

# 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.



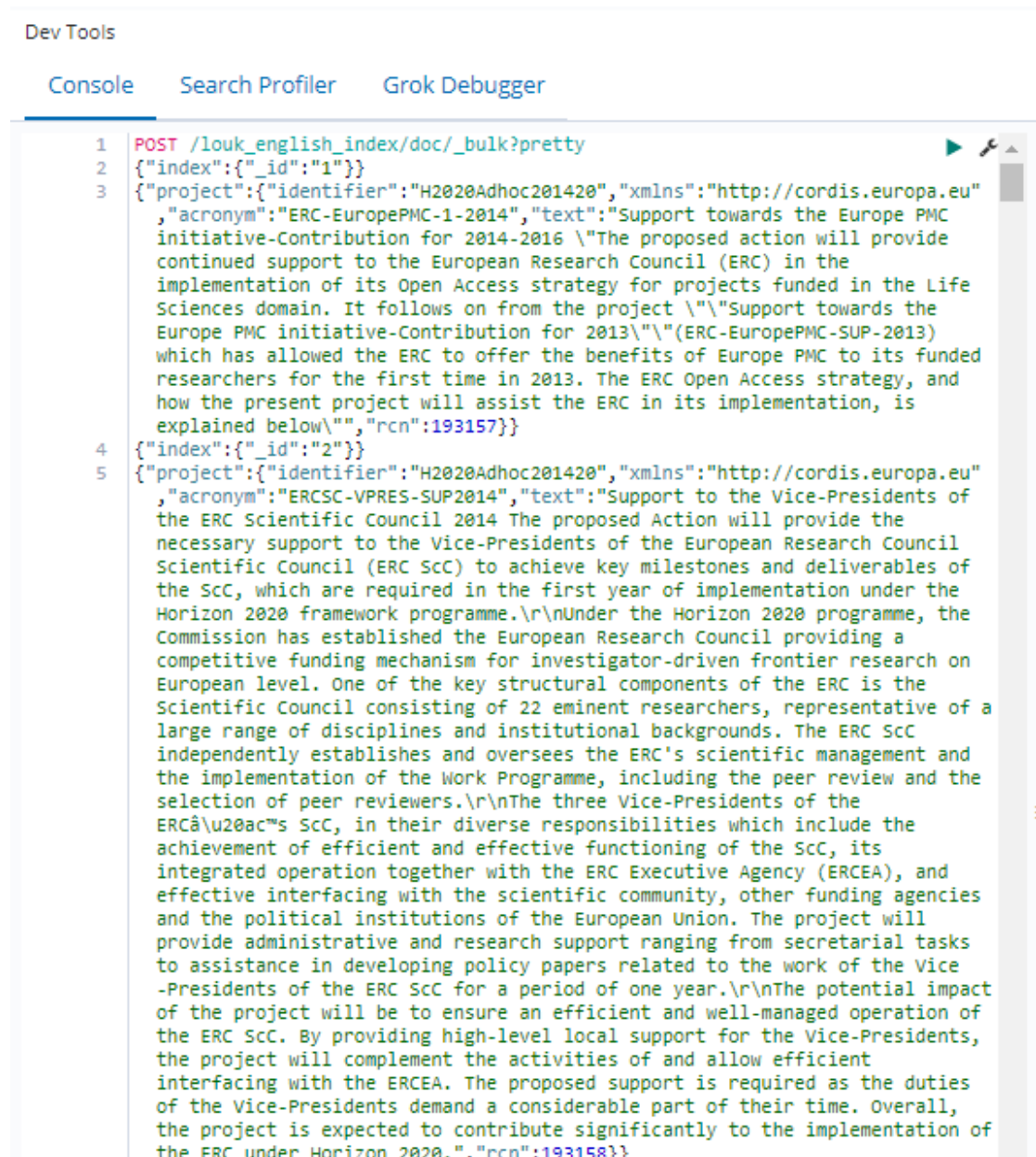
```
1 PUT /louk_english_index
2 {
3   "settings": {
4     "similarity": {
5       "bm25-inverse-zero": {
6         "type": "BM25",
7         "b": 0
8       }
9     },
10    "analysis": {
11      "filter": {
12        "english_stop": {
13          "type": "stop",
14          "stopwords": "_english_"
15        },
16        "english_keywords": {
17          "type": "keyword_marker",
18          "keywords": ["example"]
19        },
20        "english_stemmer": {
21          "type": "stemmer",
22          "language": "english"
23        },
24        "english_possessive_stemmer": {
25          "type": "stemmer",
26          "language": "possessive_english"
27        }
28      },
29      "analyzer": {
30        "rebuilt_english": {
31          "tokenizer": "standard",
32          "filter": [
33            "english_possessive_stemmer",
34            "lowercase",
35            "english_stop",
36            "english_keywords",
37            "english_stemmer"
38          ]
39        }
40      }
41    }
42  }
43 }
```

