

MAC - Labo 3 : Indexing and Search with Elasticsearch

Olivier D'Ancona & Hugo Huart & Nelson Jeanrenaud

Contents

2.2 Indexing	2
D.1	3
D.2	4
D.3	5
D.4	5
D.5	5
2.3 Reading Index	6
D.6	6
D.7	6
2.4 Using different Analyzers	7
D.8	7
D.9	12
D.10	12
D.11	13
2.5 Searching	13
D.12	13
D.13	14
2.6 Custom similarity	15
D.14	15
D.15	16
D.16	20

2.2 Indexing

Using the following pipeline:

PUT _ingest/pipeline/my_pipeline

```
{
  "processors": [
    {
      "csv": {
        "field": "_row",
        "target_fields": [
          "id",
          "author",
          "title",
          "date",
          "summary"
        ],
        "separator": "\t",
        "quote": "§"
      }
    },
    {
      "split": {
        "field": "author",
        "separator": ";",
        "ignore_missing": true
      }
    },
    {
      "remove": {
        "field": "_row"
      }
    }
  ]
}
```

D.1

API requests to create `cacm_standard`:

Mappings:

PUT `/cacm_standard`

```
{
  "mappings": {
    "properties": {
      "author": {
        "type": "keyword"
      },
      "date": {
        "type": "date"
      },
      "id": {
        "type": "unsigned_long"
      },
      "summary": {
        "type": "text",
        "fielddata": true
      },
      "title": {
        "type": "text",
        "fielddata": true
      }
    }
  }
}
```

Reindex:

POST `_reindex`

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_standard",
    "pipeline": "my_pipeline"
  }
}
```

D.2

API requests to create `cacm_termvector`:

Mappings:

PUT /cacm_termvector

```
{
  "mappings": {
    "properties": {
      "author": {
        "type": "keyword"
      },
      "date": {
        "type": "date"
      },
      "id": {
        "type": "unsigned_long"
      },
      "summary": {
        "type": "text",
        "term_vector": "with_positions"
      },
      "title": {
        "type": "text"
      }
    }
  }
}
```

Reindex:

POST _reindex

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_termvector",
    "pipeline": "my_pipeline"
  }
}
```

D.3

API request to query a term vector:

```
GET /cacm_termvector/_termvectors/gNa1ZYAB7VfE5TWZZFs7
```

gNa1ZYAB7VfE5TWZZFs7 being the ID of a document that has a `summary` field.

D.4

The official documentation of Elasticsearch describes a term vector as the following:

Term vectors contain information about the terms produced by the analysis process, including:

- *A list of terms.*
- *The position (or order) of each term.*
- *The start and end character offsets mapping the term to its origin in the original string.*
- *Payloads (if they are available) — user-defined binary data associated with each term position.*

D.5

Sizes of the indexes:

- **cacm_raw : 1.34 MB**
- **cacm_standard : 1.48 MB**
- **cacm_termvector : 2.07 MB**

2.3 Reading Index

D.6

Using the following request, we observe that **Thacher Jr., H. C.** is the author with the highest number of publications. He has **38** publications.

Request:

```
GET /cacm_standard/_search
{
  "aggs": {
    "genres": {
      "terms": {
        "field": "author",
        "size": 1
      }
    }
  }
}
```

D.7

Using the following request, we observe that the top 10 terms are:

1. *of*
2. *algorithm*
3. *a*
4. *for*
5. *the*
6. *and*
7. *in*
8. *on*
9. *an*
10. *computer*

Request:

```
GET /cacm_standard/_search
{
  "aggs": {
    "genres": {
      "terms": {
        "field": "title",
        "size": 10
      }
    }
  }
}
```

2.4 Using different Analyzers

D.8

The following requests create indexes with the required analyzers.

whitespace analyzer

PUT /cacm_standard_whitespace

```
{
  "settings": {
    "analysis": {
      "analyzer": "whitespace"
    }
  },
  "mappings": {
    "properties": {
      "id":{"type": "unsigned_long"},
      "author": {"type": "keyword"},
      "title":{"type": "text", "fielddata": true},
      "date":{"type": "date"},
      "summary":{"analyzer" : "whitespace", "type": "text", "fielddata" : true}
    }
  }
}
```

POST _reindex

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_standard_whitespace",
    "pipeline": "my_pipeline"
  }
}
```

english analyzer

PUT /cacm_standard_english

