

Rosette for Solr-Based Applications

Build Cost-Effective, Multilingual, Search-Based Applications Using Open-Source Components

The same multilingual text analysis technology used by such leading enterprise and web search engines as Google, Microsoft, and Yahoo is now available to the rapidly-growing community of software developers building applications based on Apache Solr and Apache Lucene.

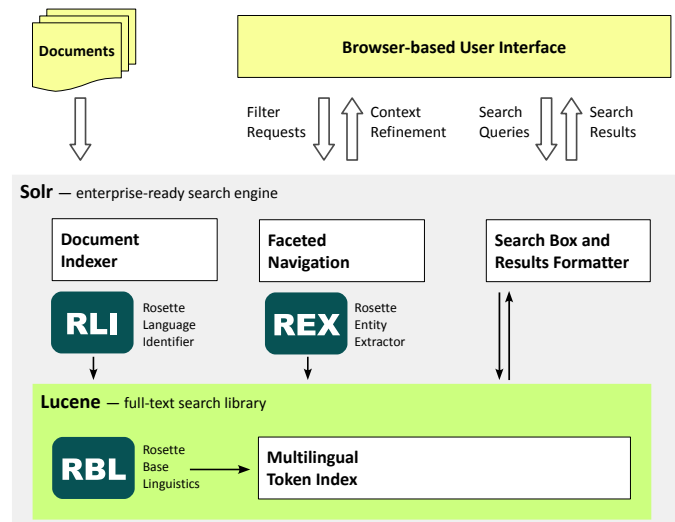
Apache Solr is an open-source, enterprise-ready search server offering XML/HTTP and JSON APIs; hit highlighting; faceted search; caching; replication; RDBMS integration; and a web administration interface. A key component of Solr is Apache Lucene, an open-source information retrieval toolkit originally written in Java but now also available for other programming languages including Delphi, Perl, C#, C++, Python, Ruby, and PHP. Lucene indexes are portable across platforms, making it easy to leverage advances in hardware and operating systems while minimizing additional development costs for faster and better search functionality.

Today, Solr and Lucene are powering thousands of large-scale search installations at such organizations as CNET, IBM, Netflix, and Wikipedia.



GETTING STARTED WITH ROSETTE & SOLR

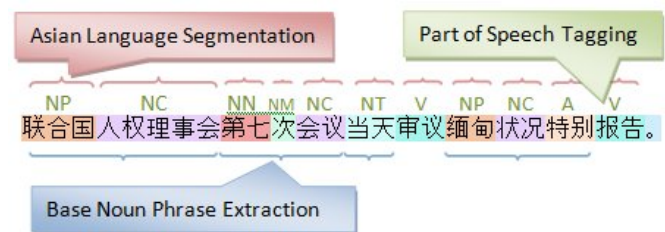
Rosette is designed to quickly connect to a new or existing project based on Solr, enabling access to robust and accurate multilingual search in days or hours rather than weeks or months. To get started, simply download and install the Rosette SDK or runtime package. The Rosette SDK enables advanced multilingual processing of text fields with only minor configuration changes. No additional effort is required for Solr to search documents containing text in any of the languages supported by Rosette.



ENABLING MULTILINGUAL SEARCH

Rosette provides convenient access to the core linguistic capabilities needed to implement a multilingual search-based application, including:

- **Language Identification**—Automatically classify documents by language or encoding.
- **Segmentation/Tokenization**—Determine the boundaries of lexical tokens (including punctuation and special characters) within the input stream.
- **Lemmatization**—Derive the dictionary base form from the inflected form of a verb or adjective.
- **Noun Decomposition**—Divide compound nouns into individual components to enable flexible information retrieval.
- **Part-of-Speech Tagging**—Classify words in an input stream according to grammatical function, such as noun, verb, or preposition.



