Highlight of text patterns entrances – functionality general specification definitions

Highlight is an algorithm of text processing that gets the search query string and textual context on input and returns textual content with marks of entrances of patterns from search query and additional stat information. Patterns usually are lexical words, but depend on stemming and tokenizing processes can be more complex constructions.

Input parameters:

Process can use some additional properties that vary of properties and behavior of algorithm. Properties list:

1. Highlighted pattern marker string “begin” for all cases.
2. Highlighted pattern marker string “end” for all cases.
3. Max number of highlighted entrances for all patterns. Not less than sum of maximums for each pattern if set.
4. Max numbers of highlighted entrances for each pattern from search string. (future implementation)
5. List of delimiter characters for tokenizer. If set replaces default delimiters list: **~`!@#$%^&*()_+=-{}[]<>,.?/\|** as well as ASCII characters with code value from 0 to 32 inclusively. Not implemented.
6. List of acronym patterns that must not be split by tokenizer. Not implemented.
7. Treat numbers as lexical words, Boolean flag. Not implemented.

Output result fields:

1. Textual content with highlight marker patterns “begin” and “end” inserted.
2. Counter – total number of single patterns and phrases from search string highlighted.
3. Counter – total number of single patterns (formally lexical words) found.

Different forms of patterns to highlight

Single words forms varied

It is formally lexical word that can have different morphological forms and can be identified by stemmer as equal lexical roots. For example, single and plural of nouns, time ending of verbs and so on.

Test:

Search string: test
Textual content: This is test content that can be used for many different tests.
Highlighted content: This is **test** content that can be used for many different **tests**.

Single words forms exact matches

Exact match word forms are completely the same as searched patterns.

Test:

Search string: “test”
Textual content: This is test content that can be used for many different tests.
Highlighted content: This is **test** content that can be used for many different tests.

Phrases

Phrases are several single words patterns. Words from phrase are highlighted separately by each word pattern independent way.

Test:

Search string: may tests used
Textual content: This is test content that can be used for many different tests.
Highlighted content: This is **test** content that can be **used** for **many** different **tests**.

Phrases exact match

Exact matches phrases are highlighted the same way as regular phrases, but words from phrase obey to rules of single words exact matches. Additionally, words of phrase need to be located one by one sequentially in exact order.

Test:

Search string: “test content”
Textual content: This is test content that can be used for many different tests.
Highlighted content: This is **test content** that can be used for many different tests.

Phrases near exact match

Near exact matches phrases are highlighted the same way as exact matches, but words from phrase can change order.

Test:

Search string: ‘content test’
Textual content: This is test content that can be used for many different tests.
Highlighted content: This is **test content** that can be used for many different tests.

Mixed phrases

This type can have any recombination of cases that are described above.

Test:

Search string: many “test content” use ‘tests different’
Textual content: This is test content that can be used for many different tests.
Highlighted content: This is **test content** that can be **used** for **many different tests**.