Elasticsearch

pesquisando e analisando seus dados

Quebrando textos com Analyzers

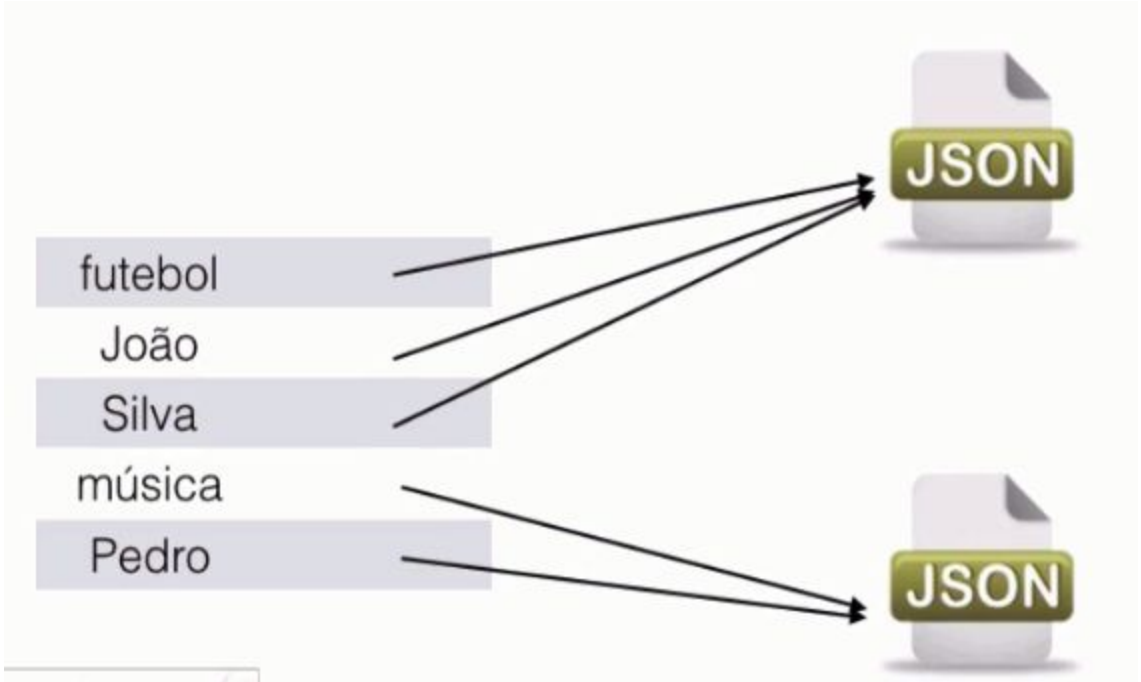
GET catalogo/_search?q=musica

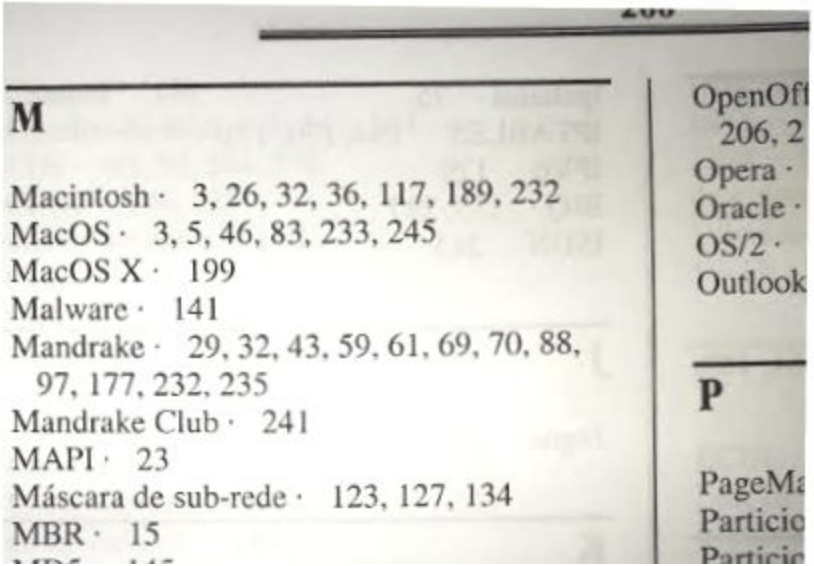

{
"took" : 477,
"timed_out" : false,
"_shards" : {
"total" : 1,
"successful" : 1,
"skipped" : 0,
"failed" : 0
},
"hits" : {
"total" : {
"value" : 0,
"relation" : "eq"
},
"max_score" : null,
"hits" : [ ]
}
}