```
{
  "settings": {
    "analysis": {
      "analyzer": "english"
    }
  },
  "mappings": {
    "properties": {
      "id": { "type": "unsigned_long" },
      "author": { "type": "keyword" },
      "title": { "type": "text", "fielddata": true },
      "date": { "type": "date" },
      "summary": { "analyzer": "english", "type": "text", "fielddata": true }
    }
  }
}
```

POST _reindex

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_standard_english",
    "pipeline": "my_pipeline"
  }
}
```


standard analyzer with shingles of size 1 and 2

PUT /cacm_standard_myalyzer1

```
{
  "settings": {
    "analysis": {
      "analyzer": {
        "my_analyzer1": {
          "type": "custom",
          "tokenizer": "standard",
          "filter": [
            "lowercase",
            "custom_shingle"
          ]
        }
      },
      "filter": {
        "custom_shingle": {
          "type": "shingle",
          "max_shingle_size": 2
        }
      }
    }
  },
  "mappings": {
    "properties": {
      "id": { "type": "unsigned_long" },
      "author": { "type": "keyword" },
      "title": { "type": "text", "fielddata": true },
      "date": { "type": "date" },
      "summary": { "analyzer": "my_analyzer1", "type": "text", "fielddata": true }
    }
  }
}
```

POST _reindex

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_standard_myalyzer1",
    "pipeline": "my_pipeline"
  }
}
```

standard analyzer with shingles of size 3

PUT /cacm_standard_myalyzer2

```
{
  "settings": {
    "analysis": {
      "analyzer": {
        "my_analyzer2": {
          "type": "custom",
          "tokenizer": "standard",
          "filter": [
            "lowercase",
            "custom_shingle"
          ]
        }
      },
      "filter": {
        "custom_shingle": {
          "type": "shingle",
          "min_shingle_size": 3,
          "max_shingle_size": 3,
          "output_unigrams": false
        }
      }
    },
    "mappings": {
      "properties": {
        "id": { "type": "unsigned_long" },
        "author": { "type": "keyword" },
        "title": { "type": "text", "fielddata": true },
        "date": { "type": "date" },
        "summary": { "analyzer": "my_analyzer2", "type": "text", "fielddata": true }
      }
    }
  }
}
```

POST _reindex

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_standard_myalyzer2",
    "pipeline": "my_pipeline"
  }
}
```

stop analyzer

PUT /cacm_standard_stopwords

```
{
  "settings": {
    "analysis": {
      "analyzer": {
        "stopwords": {
          "tokenizer": "lowercase",
          "filter": [ "custom_stopwords" ]
        }
      },
      "filter" : {
        "custom_stopwords" : {
          "type" : "stop",
          "stopwords_path" : "data/common_words.txt"
        }
      }
    },
    "mappings": {
      "properties": {
        "id": { "type": "unsigned_long" },
        "author": { "type": "keyword" },
        "title": { "type": "text", "fielddata": true },
        "date": { "type": "date" },
        "summary": { "analyzer" : "stopwords", "type": "text", "fielddata": true }
      }
    }
  }
}
```

POST _reindex

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_standard_stopwords",
    "pipeline": "my_pipeline"
  }
}
```

D.9

Explanation of the analyzers, according to the Elasticsearch documentation:

- **whitespace** : Breaks text into terms whenever a whitespace is encountered.
- **english** : Targeted for English text. It features relevant stop words, plural to singular conversion and other similar language-specific filters.
- **standard** with shingles of size 1 and 2 : Produce shingles (or word n-gram) up to a size of two:
The text "I Love MAC" would produce ["I", "I Love", "Love", "Love MAC", "MAC"].
- **standard** with shingles of size 3 only : Produce shingles (or word n-gram) of size 3:
The text "I Love MAC" would produce ["I", "I Love MAC", "Love", "MAC"].
- **stop** : Uses a list of words as stop words that will be removed from the the requested text.

D.10

Using the Index stats and search APIs with the following types of requests:

