# Lowercase, Tokenize, Stem, and Stopword Text

### *Description*

Normalizes free-form text from selected columns within a given table. For example, the input text "Emergence of Scaling in Random Networks" becomes "emerg|scale|random|network" where we have chosen "|" as the character separating individual items of the list.

From this example you can follow the four normalization steps:

1. Lowercase: The example text becomes "emergence of scaling in random networks".
2. Tokenize: The text blob is split into a list of individual words. The example text becomes "emergence|of|scaling|in|random|networks".
3. Stem: Common or low-content prefixes and suffixes are removed to identify the core concept. The example text becomes "emerg|of|scale|in|random|network".
4. Stopword: Low-content tokens like "of" and "in" are removed (see the complete stopword list). The example text becomes "emerg|scale|random|network".

### *Parameters*

- *Stopword List*: The plain-text file that contains the list of stopwords to use. By default, it points to the included stopword list. If an invalid file path is specified, it will again default to the included stopword list. Stopwords are separated by line (so each line lists a single stopword).
- *New Separator*: The character that will separate items in the output lists of tokens.
- Each individual textual column of the table can be selected or not selected for normalization.

### *Applications*

This algorithm can prepare the text in a table for Burst Detection.

### *Links*

- Source code
- Stopword list

### *See Also*