GET catalogo/_search?q=música


"max_score" : 0.7411202,
"hits" : [
"_index" : "catalogo",
"_type" : "_doc",
"_id" : "eSIgcW4B0aSwEs1PmGSV",
"_score" : 0.7411202,
"_source" : {
"nome" : "João Silva",
"interesses" : [
"futebol",
"música",
"literatura"
],
"cidade" : "São Paulo",
"formação" : "Letras",
"estado" : "SP",
"país" : "Brasil"
}
termo
encontrado
com acento


},
"max_score" : 0.7411202,
"hits" : [
{
"_index" : "catalogo",
"_type" : "_doc",
"_id" : "eSIgcW4B0aSwEs1PmGSV",
"_score" : 0.7411202,
"_source" : {
"nome" : "João Silva",
"interesses" : [
"futebol",
"música",
"literatura"
],
"cidade" : "São Paulo",
"formação" : "Letras",
"estado" : "SP",
"país" : "Brasil"
}


futebol
João
Silva
música
Pedro
JSON
JSON



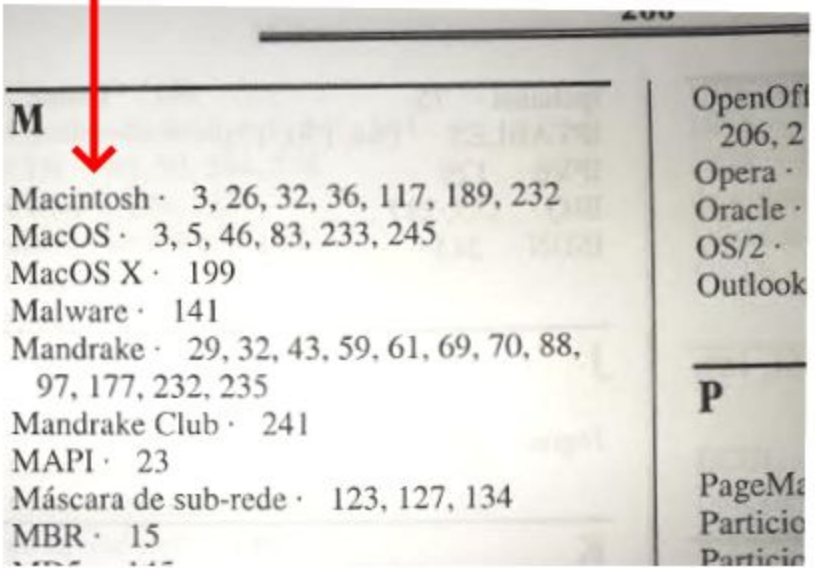
M

OpenOff
206, 2
Opera ·
Oracle ·
OS/2 ·
Outlook

P

PageMa
Particio

