

Adicionando sinônimos ao nosso catálogo

Agora que entendemos um pouco melhor o uso de sinônimos, precisamos tomar uma decisão importante. Em qual atributo queremos dar suporte aos sinônimos? Boa parte dos atributos já possuem analyzers. Tomemos como exemplo o campo `_all` :

```
"_all": {  
  "type": "string",  
  "index": "analyzed",  
  "analyzer": "portuguese"  
}
```

Veja que estamos contentes com nosso analyzer atual e nosso real objetivo é poder adicionar um certo nível de customização ao que já existe. Em quais atributos queremos utilizar sinônimos?

Podemos tanto utilizar o atributo `_all` como o atributo `interesses` . Note que ambos já utilizam o analyzer *portuguese*. Vale lembrar que esta decisão é muito particular e depende muito dos casos de negócio que estamos interessados. Sinônimos no atributo `_all` dão muita flexibilidade para buscas ao estilo Google (`?q=termos`), porém, como comentado acima, podem retornar documentos que na prática tem pouca relevância com o objetivo da busca. Note que, no início deste capítulo, fomos bem específicos na motivação: "queremos saber os interesses das pessoas". Logo, vamos utilizar sinônimos apenas no atributo `interesses` .

Como queremos que os sinônimos afetem nosso índice invertido? Em outras palavras, queremos/precisamos usar os sinônimos durante a indexação de documentos ou apenas durante a busca?

Alguém quer um índice inchado que, quando tivermos milhões de documentos, vai dar trabalho para manter ou gerar resultados desejados em nossas buscas? Pelo que vimos até então, seremos muito bem atendidos com a abordagem contracionista que deve ser utilizada durante a busca e não durante a indexação de um documento.

Qual a interação que queremos com analyzers existentes? Em outras palavras, como nossos sinônimos devem se comportar junto a *stemmers*, *stop words* e outros filtros que os atributos já possuem?

Queremos nos preocupar com letra maiúscula ou minúscula? Com plural ou singular na busca? Queremos o mesmo resultado quando buscamos por 'Esporte', 'Esportes', 'esporte' ou 'esportes'.

Nosso próximo passo é entender como devemos criar nossa lista de sinônimos que associe 'esportes' a uma lista de esportes, 'exatas' a uma lista de sinônimos relacionados à área de exatas e 'artes' a uma lista de sinônimos relacionados à artes. Nossa estratégia será a seguinte. Vamos identificar os *tokens* gerados pelo analyzer *portuguese* e usá-los na construção da lista de sinônimos.

Identificando os termos que precisamos

Vamos executar o analyzer *portuguese* no índice `catalogo_v2` para os termos:

- Esporte
- Esportes
- Basquete

- Volei
- Futebol
- Exatas
- Matemática
- Física
- Computação
- Artes
- Pintura
- Teatro
- Música
- Cinema

Os resultados são, respectivamente:

- esport
- esport
- basquet
- volei
- futebol
- exat
- matematic
- fisic
- computaca
- arte
- pintur
- teatr
- music
- cinem

Mais sinônimos

Agora que sabemos como tirar vantagem do analyzer **portuguese**, podemos construir nossa lista de sinônimos:

```
"futebol => futebol,society",  
"society => society,futebol",  
"volei,voleibol,volleyball",  
"esport => esport,futebol,society,volei,basquet",  
"exat => exat,matematic,fisic,computaca",  
"arte => arte,pintur,teatr,music,cinem"
```

Vamos criar um novo índice **catalogo_v3** e configurar nossos sinônimos:

```
PUT /catalogo_v3  
{  
  "settings": {  
    "index": {  
      "number_of_shards": 3,  

```

```
"number_of_replicas": 0
},
"analysis": {
  "filter": {
    "portuguese_stop": {
      "type": "stop",
      "stopwords": "_portuguese_"
    },
    "portuguese_stemmer": {
      "type": "stemmer",
      "language": "light_portuguese"
    },
    "filtro_de_sinonimos": {
      "type": "synonym",
      "synonyms": [
        "futebol => futebol,society",
        "society => society,futebol",
        "volei,voleibol,volleyball",
        "esport => esport,futebol,society,volei,basquet",
        "exat => exat,matematic,fisic,computaca",
        "arte => arte,pintur,teatr,music,cinem"
      ]
    }
  },
  "analyzer": {
    "sinonimos": {
      "tokenizer": "standard",
      "filter": [
        "lowercase",
        "portuguese_stop",
        "portuguese_stemmer",
        "filtro_de_sinonimos"
      ]
    }
  }
},
"mappings": {
  "pessoas": {
    "_all": {
      "type": "string",
      "index": "analyzed",
      "analyzer": "portuguese"
    },
    "properties": {
      "cidade": {
        "type": "string",
        "index": "analyzed",
        "analyzer": "portuguese"
      },
      "estado": {
        "type": "string"
      },
      "formação": {
        "type": "string",
        "index": "analyzed",
        "analyzer": "portuguese"
      }
    }
  }
}
```

```
"interesses": {
  "type": "string",
  "index": "analyzed",
  "analyzer": "portuguese",
  "search_analyzer": "sinonimos"
},
"nome": {
  "type": "string",
  "index": "analyzed",
  "analyzer": "portuguese"
},
"país": {
  "type": "string",
  "index": "analyzed",
  "analyzer": "portuguese"
}
}
}
}
```

