Lucene Word Stemming

Customer Self Service with integrated Lucene Search offers users fast, powerful search queries. One of Lucene's optional features is the use of "word stemming" for indexing and search queries.

Word stemming essentially chops off known English suffixes to get back to the word stem.

For example, lamin**ate**, lamin**ator**, lamin**ating**, and lamin**ated** are all stemmed to "lamin".

This stemming is important for both indexing and searching. Using the example above, a product with a description of "A4 laminating machine" would be indexed as "A4 lamin machine". If a user searches for "laminator", Lucene will actually search for "lamin" - and so the product indexed as "A4 lamin machine" will be found. This is how a search for "laminator" can return a product with a description of "A4 laminating machine".

It is important to note, however, that "lamina" is not stemmed to "lamin". Stemming won't simply chop off an "a" – it will treat "lamina" as a unique word and therefore will not find a product indexed as "A4 lamin machine".

With stemming turned off, a search for "lamina" will return anything starting with lamina, because Lucene will not have stemmed the words when indexing. So the example product above would be indexed as "A4 laminating machine" and will be found when you search for "lamina". However, searching for "laminator" will no longer return products containing words like "laminating, laminated", etc.

Commerce Vision recommend that sites utilising Lucene Search leave stemming turned on – the benefit far outweighs the potential partial word issue, as generally users will keep typing and get their results.

Stemming can be enabled by choosing the "Snowball Analyser" when setting up the Lucene Index fields.

ⓘ   For further information on Lucene, see Product Search with the Lucene Search Engine.

Related articles

- Product Search with Lucene Search Text Engine
- Lucene Introduction
- Lucene Query Translation Maintenance
- Predictive Search
- Build or Rebuild the Lucene Search Index