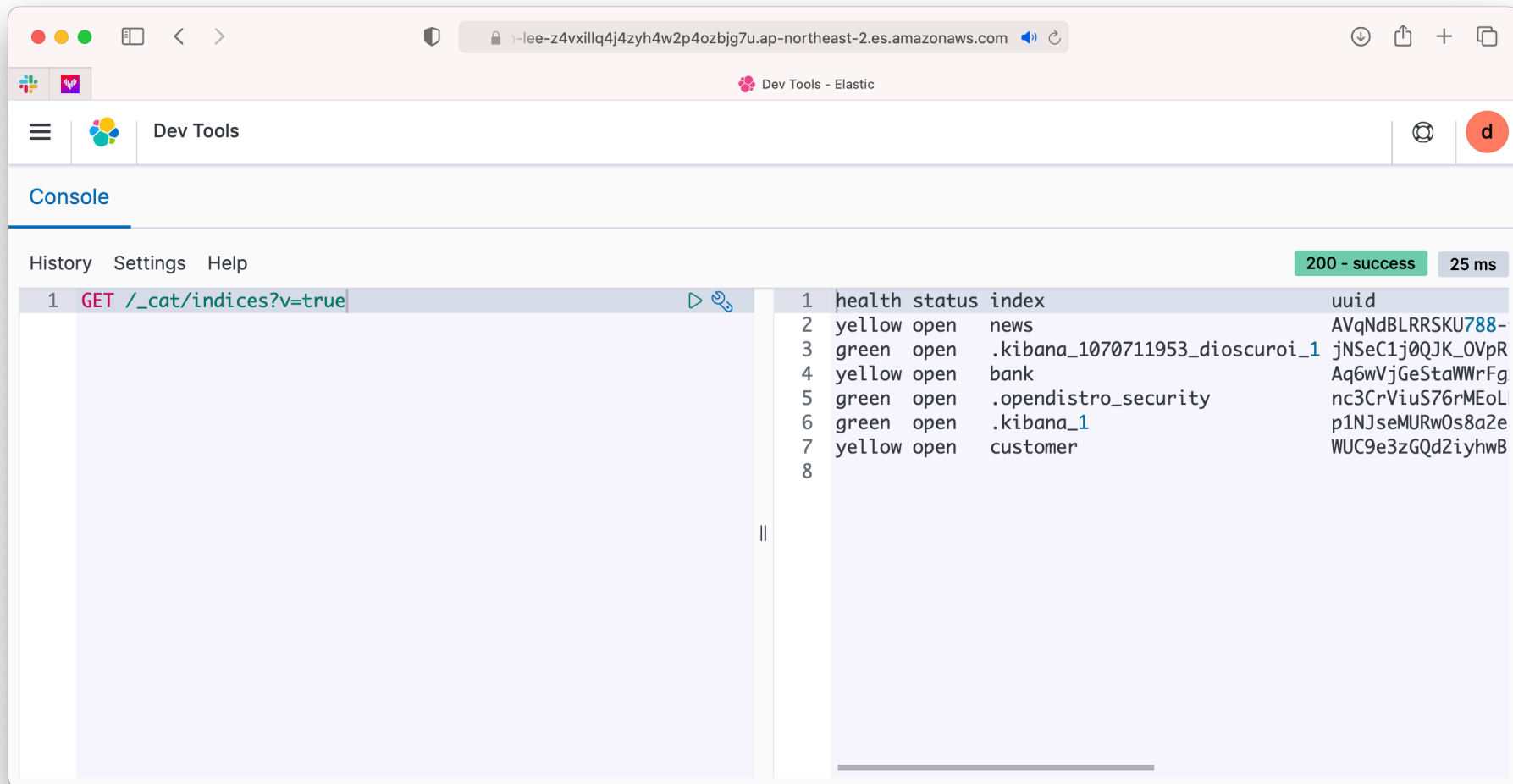
Regular Expression (Regex)

- Special characters
 - [] : set of characters
 - () : indicates a group
 - | : A|B indicates A or B

Regular Expression (Regex)