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:



```
Dev Tools

Console Search Profiler Grok Debugger

1 POST /louk_english_index/doc/_bulk?pretty
2 {"index":{"_id":"1"}}
3 {"project":{"identifier":"H2020Adhoc201420","xmlns":"http://cordis.europa.eu",
  "acronym":"ERC-EuropePMC-1-2014","text":"Support towards the Europe PMC
  initiative-Contribution for 2014-2016 \"The proposed action will provide
  continued support to the European Research Council (ERC) in the
  implementation of its Open Access strategy for projects funded in the Life
  Sciences domain. It follows on from the project \"\"Support towards the
  Europe PMC initiative-Contribution for 2013\"\"(ERC-EuropePMC-SUP-2013)
  which has allowed the ERC to offer the benefits of Europe PMC to its funded
  researchers for the first time in 2013. The ERC Open Access strategy, and
  how the present project will assist the ERC in its implementation, is
  explained below\"\", \"rcn\":\"193157\"}}
4 {"index":{"_id":"2"}}
5 {"project":{"identifier":"H2020Adhoc201420","xmlns":"http://cordis.europa.eu",
  "acronym":"ERCSC-VPRES-SUP2014","text":"Support to the Vice-Presidents of
  the ERC Scientific Council 2014 The proposed Action will provide the
  necessary support to the Vice-Presidents of the European Research Council
  Scientific Council (ERC ScC) to achieve key milestones and deliverables of
  the ScC, which are required in the first year of implementation under the
  Horizon 2020 framework programme.\\r\\nUnder the Horizon 2020 programme, the
  Commission has established the European Research Council providing a
  competitive funding mechanism for investigator-driven frontier research on
  European level. One of the key structural components of the ERC is the
  Scientific Council consisting of 22 eminent researchers, representative of a
  large range of disciplines and institutional backgrounds. The ERC ScC
  independently establishes and oversees the ERC's scientific management and
  the implementation of the Work Programme, including the peer review and the
  selection of peer reviewers.\\r\\nThe three Vice-Presidents of the
  ERC's ScC, in their diverse responsibilities which include the
  achievement of efficient and effective functioning of the ScC, its
  integrated operation together with the ERC Executive Agency (ERCEA), and
  effective interfacing with the scientific community, other funding agencies
  and the political institutions of the European Union. The project will
  provide administrative and research support ranging from secretarial tasks
  to assistance in developing policy papers related to the work of the Vice-
  Presidents of the ERC ScC for a period of one year.\\r\\nThe potential impact
  of the project will be to ensure an efficient and well-managed operation of
  the ERC ScC. By providing high-level local support for the Vice-Presidents,
  the project will complement the activities of and allow efficient
  interfacing with the ERCEA. The proposed support is required as the duties
  of the Vice-Presidents demand a considerable part of their time. Overall,
  the project is expected to contribute significantly to the implementation of
  the ERC under Horizon 2020.", \"rcn\":\"193158\"}}
```

### 4) Χρήση queries

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

Μορφή των query:

```
1 GET louk_english_index/_search
2 {
3   "from": 1,
4   "size": 20,
5   "query": {
6     "query_string": {
7       "query": "Query που θέλουμε να τρέξουμε"
8     }
9   }
10 }
```

Ενδεικτικό παράδειγμα:

```
1 GET louk_english_index/_search
2 {
3   "from": 1,
4   "size": 20,
5   "query": {
6     "query_string": {
7       "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"
8     }
9   }
10 }
```

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" : 10314,
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 DIV 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://mailzone.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 DIV 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 DIV 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 DIV 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    ]
29  }
}

```

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

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 βρήσκειται μέσα στο αρχείο την εργασίας με όνομα: trec\_eval\_result.