Lucene term documents and term positions

Introduction

Term documents

For each term T, there are (doc frequency of the term) tuples of <doc ID, freq of T in this doc>.

This information is stored in the .frq file and accessible via the TermDocs interface.

Term positions

For each term T, there are (doc frequency of the term) tuples of <doc ID, freq of T in this doc, (term freq of T in this doc) counts of positions of T in this doc>.

This information is stored in the .prx file and accessible via the TermPositions interface.

Using the information

Via Query

If you use Query classes, they get and make use of term documents and term position so you do not have to worry about them. Non-span queries other than phrase queries use term documents only. Phrase queries uses term documents + term positions to make sure that a document actually have the terms, say, right next to each other and in order. This makes phrase queries slower than term queries, i.e. searching for the phrase “southern california” show be slower than searching for required words “southern california”.

Lower-level than queries is the spans API. The Spans class is still higher-level than using TermPositions directly.

Via the interfaces

Note that the document number (returned by doc()) and frequency (returned by freq()) of a TermDocs object is undefined until next() is called the first time. This is unclear from the API reference but I have found this out by experiment.

Revision #1
Created Fri, Aug 29, 2008 4:57 PM by Chan, Wing Kai
Updated Fri, Aug 29, 2008 4:57 PM by Chan, Wing Kai