- Examples
 - `re.findall(r'([\wㄱ-힣]+)\s*기자', body_text)`
 - `re.findall(r'[\w\.]++@[\w\.]++', body_text)`
 - `re.sub(r'동영상 뉴스\s*', '', body_text)`
 - `re.sub(r'▶\s.*\n', '', body_text)`

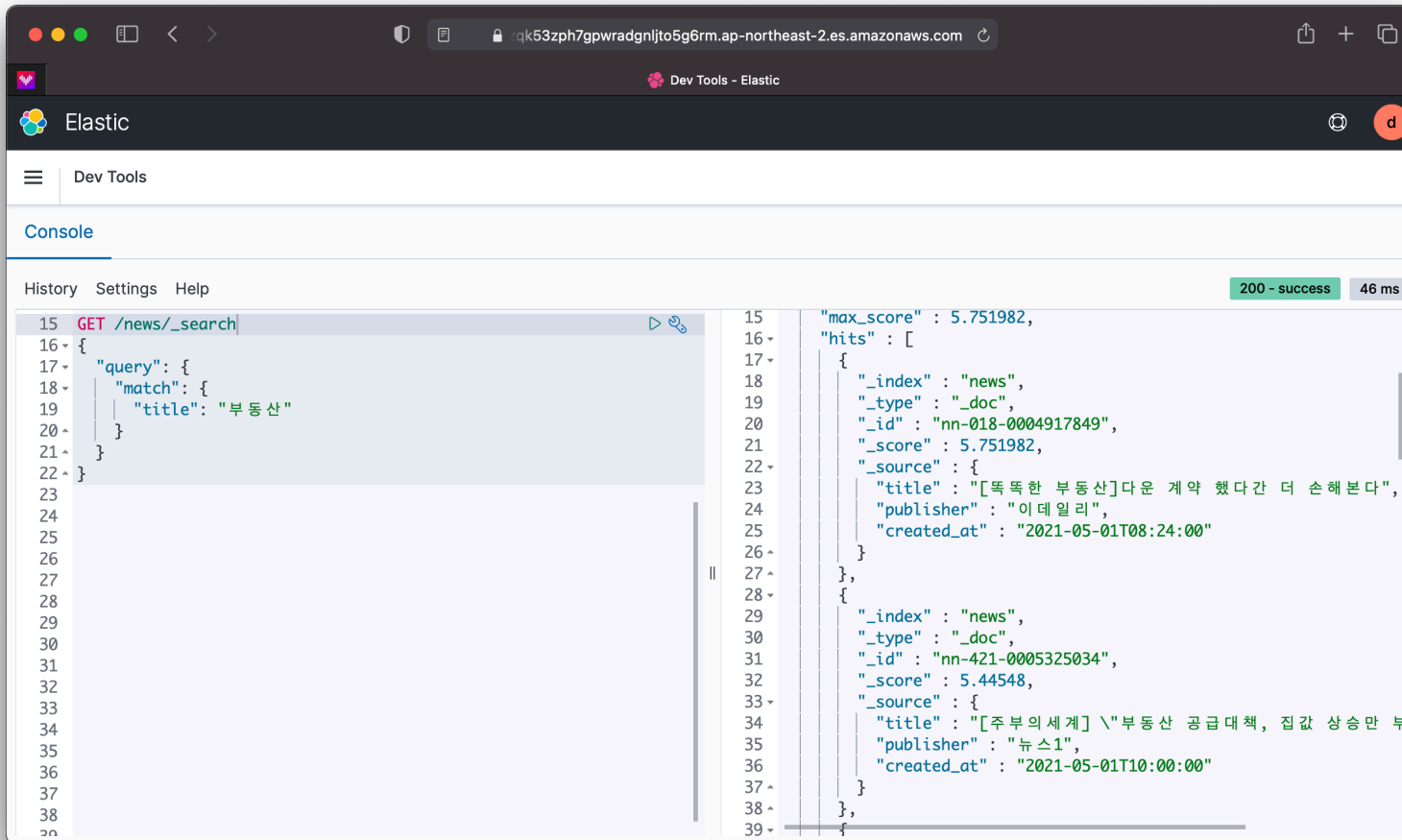
ElasticSearch: 수집된 데이터 용량 확인



The screenshot shows the Elastic Dev Tools console with the following data:

health	status	index	uuid
yellow	open	news	AVqNdBLRRSKU788-
green	open	.kibana_1070711953_dioscuroi_1	jNSeC1j0QJK_OVpR
yellow	open	bank	Aq6wVjGeStaWwRfg
green	open	.opendistro_security	nc3CrViuS76rMEoL
green	open	.kibana_1	p1NJseMURw0s8a2e
yellow	open	customer	WUC9e3zGQd2iyhwB

