Elastic Search: Analyzer[1]

How Text are Tokenized, Selected, and Normalized

Weimao Ke

Drexel University

Table of contents

Text Analyzer

What is an Analyzer

An Analyzer:

- Converts text into tokens or terms
- May apply filtering, e.g. stopword removal
- May perform normalization, e.g. case folding
- Built-in or custom

What is an Analyzer

Two types of text analyzers:

- Index time analyzer: converts document text to terms in the inverted index
- Search time analyzer: converts query text to terms for matching

Note:

- Query terms and document terms should be in the same form
- Ideally, use the same index time analyzer and search time analyzer

Index Time Analyzer

Default Analyzer

If the analyzer is not specified before indexing a document:

```
PUT /my_index/_doc/1
{
  "message": "The QUICK brown foxes
              jumped over the lazy dog!"
}
```

The default *english* analyzer will be used, which:

- Lowercases each token, e.g. *QUICK* $\rightarrow$ *quick*
- Removes frequent stopwords, e.g. *the*
- Reduces each term to its word stem, e.g. *foxes* $\rightarrow$ *fox*

And the following will be used in the inverted index:

```
[ quick, brown, fox, jump, over, lazi, dog ]
```

Analyzer Specification

You can specify the index time analyzer with mapping:

```
PUT /my_index
{
  "mappings": {
    "properties": {
      "message": {
        "type": "text",
        "analyzer": "standard"
      }
    }
  }
}
```

Analyzer Specification

This line:

```
7   "analyzer": "standard"
```

specifies the *standard* analyzer, which:

- Provides a grammar based tokenization
- Is based on Unicode Text Segmentation
- Works well for many languages

Analyzer Specification

Example:

```
POST _analyze
{
  "analyzer": "standard",
  "text": "The 2 QUICK Brown-Foxes jumped
           over the lazy dog's bone."
}
```

The result of the *standard* analyzer:

```
[ the, 2, quick, brown, foxes, jumped,
          over, the, lazy, dog's, bone ]
```

Analyzer Specification

The *standard* analyzer consists of:

- A standard tokenizer (Unicode Text Segmentation)
- Token filters
  - Lower Case Token Filter
  - Stop Token Filter (disabled by default)

It accepts parameters:

- *max_token_length*: a token must be within the limit or be split.
- *stopwords*: pre-define stopwords (e.g. _*english*_) or a custom array/list.
- *stopwords_path*: the path to a file containing stop words.

Analyzer Specification

Example:

```
PUT my_index
{
  "settings": {
    "analysis": {
      "analyzer": {
        "my_english_analyzer": {
          "type": "standard",
          "max_token_length": 5,
          "stopwords": "_english_"
        }
      }
    }
  }
}
```

Analyzer Specification

Now if you use the custom *my_english_analyzer*:

```
POST my_index/_analyze
{
  "analyzer": "my_english_analyzer",
  "text": "The 2 QUICK Brown-Foxes jumped over the
     lazy dog's bone."
}
```

The following terms will be produced:

```
[ 2, quick, brown, foxes, jumpe, d,
            over, lazy, dog's, bone ]
```

Note that:

- The stopword "the" has been removed.
- The term "jumped" is split into "jumpe" and "d".

Search Time Analyzer

Search Time Analyzer

With a full-text *match* query, for example:

```
GET /my_index/_search
{
    "query": {
        "match" : {
            "message" : {
                "query" : "a quick fox"
            }
        }
    }
}
```

The same analyzer (as the index time analyzer) will be used.

Search Time Analyzer

It is possible, though, to use a different search time analyzer.

To specify a search time analyzer within a query:

```
GET /my_index/_search
{
    "query": {
        "match" : {
            "message" : {
                "query" : "a quick fox",
                "analyzer": "standard",
            }
        }
    }
}
```

Search Time Analyzer

You may also use a specific search time analyzer with mapping:

```
PUT /my_index
{
  "mappings": {
    "properties": {
      "message": {
        "type": "text",
        "analyzer": "standard",
        "search_analyzer": "english"
      }
    }
  }
}
```

References

[1] elastic.co. Elasticsearch reference [7.5]: Text analysis. https://www.elastic.co/guide/en/elasticsearch/reference/7.5/analysis.html. Accessed: 2020-1-16.