Συστήματα Ανάκτησης Πληροφοριών

Σπανίδης Λουκάς

1) **XML to JSON**
   Αρχικά έπρεπε να τροποποιήσουμε τα αρχεία XML σε String έτσι ώστε να ενώσουμε σε μία καινούργια μεταβλητή **text** τις ήδη υπάρχουσες **title** και **objective**. Έπειτα, έχοντας σε αυτή την μορφή τα αρχεία μας, τα μετατρέπουμε σε JSON. Τέλος, γράφουμε όλα αυτά τα αρχεία σε ένα file βάζοντας πριν από το κάθε αρχείο ένα index ώστε να μπορεί το elasticsearch να ξεχωρίζει τα διαφορετικά αρχεία.

2) **Δημιουργία ευρετηρίου**
   Έπειτα έπρεπε να φτιάξουμε το index.

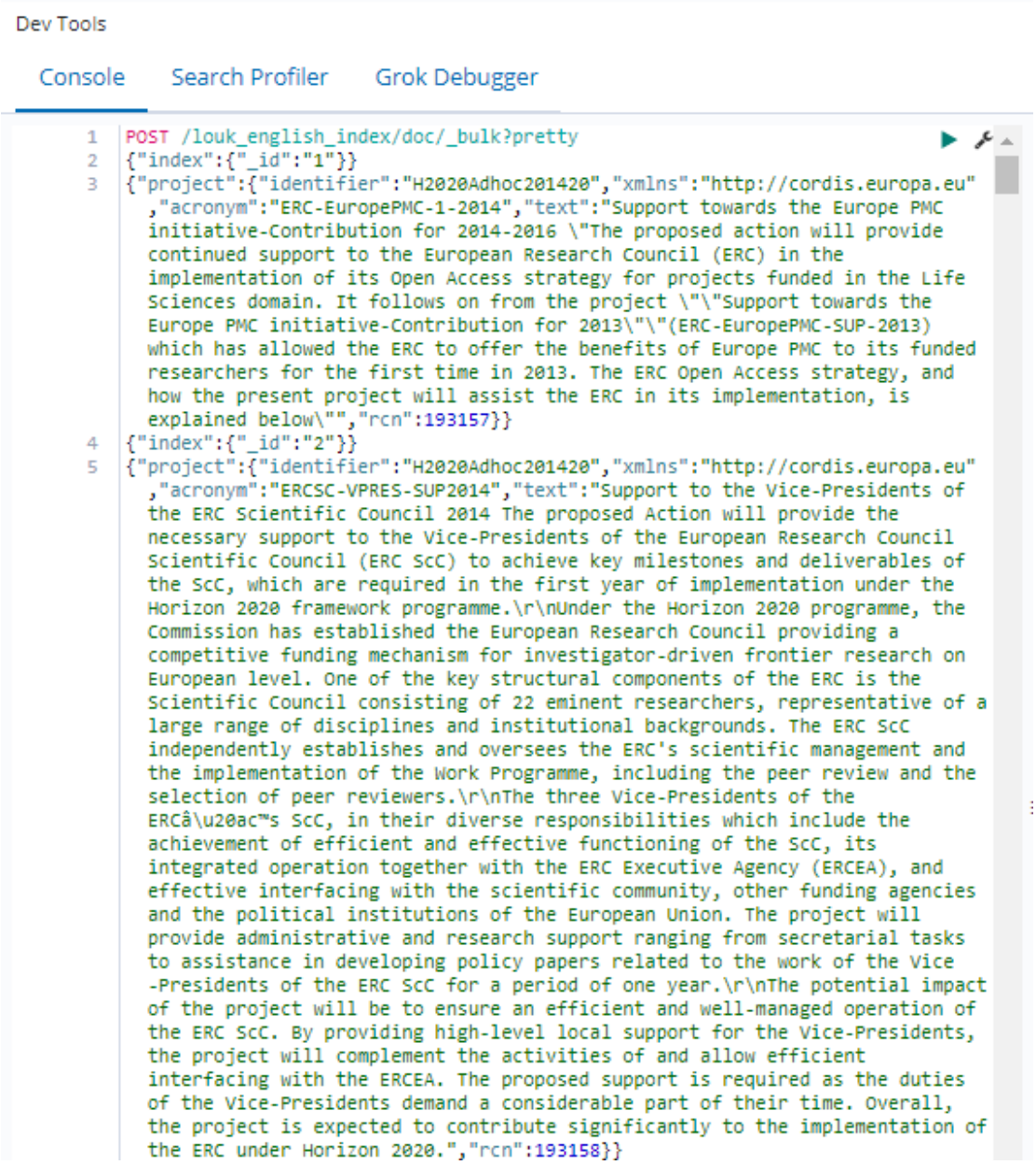
```
PUT /louk_english_index
{
  "settings": {
    "similarity": {
          "bm25-inverse-zero": {
              "type": "BM25",
              "b": 0
          }
        },
    "analysis": {
      "filter": {
        "english_stop": {
          "type":       "stop",
          "stopwords":  "_english_"
        },
        "english_keywords": {
          "type":       "keyword_marker",
          "keywords":   ["example"]
        },
        "english_stemmer": {
          "type":       "stemmer",
          "language":   "english"
        },
        "english_possessive_stemmer": {
          "type":       "stemmer",
          "language":   "possessive_english"
        }
      },
      "analyzer": {
        "rebuilt_english": {
          "tokenizer":  "standard",
          "filter": [
            "english_possessive_stemmer",
            "lowercase",
            "english_stop",
            "english_keywords",
            "english_stemmer"
          ]
        }
      }
    }
  }
}
```