Observemos que:

Usamos `"search_analyzer": "sinonimos"` para indicar que, durante buscas no atributo `interesses`, queremos aplicar o analyzer `sinonimos`.

O analyzer `sinonimos` utiliza filtros existentes e filtros que copiamos do analyzer `portuguese`, deste modo simulamos o mesmo comportamento do analyzer e ainda adicionamos os sinônimos na lista.

Para testar nossa estratégia, vamos adicionar os registros a seguir:

```
POST /catalogo_v3/pessoas/1
{
  "nome": "João Silva",
  "interesses": ["futebol", "música", "literatura"],
  "cidade": "São Paulo",
  "formação": "Letras",
  "estado": "SP",
  "país": "Brasil"
}
```

```
POST /catalogo_v3/pessoas/2
{
  "nome": "Maria Silva",
  "interesses": ["pintura", "literatura", "teatro"],
  "cidade": "Diamantina",
  "formação": "Artes Plásticas",
  "estado": "MG",
  "país": "Brasil"
}
```

```
POST /catalogo_v3/pessoas/3
{
  "nome": "Richard Edward",
  "interesses": ["matemática", "física", "música"],
}
```

```
"cidade": "Boston",
"formação": "Física",
"estado": "MA",
"país": "Estados Unidos"
}

POST /catalogo_v3/pessoas/4
{
  "nome": "Patrick von Steppat",
  "interesses": ["computação", "culinária", "cinema"],
  "cidade": "Rio de Janeiro",
  "formação": "Gastronomia",
  "estado": "RJ",
  "país": "Brasil"
}

POST /catalogo_v3/pessoas/5
{
  "nome": "Paulo Eduardo de Azevedo Silveira",
  "interesses": ["computação", "literatura"],
  "cidade": "São Paulo",
  "formação": "Computação",
  "estado": "SP",
  "país": "Brasil"
}

POST /catalogo_v3/pessoas/6
{
  "nome": "Michael Jordan",
  "interesses": ["basquete"],
  "cidade": "Chicago",
  "formação": "Artes",
  "estado": "IL",
  "país": "Estados Unidos"
}

POST /catalogo_v3/pessoas/7
{
  "nome": "Marcelo Negrão",
  "interesses": ["vôlei"],
  "cidade": "São Paulo",
  "formação": "Administração",
  "estado": "SP",
  "país": "Brasil"
}
```

E buscar por todas as pessoas que tenham interesse em **esportes**:

```
GET /catalogo_v3/pessoas/_search?q=interesses:esportes
```

Como resultado temos:

```
{
  "took": 15,
  "timed_out": false,
```

```
"_shards": {
  "total": 3,
  "successful": 3,
  "failed": 0
},
"hits": {
  "total": 3,
  "max_score": 0.69681984,
  "hits": [
    {
      "_index": "catalogo_v3",
      "_type": "pessoas",
      "_id": "6",
      "_score": 0.69681984,
      "_source": {
        "nome": "Michael Jordan",
        "interesses": [
          "basquete"
        ],
        "cidade": "Chicago",
        "formação": "Artes",
        "estado": "IL",
        "país": "Estados Unidos"
      }
    },
    {
      "_index": "catalogo_v3",
      "_type": "pessoas",
      "_id": "7",
      "_score": 0.69681984,
      "_source": {
        "nome": "Marcelo Negrão",
        "interesses": [
          "volei"
        ],
        "cidade": "São Paulo",
        "formação": "Administração",
        "estado": "SP",
        "país": "Brasil"
      }
    },
    {
      "_index": "catalogo_v3",
      "_type": "pessoas",
      "_id": "1",
      "_score": 0.14160846,
      "_source": {
        "nome": "João Silva",
        "interesses": [
          "futebol",
          "música",
          "literatura"
        ],
        "cidade": "São Paulo",
        "formação": "Letras",
        "estado": "SP",
        "país": "Brasil"
      }
    }
  ]
}
```

```
}  
]  
}  
}
```

O que aprendemos?

- Aplicação de analyzers durante indexação e busca.
- Os detalhes do analyzer portuguese.
- Como combinar um analyzer customizado para sinônimos a um analyzer existente.