Tokens



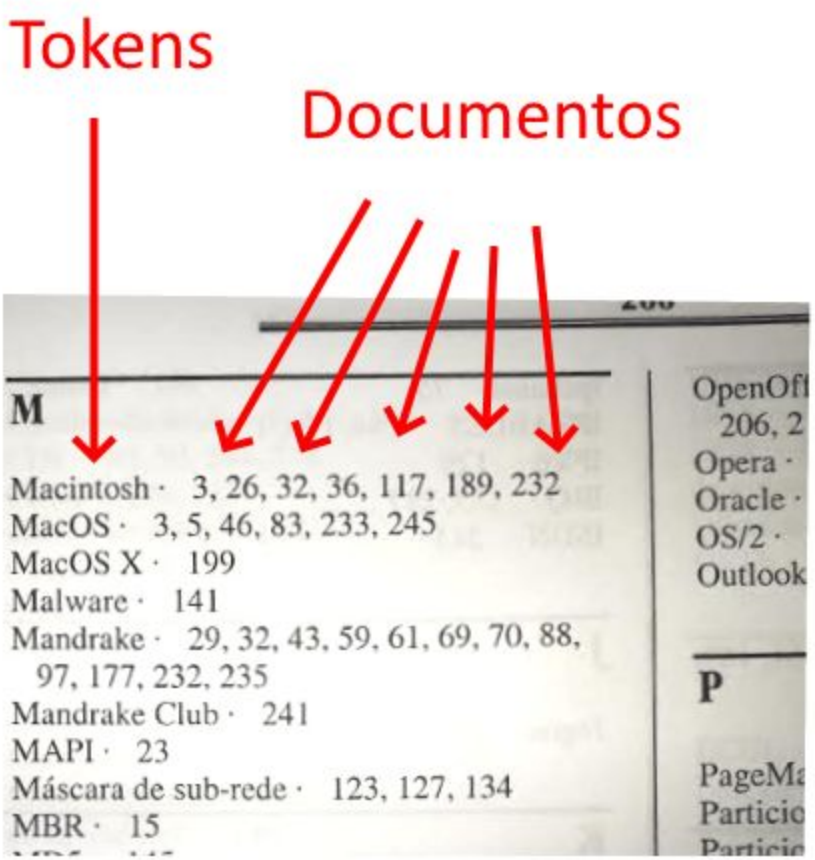
M

Macintosh · 3, 26, 32, 36, 117, 189, 232
MacOS · 3, 5, 46, 83, 233, 245
MacOS X · 199
Malware · 141
Mandrake · 29, 32, 43, 59, 61, 69, 70, 88, 97, 177, 232, 235
Mandrake Club · 241
MAPI · 23
Máscara de sub-rede · 123, 127, 134
MBR · 15

OpenOf
206, 2
Opera ·
Oracle ·
OS/2 ·
Outlook

P

PageMa
Particio


Tokens
Documentos
M
Macintosh · 3, 26, 32, 36, 117, 189, 232
MacOS · 3, 5, 46, 83, 233, 245
MacOS X · 199
Malware · 141
Mandrake · 29, 32, 43, 59, 61, 69, 70, 88, 97, 177, 232, 235
Mandrake Club · 241
MAPI · 23
Máscara de sub-rede · 123, 127, 134
MBR · 15
OpenOf
206, 2
Opera ·
Oracle ·
OS/2 ·
Outlook
P
PageMa
Particio

Por exemplo, dada a seguinte lista de documentos:

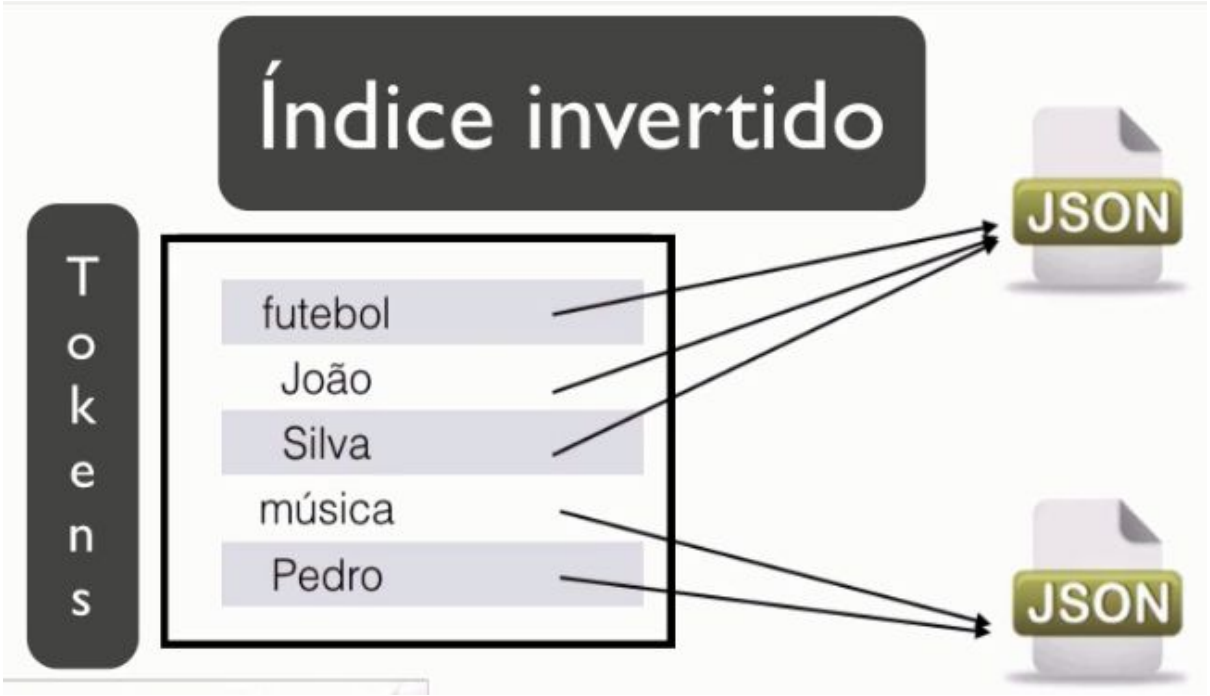
```
1: "Sei que sou"
2: "Sou o que sei"
3: "Sou uma banana"
```

