03

Analyzers com Sinônimos

Até agora conseguimos flexibilizar nosso modelo com a ajuda de analyzers. Como comentado, analyzers têm o papel de ampliar os possíveis resultados através do uso de conceitos e ideias. Tome como exemplo a ideia de transformar as palavras em sua forma mais primitiva (*stemming*), remover pontuações e a transformação do texto para caixa baixa (*lowercase*). Contudo, as vezes queremos aplicar ideias que são muito simples do ponto de vista humano, como coloquialismos ou gírias, para ampliar o escopo de buscas.

Imagine que precisamos saber quais são as pessoas que possuem algum interesse esportivo, independente do esporte. Ou então as pessoas que gostam do bom e velho society. Ou ainda, pessoas que possuem interesses na área de exatas, como computação, física ou matemática.

Sinônimos podem substituir tokens ou adicionar tokens ao índice invertido e são implementados via analyzers customizados. O ponto positivo aqui é que os analyzers padrão do ElasticSearch também são implementados desta maneira (por mais contra-intuitivo que isso possa parecer). Antes de criamos nosso analyzer customizado com sinônimos, vamos dar uma breve olhada no analyzer *portuguese* que utilizamos anteriormente:

```
{
  "settings": {
    "analysis": {
      "filter": {
        "portuguese_stop": {
          "type":       "stop",
          "stopwords":  "_portuguese_"
        },
        "portuguese_keywords": {
          "type":       "keyword_marker",
          "keywords":   []
        },
        "portuguese_stemmer": {
          "type":       "stemmer",
          "language":   "light_portuguese"
        }
      },
      "analyzer": {
        "portuguese": {
          "tokenizer":  "standard",
          "filter": [
            "lowercase",
            "portuguese_stop",
            "portuguese_keywords",
            "portuguese_stemmer"
          ]
        }
      }
    }
  }
}
```

Repare que utilizamos diversos *filters* na parte de análise. Um *synonym* nada mais é do que um filtro onde definimos o mapa de sinônimos. Vejamos o exemplo a seguir:

```
{
  "settings": {
    "analysis": {
      "filter": {
        "filtro_de_sinonimos": {  (1)
            "type": "synonym",
            "synonyms": [
                "esporte,futebol,society,futeba,pelada" (2)
            ]
        }
      },
      "analyzer": {
        "sinonimos": {
          "tokenizer":  "standard",
          "filter": [
            "lowercase",
            "filtro_de_sinonimos"  (3)
          ]
        }
      }
    }
  }
}
```

Onde:

- (1) Definimos um filtro do tipo synonym.
- (2) Definimos à lista de palavras que são consideradas semelhantes (mais sobre isso a seguir).
- (3) Aplicamos o filtro de sinônimos criado no processo de análise da informação.

Quando uma das palavras é encontrada na lista de sinônimos, ela é substituída por todas as palavras que estão na lista. Quando isso acontece durante a fase de indexação de documentos, os sinônimos são adicionados ao índice invertido que é criado para busca. Vale chamar a atenção que, esta é uma abordagem expansionista e pode colaborar para um "inchaço" do índice.

Interagindo com sinônimos

Ao invés de alteramos nosso índice, vamos primeiro entender os efeitos colaterais que o uso de sinônimos pode trazer em nossas buscas.

Vamos criar um novo índice com as configurações a seguir:

```
PUT /indice_com_sinonimo
{
  "settings": {
    "index": {
      "number_of_shards": 3,
      "number_of_replicas": 0
    },
    "analysis": {
      "filter": {
        "filtro_de_sinonimos": {
            "type": "synonym",
```

```
            "synonyms": [
                "esporte,futebol,society,futeba,pelada"
            ]
          }
        },
        "analyzer": {
          "sinonimos": {
            "tokenizer":  "standard",
            "filter": [
              "lowercase",
              "filtro_de_sinonimos"
            ]
          }
        }
      }
    }
}
```

Vamos executar a seguinte consulta para verificar os tokens marcados na *position* 4:

```
GET /indice_com_sinonimo/_analyze?analyzer=sinonimos&text=eu+gosto+de+jogar+society
```