GOING BEYOND SEARCH

Rosette offers the most complete collection of advanced linguistic capabilities within a single platform, including:

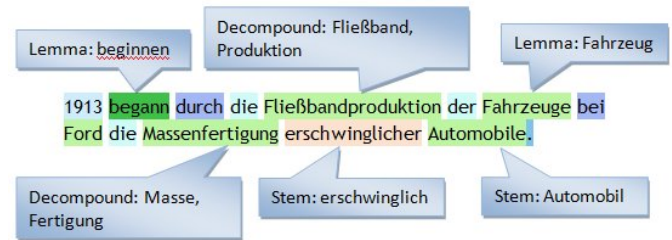
- **Language Boundary Location**—Identify the boundaries of each region of text in a distinct language so that the appropriate analyzer can be focused on each region.
- **Sentence Boundary Location**—Identify the boundaries of individual sentences within each language region.
- **Noun Phrase Extraction**—Identify the sequence of words surrounding a single noun which comprise a phrase.
- **Stop Word Removal**—Reduce index size by eliminating “noise” words.
- **User-defined Dictionaries**—Augment standard dictionaries with lists of specialized terms.
- **Chinese Script Conversion**—Implement pan-Chinese search by translating queries between simplified and traditional forms. Conversion engine is capable of handling variations at both the character level and the compound level
- **Japanese Orthographic Normalization**—Implement advanced Japanese search by recognizing and normalizing orthographic variations

Rosette incorporates a variety of algorithms so that the best approach can be applied depending upon the requirements of the language being analyzed. A combination of lexical data, heuristic rules, and statistical models are used to achieve a balance between speed and accuracy for each application.

SOLR PERFORMANCE & SCALABILITY

Solr offers a wide range of capabilities and benefits previously available only in expensive, proprietary, full-text search engines:

- Cross-platform—Windows, Linux, Unix, MacOS
- Low memory requirements
- Incremental indexing as speedy as batch indexing
- Index size of 20% to 30% of original text size
- Powerful search and ranking algorithms



LANGUAGES SUPPORTED

A single, uniform, tightly-coupled programming interface is used for indexing documents in any supported language:

Arabic	Italian
Chinese (Simplified)	Japanese
Chinese (Traditional)	Korean
Czech	Norwegian
Danish	Pushto
Dutch	Polish
English	Portuguese
Farsi/Dari	Russian
French	Spanish
German	Swedish
Greek	Thai
Hebrew	Turkish
Hungarian	Urdu

SYSTEM PLATFORMS SUPPORTED

Software development kits (SDKs) are available for popular architectures, operating systems, and development environments. Support for platforms not listed below is available by contacting your sales representative.

AIX 5.3/6.1, PPC
 HP-UX 11i, PA-RISC/IA64
 Linux CentOS 4.x/5.x, IA32/AMD64 (GCC 4.1/4.2)
 Linux Debian 5.0, IA32/AMD64
 Linux Red Hat 3.0, IA32/AMD64 (GCC 3.2/3.4/4.0)
 Linux Red Hat 4.0, IA32/AMD64 (GCC 3.4/4.0)
 Linux Red Hat 5.0, IA32/AMD64 (GCC 4.1/4.2)
 Linux Ubuntu 8.04/9.x/10.x, IA32/AMD64
 MacOS (GCC 4.0)
 Solaris 9, SPARC32/64
 Solaris 10, SPARC32/64
 Solaris 10, IA32/AMD64
 Windows XP/Vista/7, IA32 (MSVC 7.1)
 Windows XP/Vista/7, IA32/AMD64 (MSVC 8.0)

FOR MORE INFORMATION

For more information, please visit www.basistech.com. To request an evaluation copy, please write to info2010@basistech.com or call us at 617-386-2090 or 800-697-2062.



One Alewife Center
Cambridge, MA 02140

171 Second Street
San Francisco, CA 94105

13800 Coppermine Road
Herndon, VA 20171

9-6 Nibancho, Chiyoda-ku
Tokyo 102-0084