Por exemplo, dada a seguinte lista de documentos:

```
1: "Sei que sou"
2: "Sou o que sei"
3: "Sou uma banana"
```

Obtemos a seguinte lista invertida:

```
"sei" : {1, 2}
"que" : {1, 2}
"sou" : {1, 2, 3}
"o" : {2}
"uma" : {3}
"banana" : {3}
```


Índice invertido
Tokens
futebol
João
Silva
música
Pedro
JSON
JSON

Analyzers Comuns

- Espaço em branco
- Padrão (standard)
- Simples
- Idioma (Portuguese, English, etc.)

Analyzers Comuns

- Espaço em branco
- Padrão (standard)
- Simples
- Idioma (Portuguese, English, etc.)

"Eu nasci há 10 mil (sim, isso mesmo, 10 mil) anos atrás"

```
GET catalogo/_analyze
{

}
```

```
GET catalogo/_analyze
{
  "text": "Eu nasci há 10 mil (sim, isso
mesmo, 10 mil) anos atrás"

}
```

```
{
  "tokens" : [
    {
      "token" : "eu",
      "start_offset" : 0,
      "end_offset" : 2,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "nasci",
      "start_offset" : 3,
      "end_offset" : 8,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "há",
      "start_offset" : 9,
      "end_offset" : 11,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
```

```
GET catalogo/_analyze
{
  "analyzer": "simple",
  "text":  "Eu nasci há 10 mil (sim, isso mesmo, 10 mil) anos atrás"
}
```

```
GET catalogo/_analyze
{
  "analyzer": "whitespace",
  "text":  "Eu nasci há 10 mil (sim, isso mesmo,
10 mil) anos atrás"
}
```

O analisador Portuguese

```
GET catalogo/_analyze
{
  "analyzer": "portuguese",
  "text":  "Eu nasci há 10 mil (sim,
isso mesmo, 10 mil) anos atrás"
}
```

```
GET catalogo/_analyze
{
  "analyzer": "portuguese",
  "text":  "Eu nasci há 10 mil (sim, isso
mesmo, 10 mil) anos atrás"
}
```


capitulo1.txt 10/12/2019 19:03
capitulo2.txt 10/12/2019 20:45
capitulo4.txt 14/12/2019 11:22
capitulo5.txt 10/12/2019 19:03
capitulo6.txt 10/12/2019 19:03
capitulo7.txt 10/12/2019 19:03
capitulo8.txt 10/12/2019 19:03

```
PUT /catalogo_v2
{
  "settings": {
    "index": {
      "number_of_shards": 3,
      "number_of_replicas": 0
    }
  },
  "mappings": {...
```

```
"formação": {
  "type": "text",
  "index": true,
  "analyzer": "portuguese"
},
"interesses": {
  "type": "text",
  "index": true,
  "analyzer": "portuguese"
},
"nome": {
  "type": "text",
  "index": true,
  "analyzer": "portuguese"
},
```

```
"estado": { //SEM ANALISADOR!
  "type": "text"
},
```

```
PUT /catalogo_v2
{
  "settings": {
    "index": {
      "number_of_shards": 3,
      "number_of_replicas": 0
    }
  },
```

```
PUT /catalogo_v2
{
  "settings": {
    "index": {
      "number_of_shards": 3,
      "number_of_replicas": 0
    }
  },
```

```
POST /catalogo_v2/_doc/1
{
    "nome": "João Silva",
    "interesses": ["futebol", "música",
"literatura"],
    "cidade": "São Paulo",
    "formação": "Letras",
    "estado": "SP",
    "país": "Brasil"
}
```

```
{
  "_index" : "catalogo",
  "_type" : "_doc",
  "_id" : "1",
  "_version" : 1,
  "result" : "created",
  "_shards" : {
    "total" : 2,
    "successful" : 1,
    "failed" : 0
  },
  "_seq_no" : 0,
  "_primary_term" : 1
}
```

GET catalogo/_search?q=música

- Música
- música
- Musica
- musica
- MUSICA
- musicA