RightFind™ XML for Mining

Help: Creating a Lucene Query Project

This guide explains how to create a project in XML for Mining using a syntactically valid Lucene query.

Query Syntax

The search engine XML for Mining uses is called Elasticsearch which is a distributed scalable real time full text search engine built on top of Apache Lucene, one of the most successful open source projects for enterprise applications.

To create a Lucene Query Project, use Lucene syntax in the free text area of the Create Project page. Specify the index field (or combination of fields through Boolean operators) and perform *keyword matching*, *wildcard matching*, *fuzzy matching*, and *proximity matching*. Lucene's query syntax also supports *range searches*, *boosts*, and *nested queries*.

Keyword and Wildcard Matching

When performing a search, you can either specify a field or use the default field. Field names and default field is implementation specific. You can search any field by entering the field name, a colon ":", and the term for which you are looking.

Assume you want to use the fields `publisherId` and `content`, with `content` as the default field. To find documents by Springer that contain the word **diabetes**, type:

```
publisherId:springer_TDM AND content:diabetes
```

or

```
pid:springer* AND diabetes
```

Since `content` is the **default field**, the field indicator is not required. The content field represents all the full text in the document. In this example also note the use of the shorter field id for the publisherId and the wildcard used to find the publisher name.

Note: The field name is only valid for the term that it directly precedes.

To search for all documents from Springer that contain a word that starts with **micro** in the abstract, perform a search similar to the following example:

```
publisherId:springer* AND abstract:micro*
```

In this example, the * symbol is the wildcard. You can also search for words that start with **foo** and end with **bar** by using the string **`foo*bar`**.

Updated: 11 January 2020

Note: Placing wildcards as the first character of a term is not supported.

To perform a single character wildcard query, use the **?** character. For example:

```
publisherId:springer* AND abstract:micro?NA
```

This query matches words that start with **micro** followed by one letter and the letters **NA**, such as **microDNA** and **microRNA**.

Fuzzy and Proximity Matching

Lucene supports fuzzy searches based on Damerau-Levenshtein distance. To perform a fuzzy search, use the tilde symbol (**~**) at the end of a single word term. For example to search for a term similar in spelling to **apoplexia**, use the fuzzy search:

```
apoplexia~
```

This search finds terms such as **apoplexia** and **pagoplexia**.

To specify the maximum number of edits allowed, add a parameter between 0 and 2. If the parameter is omitted, the number of edits defaults to 2.

Lucene supports proximity searches that find words that are a specific distance away from each other. To perform a proximity search, use the tilde symbol (**~**) at the end of a phrase. For example, to search for Springer documents that contain the word **diabetes** and **treatment** four words apart from each other, specify the following query in the abstract field:

```
publisherId:springer* AND abstract:"diabetes treatment"~4
```

Range Searching

Range queries let you match documents whose field values are between the lower and upper bound specified by the **Range Query**.

Range queries can be inclusive or exclusive of the upper and lower bounds. Sorting is performed lexicographically. Inclusive range queries are denoted by square brackets. Exclusive range queries are denoted by curly brackets.

For example:

```
date:[2014-01-01 TO 2015-01-01]
```

finds documents whose `mod_date` fields have values between **2014-01-01** and **2015-01-01**, inclusive where the date format is YYYY-MM-DD.

Range Queries are not reserved for date fields. You can use range queries with non-date fields. For example:

```
metadata_title:{Aida TO Carmen}
```

finds all documents whose titles are between **Aida** and **Carmen**, but not including **Aida** and **Carmen**.

Boosting Terms

Lucene provides the relevance level of matching documents based on the terms found. To boost a term, use the caret symbol (**^**) and a numerical boost factor at the end of the term you are searching. The boost factor must be a positive number. Its default value is 1 but it can be less than 1 (for example, 0.2). The higher the boost factor, the more relevant the term will be. In the case of a boost value less than 1, the term's relevancy is lower than the default.

Boosting lets you control the relevance of a document by boosting its term. For example, if you are searching for **jakarta apache** and want the term **jakarta** to be more relevant, boost it by using the **^** symbol along with the boost factor next to the term.

For example:

```
jakarta^4 apache
```

makes documents with the term **jakarta** appear more relevant. You can also boost *phrase terms*, for example:

```
"jakarta apache"^4 "Apache Lucene"
```

Synonym-Based Query Expansion

While the Search Query Analysis project type gives users the option of applying the NCI Thesaurus or MeSH synonym list to expand their query, the Lucene Query project type does not. Rather, users should specify all permutations of phrase synonyms they expect to be applied, or use wildcards where appropriate.

Medical Subject Heading (MeSH) search

Use the field **mesh_tags** to perform a search of the MeSH headings applied to a given article. The following syntax can be used:

Lucene syntax	Query description
mesh_tags:"YYY/ZZZ"	Requires the exact descriptor/qualifier string YYY/ZZZ.
mesh_tags:/YYY\/.*/	Requires the exact descriptor YYY with any qualifier (must be at least one qualifier).
mesh_tags:/YYY.*/	Requires the string YYY at the beginning of the descriptor, with or without any qualifier.
mesh_tags:/YYY/	Requires the exact descriptor YYY (without a qualifier).
mesh_tags:YYY	Requires the exact descriptor YYY (without a qualifier).

mesh_tags:”YYY ZZZ”	Requires the exact descriptor YYY ZZZ (without a qualifier).
mesh_tags:/YYY/ OR mesh_tags:/YYY\/.*/	Requires the exact descriptor YYY, with or without any qualifier.

