http://mail-archives.apache.org/mod\_mbox/lucene-dev/200110.mbox/%3c010901c15864$e77a7260$0201a8c0@orca%3e

From lucene-dev-return-255-apmail-jakarta-lucene-dev-archive=jakarta.apache.org@jakarta.apache.org Fri Oct 19 06:12:36 2001

Return-Path: <lucene-dev-return-255-apmail-jakarta-lucene-dev-archive=jakarta.apache.org@jakarta.apache.org>

Delivered-To: apmail-jakarta-lucene-dev-archive@jakarta.apache.org

Received: (qmail 82449 invoked by uid 500); 19 Oct 2001 06:12:35 -0000

Mailing-List: contact lucene-dev-help@jakarta.apache.org; run by ezmlm

Precedence: bulk

Reply-To: lucene-dev@jakarta.apache.org

list-help: <mailto:lucene-dev-help@jakarta.apache.org>

list-unsubscribe: <mailto:lucene-dev-unsubscribe@jakarta.apache.org>

list-post: <mailto:lucene-dev@jakarta.apache.org>

Delivered-To: mailing list lucene-dev@jakarta.apache.org

Received: (qmail 82417 invoked from network); 19 Oct 2001 06:12:35 -0000

Message-ID: <010901c15864$e77a7260$0201a8c0@orca>

From: "Maurits van Wijland" <m.vanwijland@quicknet.nl>

To: <lucene-dev@jakarta.apache.org>

References: <3BCF4393.9060201@earthlink.net>

Subject: Re: Solution for "unstemming" terms

Date: Fri, 19 Oct 2001 08:11:24 +0200

MIME-Version: 1.0

Content-Type: text/plain;

charset="iso-8859-1"

Content-Transfer-Encoding: 7bit

X-Priority: 3

X-MSMail-Priority: Normal

X-Mailer: Microsoft Outlook Express 5.50.4133.2400

X-MimeOLE: Produced By Microsoft MimeOLE V5.50.4133.2400

X-Spam-Rating: daedalus.apache.org 1.6.2 0/1000/N

Dmitry,

Can you send us the code? This is very usefull!

I would like to experiment with this...

Maurits.

----- Original Message -----

From: "Dmitry Serebrennikov" <dmitrys@earthlink.net>

To: <lucene-dev@jakarta.apache.org>

Sent: Thursday, October 18, 2001 11:03 PM

Subject: Solution for "unstemming" terms

> I've found a pretty good solution for retrieving un-stemmed version of

> index terms, in case anyone is interested. This uses only the features

> already in 1.2-rc1 release.

>

> The trick is to create an additional field on each document (say "dict"

> for dictionary) and set it to contain a list of space-separated strings

> like this:

>

> cat:cats likeli:likeley

>

> And so on. So each term is composed of the stem, ':' and the unstemmed

> token. I had to create a custom Tokenizer that would split this string

> on spaces alone and not split the words at the ':' position. But there

> may be a different charachter that would work fine for one of the

> standard tokenizers.

>

> When you need to retrieve all unstemmed forms for a particular stem, you

> simply open up a TermEnum for a term <dict:stem:> like this:

> TermEnum te = reader.terms(new Term("dict", stem + ':'));

>

> The you just read the first one or all of the ones that startWith your

> stem. This works very fast because TermEnums are fast. You even get the

> unstemmed forms in a sorted order for free!

>

> - Dmitry

>

>