Usamos nosso novo índice (*indice_com_sinonimo*) passando como parâmetro o nome do analyzer e o texto para analisar. Como resultado devemos receber:

```
{
  "tokens": [
    {
      "token": "eu",
      "start_offset": 0,
      "end_offset": 2,
      "type": "<ALPHANUM>",
      "position": 0
    },
    {
      "token": "gosto",
      "start_offset": 3,
      "end_offset": 8,
      "type": "<ALPHANUM>",
      "position": 1
    },
    {
      "token": "de",
      "start_offset": 9,
      "end_offset": 11,
      "type": "<ALPHANUM>",
      "position": 2
    },
    {
      "token": "jogar",
      "start_offset": 12,
      "end_offset": 17,
      "type": "<ALPHANUM>",
      "position": 3
```

```
      },
      {
        "token": "society",
        "start_offset": 18,
        "end_offset": 25,
        "type": "<ALPHANUM>",
        "position": 4
      },
      {
        "token": "esporte",
        "start_offset": 18,
        "end_offset": 25,
        "type": "SYNONYM",
        "position": 4
      },
      {
        "token": "futebol",
        "start_offset": 18,
        "end_offset": 25,
        "type": "SYNONYM",
        "position": 4
      },
      {
        "token": "futeba",
        "start_offset": 18,
        "end_offset": 25,
        "type": "SYNONYM",
        "position": 4
      },
      {
        "token": "pelada",
        "start_offset": 18,
        "end_offset": 25,
        "type": "SYNONYM",
        "position": 4
      }
    ]
}
```


Agora vamos fazer outra consulta para verificar os tokens marcados na *position* 4:

```
GET /indice_com_sinonimo/_analyze?analyzer=sinonimos&text=eu+gosto+de+praticar+esporte
```

E o resultado:


```
{
  "tokens": [
    {
      "token": "eu",
      "start_offset": 0,
      "end_offset": 2,
      "type": "<ALPHANUM>",
      "position": 0
    },
    {
```

```
      "token": "gosto",
      "start_offset": 3,
      "end_offset": 8,
      "type": "<ALPHANUM>",
      "position": 1
    },
    {
      "token": "de",
      "start_offset": 9,
      "end_offset": 11,
      "type": "<ALPHANUM>",
      "position": 2
    },
    {
      "token": "praticar",
      "start_offset": 12,
      "end_offset": 20,
      "type": "<ALPHANUM>",
      "position": 3
    },
    {
      "token": "esporte",
      "start_offset": 21,
      "end_offset": 28,
      "type": "<ALPHANUM>",
      "position": 4
    },
    {
      "token": "futebol",
      "start_offset": 21,
      "end_offset": 28,
      "type": "SYNONYM",
      "position": 4
    },
    {
      "token": "society",
      "start_offset": 21,
      "end_offset": 28,
      "type": "SYNONYM",
      "position": 4
    },
    {
      "token": "futeba",
      "start_offset": 21,
      "end_offset": 28,
      "type": "SYNONYM",
      "position": 4
    },
    {
      "token": "pelada",
      "start_offset": 21,
      "end_offset": 28,
      "type": "SYNONYM",
      "position": 4
    }
  ]
}
```

Por fim, vamos testar os tokens marcados na *position* 2:

```
GET /indice_com_sinonimo/_analyze?analyzer=sinonimos&text=arvore+praticamente+pelada
```

E o resultado:


```
{
  "tokens": [
    {
      "token": "arvore",
      "start_offset": 0,
      "end_offset": 6,
      "type": "<ALPHANUM>",
      "position": 0
    },
    {
      "token": "praticamente",
      "start_offset": 7,
      "end_offset": 19,
      "type": "<ALPHANUM>",
      "position": 1
    },
    {
      "token": "pelada",
      "start_offset": 20,
      "end_offset": 26,
      "type": "<ALPHANUM>",
      "position": 2
    },
    {
      "token": "esporte",
      "start_offset": 20,
      "end_offset": 26,
      "type": "SYNONYM",
      "position": 2
    },
    {
      "token": "futebol",
      "start_offset": 20,
      "end_offset": 26,
      "type": "SYNONYM",
      "position": 2
    },
    {
      "token": "society",
      "start_offset": 20,
      "end_offset": 26,
      "type": "SYNONYM",
      "position": 2
    },
    {
      "token": "futeba",
      "start_offset": 20,
      "end_offset": 26,
      "type": "SYNONYM",
```

```
        "position": 2
      }
    ]
  }
```

Caso estivéssemos utilizando a configuração de sinônimos acima, o que aconteceria com as buscas com termos como 'pelada' e 'esporte'?