ElasticSearch: 부동산 뉴스 검색



The screenshot shows the Elastic Dev Tools interface. The console displays a GET request to `/news/_search` with a query for "부동산". The response shows two hits from the "news" index, each with a score and source details.

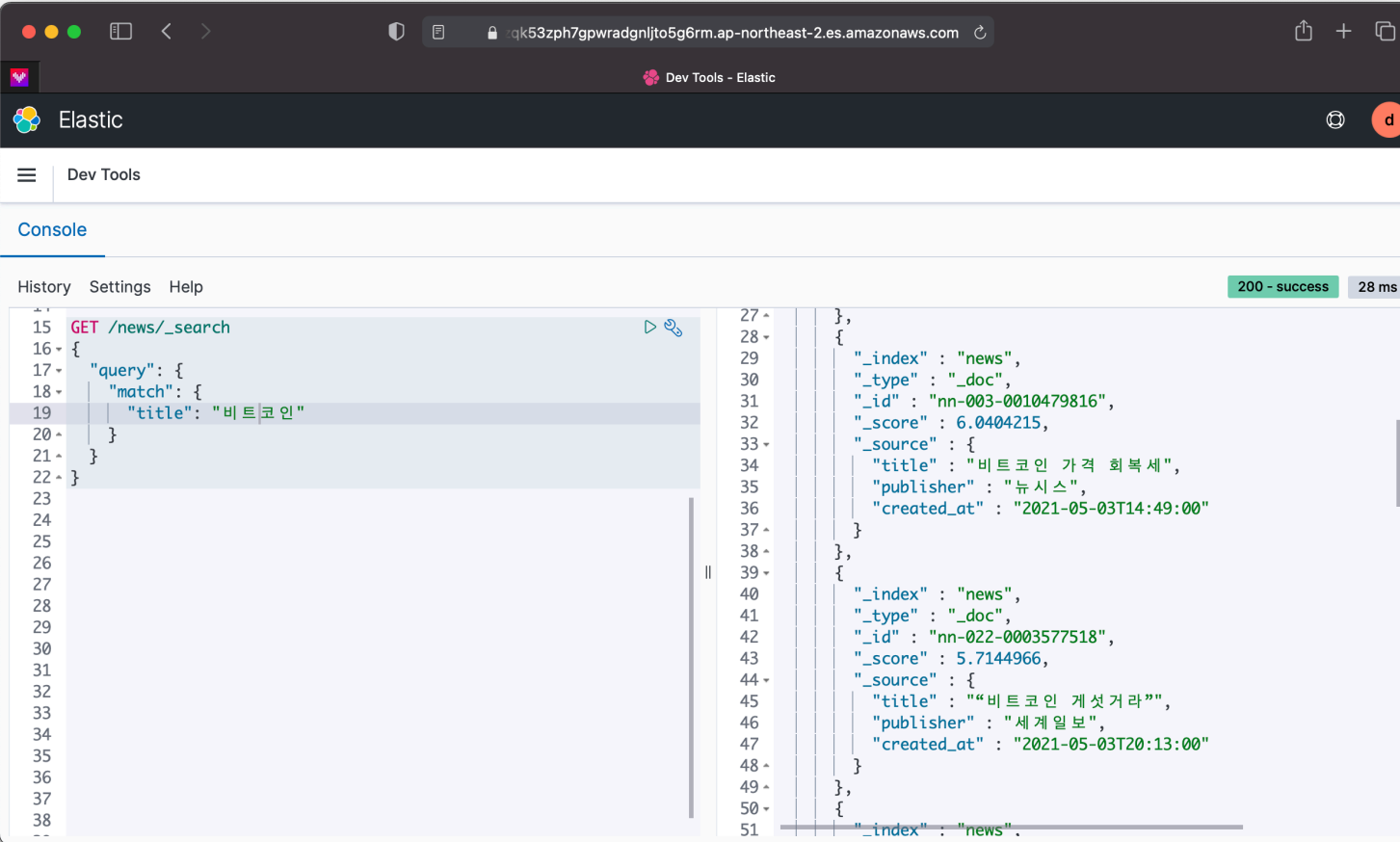
```
15 GET /news/_search
16 {
17   "query": {
18     "match": {
19       "title": "부동산"
20     }
21   }
22 }

23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
```

```
15 "max_score" : 5.751982,
16 "hits" : [
17   {
18     "_index" : "news",
19     "_type" : "_doc",
20     "_id" : "nn-018-0004917849",
21     "_score" : 5.751982,
22     "_source" : {
23       "title" : "[독특한 부동산]다운 계약 했다간 더 손해본다",
24       "publisher" : "이데일리",
25       "created_at" : "2021-05-01T08:24:00"
26     }
27   },
28   {
29     "_index" : "news",
30     "_type" : "_doc",
31     "_id" : "nn-421-0005325034",
32     "_score" : 5.44548,
33     "_source" : {
34       "title" : "[주부의세계] \"부동산 공급대책, 집값 상승만 부
35       "publisher" : "뉴스1",
36       "created_at" : "2021-05-01T10:00:00"
37     }
38   },
39 ]
```