```
GET /${INDEX_NAME}/_stats
```

```
GET /${INDEX_NAME}/_search
```

The results are:

Analyzer type:	whitespace	english	standard shingles 1-2	standard shingles 3	stop
a)	3'202 docs	3'202 docs	3'202 docs	3'202 docs	3'202 docs
b)	103'275 terms	72'298 terms	237'189 terms	144'518 terms	59'988 terms
c)	<i>of</i>	<i>which</i>	<i>the</i>	<i>in this paper</i>	<i>computer</i>
	<i>the</i>	<i>us</i>	<i>of</i>	<i>the use of</i>	<i>system</i>
	<i>is</i>	<i>comput</i>	<i>a</i>	<i>the number of</i>	<i>paper</i>
	<i>and</i>	<i>program</i>	<i>is</i>	<i>it is shown</i>	<i>presented</i>
	<i>a</i>	<i>system</i>	<i>and</i>	<i>a set of</i>	<i>time</i>
	<i>to</i>	<i>present</i>	<i>to</i>	<i>in terms of</i>	<i>program</i>
	<i>in</i>	<i>describ</i>	<i>in</i>	<i>the problem of</i>	<i>data</i>
	<i>for</i>	<i>paper</i>	<i>for</i>	<i>is shown that</i>	<i>method</i>
	<i>The</i>	<i>can</i>	<i>are</i>	<i>a number of</i>	<i>algorithm</i>
	<i>are</i>	<i>gener</i>	<i>of the</i>	<i>as well as</i>	<i>discussed</i>
d)	13'542'719 B	19'859'606 B	2'597'942 B	3'103'118 B	2'833'980 B
e)	350 ms	250 ms	340 ms	300 ms	340 ms

Note: the timings of point e) may vary significantly by 20 to 50ms

D.11

Several statements can be made regarding the previous results, here are our 3 concluding ones:

1. All the indexes have the same number of documents, the presentation of the documents is not altered in that regard.
2. The shingle-based indexes have the most terms. This make sense because they are the only indexes that add new terms in summary (the shingles).
3. The custom stop words provided in **stop** are more restrictive than the default ones of **english**. This is confirmed by the lower number of terms in the **stop** index.

2.5 Searching

D.12

Here are the API requests of the relevant queries:

1.

```
GET /cacm_standard_english/_search
{
  "query" : {
    "query_string" : {
      "query" : "Information Retrieval",
      "default_field": "summary"
    }
  },
  "_source": "id"
}
```

2.

```
GET /cacm_standard_english/_search
{
  "query" : {
    "query_string" : {
      "query" : "Information AND Retrieval",
      "default_field": "summary"
    }
  },
  "_source": "id"
}
```

3.

```
GET /cacm_standard_english/_search
{
  "query" : {
    "query_string" : {
      "query" : "Retrieval AND NOT Database",
      "default_field": "summary"
    }
  },
  "_source": "id"
}
```

Note: The query was simplified using boolean algebra.

4.

```
GET /cacm_standard_english/_search
{
  "query" : {
    "query_string" : {
      "query" : "Info*",
      "default_field": "summary"
    }
  },
  "_source": "id"
}
```

5.

```
GET /cacm_standard_english/_search
{
  "query" : {
    "query_string" : {
      "query" : "\"Information Retrieval\"~5",
      "default_field": "summary"
    }
  },
  "_source": "id"
}
```

D.13

Here are the results of the previous API requests:

1. **240** hits
2. **36** hits
3. **69** hits
4. **205** hits
5. **30** hits

2.6 Custom similarity

D.14

The new index with the custom scoring method is created with the following API request:

PUT /cacm_standard_score

```
{
  "settings": {
    "number_of_shards": 1,
    "similarity": {
      "scripted_tfidf": {
        "type": "scripted",
        "script": {
          "source": "double tf = 1 + Math.log(doc.freq);
                     double idf = Math.log((field.docCount) / (term.docFreq + 1.0)) + 1.0;
                     double norm = 1;
                     return tf * idf * norm * query.boost;"
        }
      }
    }
  },
  "mappings": {
    "properties": {
      "author": { "type": "keyword" },
      "date": { "type": "date" },
      "id": { "type": "unsigned_long" },
      "summary": { "similarity": "scripted_tfidf", "type": "text", "fielddata": true },
      "title": { "type": "text", "fielddata": true }
    }
  }
}
```

POST _reindex

```
{
  "source": {
    "index": "cacm_raw"
  },
  "dest": {
    "index": "cacm_standard_score",
    "pipeline": "my_pipeline"
  }
}
```

D.15

Top 10 results of the `compiler` program query with and without the custom scoring system:

Default scoring:

