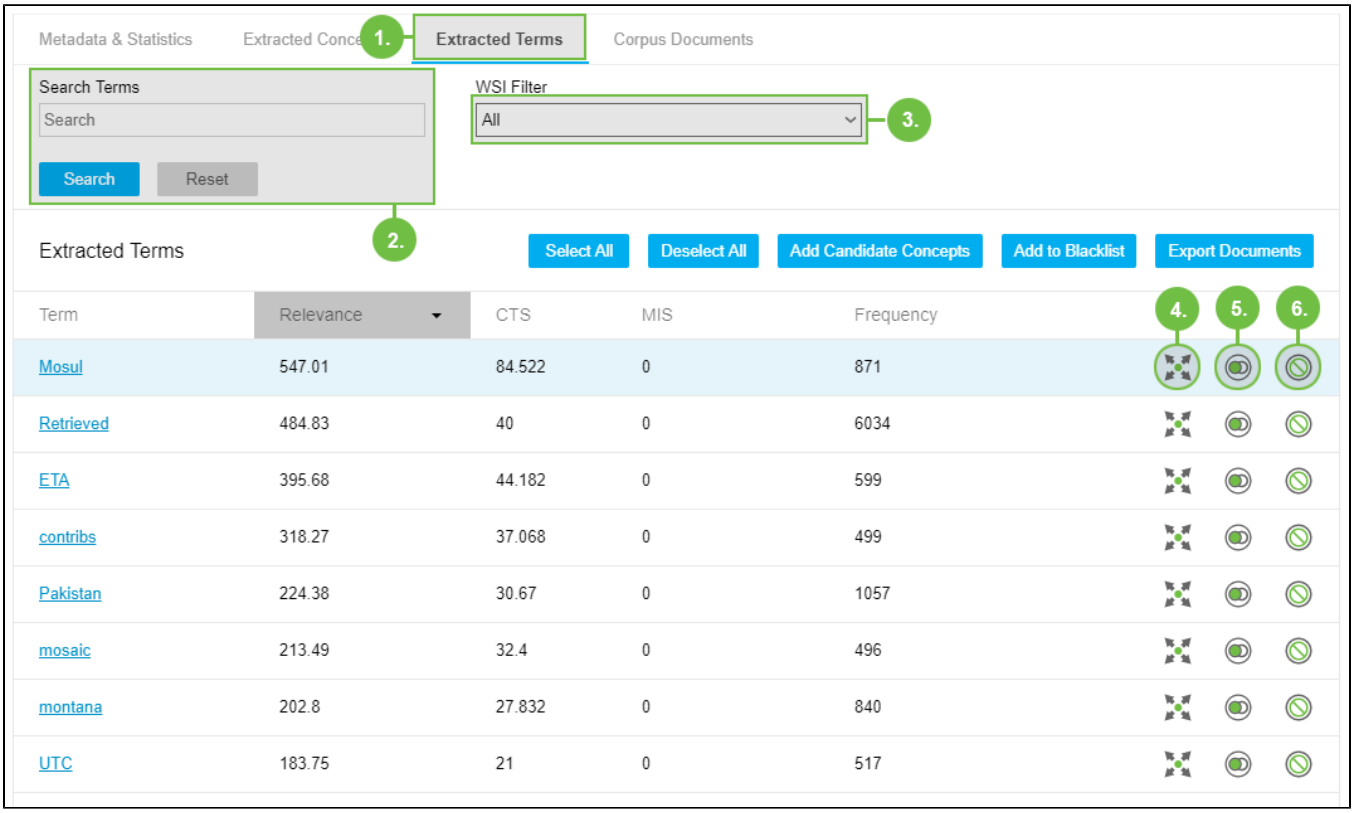
Extracted Terms List

This section contains a short guide on how to use the **Extracted Terms** list of the corpus management.

The **Extracted Terms** list shows terms which have been extracted from the documents in the corpus and which are not yet part of the thesaurus.

In your opened corpus, click the **Extracted Terms** (1) tab to open the **Extracted Terms** list. Details about available options find below.



Available Options

- Use **Search Terms** (2) to filter for concepts, click **Search** to start the search. A list of results will be displayed. Click **Reset** to display the whole list.
- The **WSI Filter** (3) refers to the **Word Sense Induction** function in PoolParty. You can use it to filter the terms for ambiguous or unambiguous terms PoolParty has extracted from the corpus.
- Use **Select All** and **Deselect All** for selecting and deselecting the whole list of terms.
- Use **Export Documents** to export all corpus documents.
- You can select candidate terms and link them to existing concepts as synonyms or as narrower concepts. Select one or more terms in this list using the default keys on your keyboard and mouse clicks, then click **Add Candidate Concepts**.
- The **Word Senses** icon (4, display is optional, **WSI** needs to be enabled) opens a list of terms whose senses are similar, ambiguous or unambiguous in the context of the corpus and the thesaurus, PoolParty calculates. Thus you can specify precisely, which additional terms could be added to the list of candidate concepts. In the dialogue that will open you also see word senses of similar terms that have been deduced from the corpus, as well as term and concept suggestions. You can decide if you want to merge senses, use the terms as candidate concepts or just return to the main window.
- Use the the **Similar Terms** icon (5) to generate a list of terms that contains only terms which are similar to the respective term. Details find in this topic: Extracted Terms - Use the Similar Terms Function.
- Use the **Add to Blacklist** icon (6) to exclude terms from extraction.

Extracted Terms List

This list provides an overview of statistical relevant terms that were found in the document corpus.

You can sort the results by clicking the table headers. The image above shows a table sorted by **Relevance** column:

Table Columns

You can use the following four table columns to sort terms by these relevance factors:

- **Relevance:** The Mutual Information Score, the Content Term Score and the term frequency were combined into one score that gives an overall relevance that is normally a good starting point for going through the list of extracted terms.
- **Mutual Information Score (MIS)**: Mutual information (MI) provides information about dependency of variables and can be used to estimate if two or more consecutive words in a text should be considered a compound term that is formed by those words. The idea is that if words are

independent that they will occur together just by chance. On the other hand, if they are observed together more often than expected, they are dependent and are candidates for terms. This score ranks multi term phrases higher.

- **Content Term Score (CTS)**: Content terms are enriched in documents where they appear, which means they are not the most frequent terms in the document set. But when they occur in a document they tend to occur very often, which indicates their importance for defining the content of the document. This score ranks single term phrases higher.
- **Frequency**: The total number of occurrences of a term in the corpus.