200 - success 46 ms

ElasticSearch: 비트코인 뉴스 검색



The screenshot shows the Elastic Dev Tools interface in a web browser. The address bar displays the URL `qk53zph7gpwradgnljto5g6rm.ap-northeast-2.es.amazonaws.com`. The page title is "Elastic". The "Dev Tools" tab is active, and the "Console" sub-tab is selected. The console shows a successful HTTP GET request to `/news/_search` with a status of "200 - success" and a response time of "28 ms".

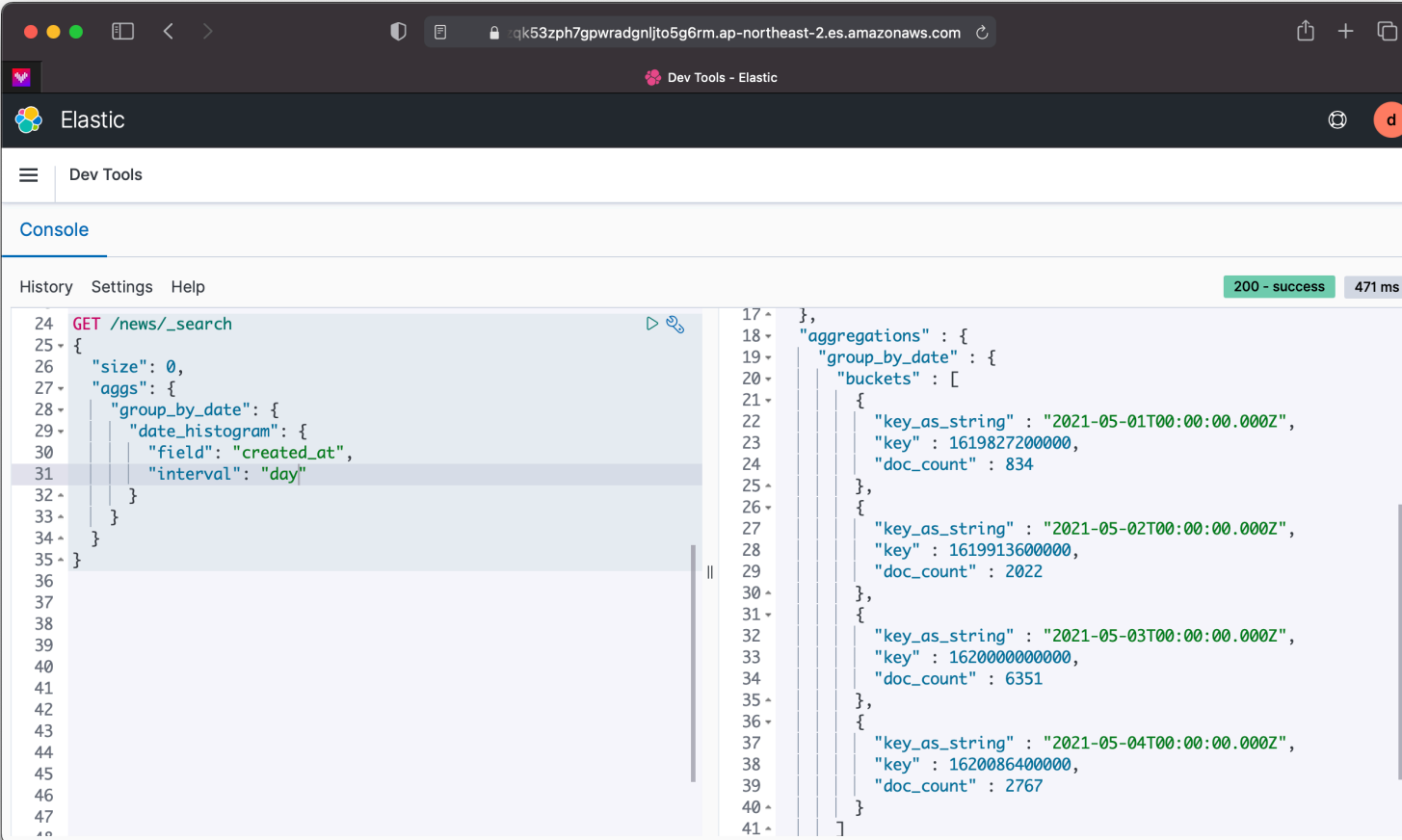
The request body (lines 15-22) is a JSON object:

```
15 GET /news/_search
16 {
17   "query": {
18     "match": {
19       "title": "비트코인"
20     }
21   }
22 }
```

The response body (lines 27-51) is a JSON array of search results:

```
27 {
28   {
29     "_index": "news",
30     "_type": "_doc",
31     "_id": "nn-003-0010479816",
32     "_score": 6.0404215,
33     "_source": {
34       "title": "비트코인 가격 회복세",
35       "publisher": "뉴시스",
36       "created_at": "2021-05-03T14:49:00"
37     }
38   },
39   {
40     "_index": "news",
41     "_type": "_doc",
42     "_id": "nn-022-0003577518",
43     "_score": 5.7144966,
44     "_source": {
45       "title": "“비트코인 게섯 거라”",
46       "publisher": "세계일보",
47       "created_at": "2021-05-03T20:13:00"
