Analyzer

A Combination of **Character Filters**, **Tokenizers**, **Token Filters**

Processing pipeline

Input --*String*--> **Character Filters** --*String*--> **Tokenizer** --*Tokens*---> **Token Filters** --*Tokens*--> **Output**

Character Filters

The Character filter removes words from the input Strings.
e.g. **html_strip char filter**:
Input: "*<p> An example
</p>*"
Output: "*An example*"

Tokenizer

The way the input is divided into parts (tokens)
e.g. **whitespace tokenizer**
Input: *The quick brown fox*
Tokens: [*The, quick, brown, fox*].

Token Filters

A Tokenfilter returns a subset of given tokens.
e.g. **stop word filter** that removes all the **stop words** from the tokens, such as *"the"*, *"and"*, ...