How to sort by Lucene.Net field and ignore common stop words such as 'a' and 'the'?

Asked 16 years, 3 months ago Modified 15 years, 4 months ago

Viewed 3k times

1

I've found how to sort query results by a given field in a Lucene.Net index instead of by score; all it takes is a field that is indexed but not tokenized. However, what I haven't been able to figure out is how to sort that field while ignoring stop words such as "a" and "the", so that the following book titles, for example, would sort in ascending order like so:

1. The Cat in the Hat
2. Horton Hears a Who

Is such a thing possible, and if yes, how?

I'm using Lucene.Net 2.3.1.2.

lucene lucene.net

Share

Improve this question

Follow

asked Sep 15, 2008 at 19:37

Peaeater

636 • 5 • 19

5 Answers

Sorted by: Highest score (default)

1

I wrap the results returned by Lucene into my own collection of custom objects. Then I can populate it with extra info/context information (and use things like the highlighter class to pull out a snippet of the matches), plus add paging. If you took a similar route you could create a "result" class/object, add something like a SortBy property and grab whatever field you wanted to sort by, strip out any stop words, then save it in this property. Now just sort the collection based on that property instead.

Share Improve this answer

Follow

answered Sep 15, 2008 at 20:42

Paul Mrozowski

6,734 ● 9 ● 37 ● 49

> I think that's how it would have to be done, yes. I do create a collection of custom objects with the Lucene results so it shouldn't be too hard. Thanks. – Peaeater Sep 16, 2008 at 1:51

0

When you create your index, create a field that only contains the words you wish to sort on, then when retrieving, sort on that field but display the full title.

Share Improve this answer

Follow

answered Sep 15, 2008 at 19:40

John Sheehan

78.1k ● 30 ● 161 ● 194

Well, that's the trick, isn't it? You can't sort by a tokenized field, and its the tokenizing that analyzes the field for stop words and punctuation, as I understand it. So how to strip those stop words but keep the field un-tokenized?
– Peaeater Sep 15, 2008 at 19:57

In your code, strip out the stop words. You'll have to maintain your own list. – John Sheehan Sep 16, 2008 at 15:47

0

It's been a while since I used Lucene but my guess would be to add an extra field for sorting and storing the value in there with the stop words already stripped. You can probably use the same analyzers to generate this value.

Share Improve this answer

Follow

answered Sep 15, 2008 at 19:41

David Thibault
8,736 3 39 51

0

There seems to be a catch-22 in that you must tokenize a field with an analyzer in order to strip punctuation and stop words, but you can't sort on tokenized fields. How then to strip the stop words without tokenizing?

Share Improve this answer

Follow

answered Sep 15, 2008 at 20:02

Peaeater
636 5 19

1 Don't rely on Lucene to strip them, do it yourself.
– John Sheehan Sep 16, 2008 at 15:48

0

For search, I found search lucene .net index with sort option link interesting to solve ur problem

Share Improve this answer

Follow

answered Jul 29, 2009 at 13:57

logiclabz