



Indri is a C++ text search engine, built to handle large collections of structured documents. It combines language modeling estimates with the inference network framework to create a powerful query language that is capable of expressing detailed information needs.

Indri is capable of running simple disjunctive bag-of-words queries, like this:

```
white house
```

It can also incorporate term proximity and document structure, as in the following query. This query looks for the exact phrase 'white house', that was tagged as a 'place' by a named entity tagger.

```
#1(white house).place
```

Indri can also retrieve passages. This query (intended for a corpus of Shakespeare plays) looks for lines spoken by Juliet where she mentions Romeo. This query looks for passages of text surrounded by 'speech' tags.

```
#combine[speech]( juliet.speaker romeo )
```

Indri can also combine many sources of evidence in a weighted combination, like this query, which assigns 5 times the weight to the phrase 'white house' as it does to the term 'washington'.

```
#weight( 5.0 #1( white house ) 1.0 washington )
```

The system is efficient, and flexible enough to support a many retrieval tasks. For more information, or to use our online demo, please visit our website.

<http://www.lemurproject.org/indri>

The Lemur project is a collaboration between the University of Massachusetts, Amherst and Carnegie Mellon University, and is supported by ARDA and NSF.

# INDRI

## Parsing

Parses PDF, XML, HTML, TREC, Word, PowerPoint

Supports UTF-8 encoded text

## Indexing and retrieval

Multithreaded

Supports indexing and retrieval simultaneously

Supports retrieval across a cluster of machines

## Query language

Similar to Inquery

Passage retrieval

Structured document retrieval

Term proximity

Weighted combinations

## Distribution

Usable from C++, Java, or PHP

Works on Linux, Windows, Solaris or Mac OS X

Includes command-line tools and a Java GUI

Open source, BSD license