Χρησιμοποίησα τον English Analyzer καθώς και βάρος BM25, όπως φαίνεται και στο παραπάνω screenshot.

3) **Upload data**

Αφού έχουμε δημιουργήσει το index είμαστε έτοιμοι να ανεβάσουμε μαζικά τα αρχεία με την εντολή **POST /louk_english_index/doc/_bulk?pretty**. Λόγου του μεγάλου αρχείου έπρεπε πριν το τρέξουμε στην κονσόλα του kibana να αλλάξουμε στο αρχείο kibana.yml το maxpayload σε μεγαλύτερο αριθμό από τον ήδη υπάρχον. Φορτώνουμε και τα 18316 αρχεία.

Ενδεικτικό screenshot:



4) **Χρήση queries**

Αφού έχουμε φορτώσει τα αρχεία θα χρησιμοποιήσουμε 10 queries για την ανάκτηση κειμένων. Δεν παίρνουμε το 1^{ο} κείμενο ως απάντηση διότι είναι ο εαυτός του, οπότε ξεκινάμε από το αμέσως επόμενο. Αυτό γίνετε με το from: 1 και παίρνουμε τα επόμενα 20 (size:20).

Μορφή των query:

```
GET louk_english_index/_search
{
  "from": 1,
  "size":20,
  "query":{
    "query_string": {
      "query":"Query που θέλουμε να τρέξουμε"
      }
  }
}
```

Ενδεικτικό παράδειγμα:

```
GET louk_english_index/_search
{
  "from": 1,
  "size":20,
  "query":{
    "query_string": {
      "query":"netCommons network infrastructure as commons Communication and
        information distribution are key components of a modern society  The advent
        of the Internet has been often invoked as a remedy for their democratization
         The truth shows a different picture  the digital divide is widening the gap
        between those who can access and take advantage of the new systems  and
        those who remain disconnected  nThe Internet's unsustainability  coupled
        with the lack of awareness of the actual complexity of the Internet's
        organisation  means that users are mostly unaware of the potentials of
        digital interaction and  most of all  of the possibility to have a bottom up
         democratic  communal organisation of it  netCommons studies an emerging
        trend  community based networking and services that can offer a complement
        to the global Internet's model  Community networks not only offer to
        citizens the access to a neutral  network infrastructure  which naturally
        increases the transparency of data flow  storage and use  but they also
        represent the archetype of networked collective cooperation and action
        Community networks are complex systems that require multiple skills to
        thrive  technical  legal  socio economic  and more  They face many
        challenges and they need means and tools to grow and produce a higher impact
        on society  netCommons follows a dual approach to achieve its goals  1  It
        works at a local level  mingling with the communities to gather relevant
        information  elaborate it  and return them advanced tools to grow and thrive
        ; 2  Starting from the hands on experience and work  it contributes to
        Internet Science by abstracting concepts  it studies and offers solutions
        and interpretations  that can be used by legislators and decision makers  to
        build global awareness of the importance of sustainability  participation
        co operation  on line information  democracy  peer production  and how to
        foster the development of community networks to generate socio economical
        opportunities based on this paradigm of Internet Science"
      }
  }
}
```