Refinando nossos sinônimos

Ainda que o uso de sinônimos seja muito útil e nos dê um nível de customização para nossas buscas, devemos ter cuidado com alguns detalhes, como o inchaço do índice invertido ou documentos que não possuem relevância para quem está fazendo a busca.

Podemos melhorar os resultados com sinônimos utilizando uma abordagem inversa a expansionista. Ao invés da sintaxe:

```
"sinonimo_1,sinonimo_2,...,sinonimo_n"
```

Podemos utilizar:

```
"sinonimo1,sinonimo2 => termo"
```

Por exemplo:

```
"esporte,futebol,basquete,esporte,society => esporte"
```

Podemos ler a configuração acima como *"os termos esporte, futebol, basquete, esporte e society 'significam' esporte, mas esporte não significa estes termos"*. Neste caso, quando indexamos documentos com qualquer palavra da lista, a entrada no índice inverso será criada para o termo "esporte". Esta abordagem é chamada de 'contracionista'.

Podemos ainda utilizar a abordagem de expansão de gênero como mostrado a seguir:

```
"esporte => futebol,basquete,society,volei"
"society => society,futebol"
"futebol => futebol,society"
```

Podemos ler a configuração acima como "o termo esporte significa futebol, basquete, society e volei, mas nenhum destes termos significa esporte". Este tipo de configuração nos permite buscar pelo esporte e encontrar documentos com as palavras futebol ou society, ou mesmo volei, mas quando buscamos pelo termo futebol, encontraremos documentos com futebol e society, mas não com volei.

*** Importante: Tanto na abordagem contracionista quanto na expansão de gênero, a análise do sinônimo será aplicada durante a indexação do documento e no momento da busca. Veremos como fazê-lo a seguir. ***

Criando índice com sinônimos

Vamos criar um novo índice com as configurações a seguir:

```
PUT /indice_com_sinonimo_2
{
  "settings": {
    "index": {
      "number_of_shards": 3,
      "number_of_replicas": 0
    },
    "analysis": {
      "filter": {
        "filtro_de_sinonimos": {
            "type": "synonym",
            "synonyms": [
        "futebol => futebol,society",
        "society => society,futebol",
        "esporte => esporte,futebol,society,volei,basquete"
            ]
        }
      },
      "analyzer": {
        "sinonimos": {
          "tokenizer":  "standard",
          "filter": [
            "lowercase",
            "filtro_de_sinonimos"
          ]
        }
      }
    }
  }
}
```

Executando a seguinte seguinte consulta:

```
GET /indice_com_sinonimo_2/_analyze?analyzer=sinonimos&text=futebol
```

Temos o seguinte resultado:

```
{
  "tokens": [
    {
      "token": "futebol",
      "start_offset": 0,
      "end_offset": 7,
      "type": "SYNONYM",
      "position": 0
    },
    {
      "token": "society",
      "start_offset": 0,
      "end_offset": 7,
      "type": "SYNONYM",
      "position": 0
```

```
      }
    ]
  }
```

Agora, executando a seguinte consulta:

```
GET /indice_com_sinonimo_2/_analyze?analyzer=sinonimos&text=esporte
```

Temos:

```
{
  "tokens": [
    {
      "token": "esporte",
      "start_offset": 0,
      "end_offset": 7,
      "type": "SYNONYM",
      "position": 0
    },
    {
      "token": "futebol",
      "start_offset": 0,
      "end_offset": 7,
      "type": "SYNONYM",
      "position": 0
    },
    {
      "token": "society",
      "start_offset": 0,
      "end_offset": 7,
      "type": "SYNONYM",
      "position": 0
    },
    {
      "token": "volei",
      "start_offset": 0,
      "end_offset": 7,
      "type": "SYNONYM",
      "position": 0
    },
    {
      "token": "basquete",
      "start_offset": 0,
      "end_offset": 7,
      "type": "SYNONYM",
      "position": 0
    }
  ]
}
```

O que acontece se executamos a consulta a seguir?

```
GET /indice_com_sinonimo_2/_analyze?analyzer=sinonimos&text=esportes
```

Obtemos um resultado que possui apenas o termo esportes! Isso significa que, utilizando nosso analyzer, o termo "esportes" não possui resultado igual ao termo "esporte".

O que aprendemos?

- Como funciona o suporte a sinônimos no ElasticSearch.
- A anatomia de um analyzer.
- Como criar um analyzer customizado para dar suporte a sinônimos.
- Como testar um analyzer customizado.
- As diferentes abordagens para o uso de sinônimos.