Index Fields

The following table describes the searchable fields within the index. These fields are the same for all customers. Use the field names in the search box or API to filter your results appropriately.

Field Names	Type	Description
abstract	String, case insensitive	Abstract of the article, if it exists.
citationsText	String, case insensitive	Citation or reference section of an article.
content or **text**	String, case insensitive	Full text of the article; generally excludes citations.
create_date	Date	Date of article when loaded into index. Format is: yyyy-mm-dd 2014-06-01
documentId_doi	String, case insensitive	DOI of the article.
documentId_medlineId	String	PubMed ID (PMID) of the article, sourced from MEDLINE.
documentId_pii	String	Publisher item ID of the article.
documentId_pmcid	String, case sensitive	PubMed Central ID of the article, sourced from the publisher.
documentId_pmid	String	PubMed ID (PMID) of the article, sourced from the publisher.
documentId_pubmedcentralId	String, case insensitive	PubMed Central ID of the article, sourced from PubMed Central.
keywords	Array of String	Subject keywords of the article.
mesh_tags	Array of String, case insensitive	Medical Subject Heading (MeSH) tags of the article. Search single or phrase terms to return articles with particular descriptors; enclose full descriptor/qualifier strings in quotes as follows to search for these exactly – "[descriptor]/[qualifier]".
metadata_authors or **author**	String, case insensitive	Author(s) of the article.
metadata_endPage or endPage	Integer	End page of the article in the journal.
metadata_issn or **issn**	String, case sensitive	ISSN of the journal containing the article.
metadata_issue or **num**	Integer	Issue of the journal containing the article.
metadata_journal or **journal**	String, case insensitive	Name of the journal containing the article.

metadata_medlinepubtype	Array of string	Publication type of the article, sourced to MEDLINE. Valid values are enumerated by NLM as part of the MeSH vocabulary, and can be found here: https://www.nlm.nih.gov/mesh/pubtypes.html
metadata_startPage or startPage	Integer	Start page of the article in the journal.
metadata_title or title	String, case insensitive	Title of the article.
metadata_volume or vol	Integer	Volume of journal containing the article.
publicationDate or date	Date	Publication Date of the article. Format is: yyyy-mm-dd 2014-06-01
publication_year	Integer	Year of publication. Format is: YYYY 2014
publisherDocumentId or docid	String, case sensitive	Canonical document ID for the article.
publisherDocumentType or doctype	String, case sensitive	Canonical document ID type for the article.
publisherId or pid	String, case sensitive	System ID for the publisher. Valid values are enumerated in an appendix of this document.
section_conclusion	String, case insensitive	Conclusion section of the article. Note: Only some articles have clearly marked section information.
section_introduction	String, case insensitive	Introduction section of the article. Note: Only some articles have clearly marked section information.
section_materials_and_methods	String, case insensitive	Materials and Methods section of the article. Note: Only some articles have clearly marked section information.
substances_tags	String, case insensitive	Chemical substance tags of the article, sourced from MEDLINE.

PublisherID Valid Values

The following table describes the valid values for the PublisherID field.

Value	*Description*
acs_TDM	Americal Chemical Society
alphamed_TDM	AlphaMed Press
ama_TDM	American Medical Association
amdiabetes_TDM	Amer. Diabetes Assoc.
annualreviews_TDM	Annual Reviews
asco_TDM	American Society of Clinical Oncology
asm_TDM	American Soc. For Microbiology
asn_TDM	American Society for Nutrition
aspet_TDM	Association of Pharm Thera
ats_TDM	American Thoracic Society
bmj_TDM	BMJ
coaction_TDM	Co-Action Publishing
cob_TDM	Company of Biologists
cup_TDM	Cambridge University Press
endo_TDM	Bioscientifica
ers_TDM	European Respiratory Society
faseb_TDM	Fed. Of Am. Soc. of Exp. Biology
futmed_TDM	Future Medicine
futsci_TDM	Future Science
georgthieme_TDM	Georg Thieme Verlag KG
hindawi_TDM	Hindawi Publishing
ieee-per_TDM	IEEE
inderscience_TDM	Inderscience
ios_TDM	IOS Press B.V.
karger_TDM	Karger
ma_healthcare_TDM	MA Healthcare Limited
maney_TDM	Maney Publishing
medline_TDM	MEDLINE
microbiology-society_TDM	Microbiology Society
nas_TDM	National Academy of Sciences
nature_TDM	Nature Publishing Group
oxford_TDM	Oxford University Press
plos_TDM	PLOS
portland_TDM	Portland Press
rcn_TDM	R C N Publishing
rsc_TDM	Royal Society of Chemistry
rup_TDM	Rockefeller University Press
sage_TDM	Sage Publications
slack_TDM	Slack Incorporated
springer_TDM	Springer Sci. and Bus. Media
taylorfrancis_TDM	Taylor & Francis
wdg_TDM	Walter de Gruyter
wiley_TDM	John Wiley & Sons
wsp_TDM	World Scientific Publishing

Updated: 11 January 2020