Ενδεικτική απάντηση:

```
1  {
2    "took" : 49,
3    "timed_out" : false,
4    "_shards" : {
5      "total" : 5,
6      "successful" : 5,
7      "skipped" : 0,
8      "failed" : 0
9    },
10   "hits" : {
11     "total" : 18314,
12     "max_score" : 759.30566,
13     "hits" : [
14       {
15         "_index" : "louk_english_index",
16         "_type" : "doc",
17         "_id" : "6009",
18         "_score" : 156.2146,
19         "_source" : {
20           "project" : {
21             "identifier" : "H2020ICT2015",
22             "xmlns" : "http://cordis.europa.eu",
23             "acronym" : "MAZI",
24             "text" : "A DIY networking toolkit for location-based collective awareness Do-It-Yourself networking refers to a conceptual approach to the use of low-cost hardware and
                wireless technology in deploying local communication networks that can operate independently from the Internet, owned and controlled by local actors. MAZI means "together"
                in Greek and MAZI [http://mazizone.eu] invests in this paradigm of technology-supported networking, as a means to bring closer together those living in physical proximity.
                Through an experienced interdisciplinary consortium, MAZI delivers a DIY networking toolkit that offers tools and guidelines for the easy deployment and customization of
                local networks and services. MAZI toolkit is designed to take advantage of particular characteristics of DIY networking: the de facto physical proximity between those
                connected; the increased privacy and autonomy; and the inclusive access. Such characteristics are used to promote information exchanges that can develop the location-based
                collective awareness, as a basis for fostering social cohesion, conviviality, knowledge sharing, and sustainable living. To achieve this objective, MAZI brings together
                partners from different disciplines: computer networks, urban planning and interdisciplinary studies, human-computer interaction, community informatics, and design
                research. These academic partners will collaborate closely with four community partners to ensure that MAZI toolkit benefits from the grounded experience of citizen
                engagement. MAZI draws from the diverse mix of competencies of its consortium to develop a transdisciplinary research framework, which will guide a series of long-term
                pilot studies in a range of environments, and enhanced by cross-fertilization events. The main goal of this process, and measure of success, is establishing DIY networking
                as a mainstream technology for enabling the development of collective awareness between those in physical proximity, and the development of surrounding research and
                theorizing of this approach.",
25             "rcn" : 199849
26           }
27         }
28       },
```


5) Αξιολόγηση με την χρήση του trec-eval

Αφού πάρουμε όλα τα αποτελέσματα (10 queries από 20 κείμενα), τα μετατρέπουμε σε ιδική μορφή για να τα εισάγουμε στο trec-eval.

Ενδεικτικό παράδειγμα:

```
1  Q01 Q0 193373 1 264.8219 STANDARD
2  Q01 Q0 205685 1 223.77287 STANDARD
3  Q01 Q0 193375 1 211.76877 STANDARD
4  Q01 Q0 193353 1 209.31947 STANDARD
5  Q01 Q0 210137 1 204.38342 STANDARD
6  Q01 Q0 193386 1 204.27719 STANDARD
7  Q01 Q0 206230 1 203.4013 STANDARD
8  Q01 Q0 211970 1 202.71054 STANDARD
9  Q01 Q0 194660 1 202.63506 STANDARD
10 Q01 Q0 211346 1 197.81853 STANDARD
11 Q01 Q0 193715 1 192.90141 STANDARD
12 Q01 Q0 206824 1 186.12343 STANDARD
13 Q01 Q0 202703 1 182.2309 STANDARD
14 Q01 Q0 193402 1 181.93422 STANDARD
15 Q01 Q0 206228 1 181.06203 STANDARD
16 Q01 Q0 211697 1 181.01686 STANDARD
17 Q01 Q0 194067 1 179.077 STANDARD
18 Q01 Q0 213250 1 174.928 STANDARD
19 Q01 Q0 205643 1 173.65123 STANDARD
20 Q01 Q0 198900 1 173.06693 STANDARD
21 Q02 Q0 210232 1 222.61356 STANDARD
22 Q02 Q0 194301 1 147.37918 STANDARD
23 Q02 Q0 206010 1 121.838425 STANDARD
24 Q02 Q0 212411 1 119.49203 STANDARD
25 Q02 Q0 198340 1 117.53799 STANDARD
26 Q02 Q0 211729 1 116.89759 STANDARD
27 Q02 Q0 206417 1 116.74827 STANDARD
28 Q02 Q0 212231 1 115.60331 STANDARD
29 Q02 Q0 193380 1 115.45932 STANDARD
30 Q02 Q0 214253 1 114.26923 STANDARD
31 Q02 Q0 213081 1 113.48167 STANDARD
32 Q02 Q0 204192 1 113.466576 STANDARD
33 Q02 Q0 200475 1 113.41677 STANDARD
34 Q02 Q0 207482 1 112.23301 STANDARD
35 Q02 Q0 197947 1 111.21204 STANDARD
36 Q02 Q0 198301 1 110.86718 STANDARD
37 Q02 Q0 194185 1 110.32379 STANDARD
38 Q02 Q0 207805 1 110.27912 STANDARD
39 Q02 Q0 211074 1 109.63233 STANDARD
40 Q02 Q0 194872 1 109.523796 STANDARD
41 Q03 Q0 204772 1 208.37727 STANDARD
42 Q03 Q0 205420 1 190.5277 STANDARD
43 Q03 Q0 214637 1 188.51483 STANDARD
44 Q03 Q0 211673 1 186.20404 STANDARD
45 Q03 Q0 209715 1 178.57896 STANDARD
46 Q03 Q0 193825 1 174.25363 STANDARD
```

Στην συνέχεια τρέχουμε την εντολή :

```
>trec_eval -q qrels.test result_trec.test
```

Όπου qrels.test οι σωστές απαντήσεις και result_trec οι δικιές μου.
Η απάντηση σε αυτή την εντολή βρήσκεται μέσα στο αρχεία την εργασίας με όνομα: trec_eval_result.