```
"hits" : [  
  {  
    "_index" : "cacm_standard",  
    "_id" : "03jFZYAB9pJXcpxGibLi",  
    "_score" : 7.0958896,  
    "_source" : {  
      "id" : "123"  
    }  
  },  
  {  
    "_index" : "cacm_standard",  
    "_id" : "ZnjFZYAB9pJXcpxGibTl",  
    "_score" : 7.0101233,  
    "_source" : {  
      "id" : "678"  
    }  
  },  
  {  
    "_index" : "cacm_standard",  
    "_id" : "NXjFZYAB9pJXcpxGir6n",  
    "_score" : 6.957088,  
    "_source" : {  
      "id" : "3189"  
    }  
  },  
  {  
    "_index" : "cacm_standard",  
    "_id" : "eXjFZYAB9pJXcpxGibfn",  
    "_score" : 6.926655,  
    "_source" : {  
      "id" : "1465"  
    }  
  },  
  {  
    "_index" : "cacm_standard",  
    "_id" : "i3jFZYAB9pJXcpxGibjn",  
    "_score" : 6.7396317,  
    "_source" : {  
      "id" : "1739"  
    }  
  },  
  {  
    "_index" : "cacm_standard",
```



```

    "_id" : "L3jFZYAB9pJXcpxGibjn",
    "_score" : 6.4866796,
    "_source" : {
      "id" : "1647"
    }
  },
  {
    "_index" : "cacm_standard",
    "_id" : "pnjFZYAB9pJXcpxGirum",
    "_score" : 6.4402065,
    "_source" : {
      "id" : "2534"
    }
  },
  {
    "_index" : "cacm_standard",
    "_id" : "N3jFZYAB9pJXcpxGirul",
    "_score" : 6.1307526,
    "_source" : {
      "id" : "2423"
    }
  },
  {
    "_index" : "cacm_standard",
    "_id" : "FnjFZYAB9pJXcpxGibTk",
    "_score" : 6.0692043,
    "_source" : {
      "id" : "598"
    }
  },
  {
    "_index" : "cacm_standard",
    "_id" : "f3jFZYAB9pJXcpxGibbm",
    "_score" : 5.9121494,
    "_source" : {
      "id" : "1215"
    }
  }
]

```

Custom scoring:

```
"hits" : [  
  {  
    "_index" : "cacm_standard_score",  
    "_id" : "pnjFZYAB9pJXcpxGirum",  
    "_score" : 14.1800995,  
    "_source" : {  
      "id" : "2534"  
    }  
  },  
  {  
    "_index" : "cacm_standard_score",  
    "_id" : "L3jFZYAB9pJXcpxGibjn",  
    "_score" : 12.428032,  
    "_source" : {  
      "id" : "1647"  
    }  
  },  
  {  
    "_index" : "cacm_standard_score",  
    "_id" : "i3jFZYAB9pJXcpxGibjn",  
    "_score" : 11.304348,  
    "_source" : {  
      "id" : "1739"  
    }  
  },  
  {  
    "_index" : "cacm_standard_score",  
    "_id" : "N3jFZYAB9pJXcpxGirul",  
    "_score" : 11.304348,  
    "_source" : {  
      "id" : "2423"  
    }  
  },  
  {  
    "_index" : "cacm_standard_score",  
    "_id" : "03jFZYAB9pJXcpxGibLi",  
    "_score" : 11.234488,  
    "_source" : {  
      "id" : "123"  
    }  
  },  
  {  
    "_index" : "cacm_standard_score",  
    "_id" : "ZnjFZYAB9pJXcpxGibTl",  
    "_score" : 11.234488,  
    "_source" : {  
      "id" : "678"  
    }  
  }  
]
```

```

    }
  },
  {
    "_index" : "cacm_standard_score",
    "_id" : "eXjFZYAB9pJXcpxGibfn",
    "_score" : 11.234488,
    "_source" : {
      "id" : "1465"
    }
  },
  {
    "_index" : "cacm_standard_score",
    "_id" : "f3jFZYAB9pJXcpxGibbm",
    "_score" : 9.900334,
    "_source" : {
      "id" : "1215"
    }
  },
  {
    "_index" : "cacm_standard_score",
    "_id" : "EXjFZYAB9pJXcpxGir2m",
    "_score" : 9.900334,
    "_source" : {
      "id" : "2897"
    }
  },
  {
    "_index" : "cacm_standard_score",
    "_id" : "xnjFZYAB9pJXcpxGibfn",
    "_score" : 9.552281,
    "_source" : {
      "id" : "1542"
    }
  }
]

```

D.16

Query with custom function score:

GET /cacm_standard/_search

```
{
  "query": {
    "function_score": {
      "query": {
        "query_string": {
          "query": "compiler program",
          "default_field": "summary"
        }
      },
      "linear": {
        "date": {
          "origin": "1970-01-01",
          "scale": "90d",
          "offset": "0d",
          "decay": 0.5
        }
      }
    }
  }
}
```