48     }
49   },
50   {
51     "_index": "news",
```

ElasticSearch: 날짜별로 입력된 뉴스 기사 갯수 세기



The screenshot shows the Elastic Dev Tools interface. The console displays a successful GET request to `/news/_search` with a date histogram aggregation. The response shows the number of documents for each day in May 2021.

```
24 GET /news/_search
25 {
26   "size": 0,
27   "aggs": {
28     "group_by_date": {
29       "date_histogram": {
30         "field": "created_at",
31         "interval": "day"
32       }
33     }
34   }
35 }
```

```
17 },
18 "aggregations" : {
19   "group_by_date" : {
20     "buckets" : [
21       {
22         "key_as_string" : "2021-05-01T00:00:00.000Z",
23         "key" : 1619827200000,
24         "doc_count" : 834
25       },
26       {
27         "key_as_string" : "2021-05-02T00:00:00.000Z",
28         "key" : 1619913600000,
29         "doc_count" : 2022
30       },
31       {
32         "key_as_string" : "2021-05-03T00:00:00.000Z",
33         "key" : 1620000000000,
34         "doc_count" : 6351
35       },
36       {
37         "key_as_string" : "2021-05-04T00:00:00.000Z",
38         "key" : 1620086400000,
39         "doc_count" : 2767
40       }
41     ]
42   }
43 }
```

200 - success 471 ms