

Extracting Term Candidates with extraTerm

Get your Terminology Started

You are lacking a company specific terminology but would like to compile one? You want to extend your existing terminology on the basis of your document collection? So far you have been building your terminology without the assistance of tools?

Our term extraction tools and services provide you with optimal support for compiling your terminology. Should you already use term extraction tools you are welcome to compare the results with those provided by extraTerm.

The results of extraTerm easily stand out in any comparison!



Linguistic Analysis

The basis of our analysis is a set of mature language technologies, amongst others LCCore. A morphological analysis component determines the lemma of each word, its syntactic category and other grammatical information such as gender and number of nouns. By lemmatisation different word forms are collected under one lemma (*configuration, configurations*). A grammatical analysis component detects the congruence of word groups and sentence patterns.

On the basis of linguistic properties term candidates are determined. In particular:

- multi-word compounds built from adjectives and nouns (*acoustic absorber, active navigation system*)
- compound nouns (*acrylester, handgrip, microsurgery*)
- derived or simplex nouns (*manoeuvrability, rubber*)

Statistical Analysis

Statistical analysis determines frequency of occurrence of term candidates as well as the number of documents in which a term candidate occurs. These statistical figures may help you in deciding whether a term candidate is to be added to your terminology or not.

Features

- lemmatised terms
- grammatical features
- different types of term formation patterns
- frequency of occurrence per term candidate
- number of documents in which a term occurs
- list of documents in which a term occurs
- sample contexts

Languages

- German
- English

No.	Lemmatized Term	Word Forms	Parts of Speech	Criterion	Frequency of Occurrence	Number of Documents	Documents
241	configurable key	configurable keys	adj noun	compound.mwu	3	1	LC_09_2015.htm
242	configurable threshold	configurable threshold	adj noun	compound.mwu	1	1	LC_09_2015.htm
243	configurable video output	Configurable video output	adj noun noun	compound.mwu	1	1	LC_09_2015.htm
244	configuration	Configuration configuration configurations	noun	simplex	36	1	LC_09_2015.htm
245	configuration data	configuration data	noun noun	compound.mwu	1	1	LC_09_2015.htm
246	configuration menu	configuration menu configuration menus	noun noun	compound.mwu	5	1	LC_09_2015.htm
247	configuration option	Configuration options	noun noun	compound.mwu	4	1	LC_09_2015.htm
248	configured button	configured button	adj noun	compound.mwu	1	1	LC_09_2015.htm

Determination of Sample Contexts

The determination of sample context per term candidate is based on a combination of statistical and linguistic methods. The sample contexts allow you for checking the distribution of a term and how it is used in different environments.

Consideration of Metadata

The consideration of metadata such as the 'sort of document' or the 'department in which a document is relevant' allows for a specific view on how terms are used across departments, different text sorts and the like.

Consideration of Existing Terminology

The term extraction process may take an existing terminology into account. Terms that are already part of the existing terminology are marked. This allows you to get an idea on how frequently existing terminology is used in your documents. It gives you a clue on which terms need to be added to the existing terminology and which terms from the existing terminology are obviously irrelevant, as they are not used in a single document.

Evaluation of Extracted Term Candidates

You may combine term extraction with terminology evaluation by **evalTerm**. This gives you information about whether term candidates contain items with incorrect spelling, which term candidates are problematic according to term formation rules (e.g. extremely complex compounds) or which term candidates possibly are variants of each other such as *cost reduction*, *reduction of costs*, or *reducing costs*.

Need more Information?

If you have special requirements regarding our term extraction service, or if you are interested in getting a deeper insight into our powerful term extraction service, do not hesitate to contact us.

IAI Linguistic Content AG

Rathausplatz 8
66111 Saarbrücken
+49 681 389510

www.iailc.de
kontakt@iailc.de