<http://mail-archives.apache.org/mod_mbox/lucene-dev/200110.mbox/%3C3BC48BCF.2020108@earthlink.net%3E>

From dmitrys@earthlink.net Wed Oct 10 17:56:45 2001

Return-Path: <dmitrys@earthlink.net>

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

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

Received: (qmail 97610 invoked from network); 10 Oct 2001 17:56:45 -0000

Received: from relay3.uswest.net (HELO relay1.uswest.net) (63.226.138.11)

by daedalus.apache.org with SMTP; 10 Oct 2001 17:56:45 -0000

Received: (qmail 86156 invoked by uid 0); 10 Oct 2001 17:56:46 -0000

Received: from unknown (HELO earthlink.net) (65.100.117.194)

by relay3.uswest.net with SMTP; 10 Oct 2001 17:56:46 -0000

Message-ID: <3BC48BCF.2020108@earthlink.net>

Date: Wed, 10 Oct 2001 11:56:31 -0600

From: Dmitry Serebrennikov <dmitrys@earthlink.net>

User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.0; en-US; rv:0.9.2) Gecko/20010726 Netscape6/6.1

X-Accept-Language: en-us

MIME-Version: 1.0

To: lucene-dev@jakarta.apache.org

Subject: Re: Token retrieval question

References: <4BC270C6AB8AD411AD0B00B0D0493DF0EE7C56@mail.grandcentral.com>

Content-Type: text/plain; charset=us-ascii; format=flowed

Content-Transfer-Encoding: 7bit

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

Doug, thanks for posting these. I may end up going in this direction in

the next few days and will use this as a blueprint. Maybe I'll end up

putting in the first pass implementation and then you can later further

tune it when you get to it.

Question on term numbers through: what would be an approach for merging

these across multiple IndexReaders for the purposes of MultiSearcher?

Doug Cutting wrote:

>Right now, Lucene does not have good support for what you're doing. Lucene

>as it stands is designed to support basic search, not other statistical text

>processing. However there are two features that I would like to add to

>Lucene that would help you.

>

[snip]

>

>This would add an IndexReader two methods:

> public TermFreqVector getTermFreqVector(int docNumber);

> public Term getTerm(int termNumber);

>The TermFreqVector class would be defined something like:

> public class TermFreqVector {

> public int[] getTermNumbers();

> public int[] getTermFrequencies();

> }

>The term number array would be sorted. The frequency of the term numbered

>getTermNumbers()[i] is getTermFrequencies()[i].

>

>Another class that would be useful is something like:

> public class TermWeightVector {

> public int[] getTermNumbers();

> public float[] getTermWeights();

>

> public void add(TermWeightVector other);

> public float distance(TermWeightVector other);

> }

>