

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, Bing, and Yahoo! is now available to the rapidly-growing community of software developers building applications based on Apache Solr and Apache Lucene.

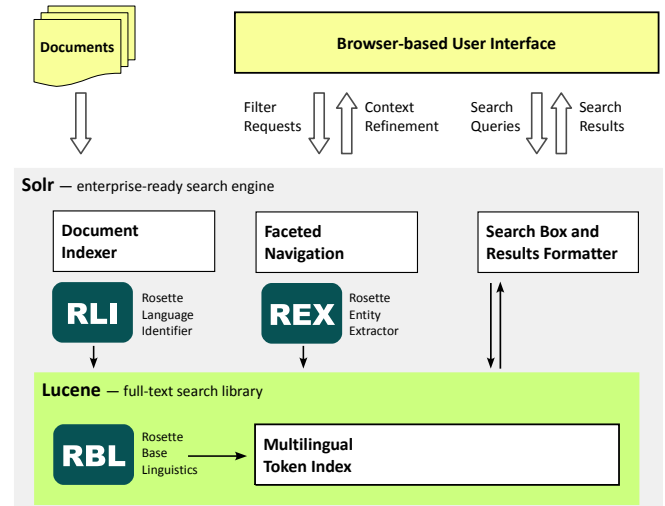
Apache Solr is an open-source, enterprise-ready search server offering XML/HTTP and JSON APIs; hit highlighting; faceted search; caching; replication; RDBMS integration; and a web administration interface. A key component of Solr is Apache Lucene, an open-source information retrieval toolkit. Lucene 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

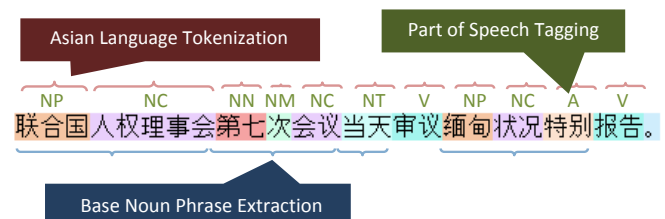
Rosette is designed to quickly connect to a new or existing project based on Solr, enabling access to robust and accurate multilingual search in days or hours rather than weeks or months. To get started, simply download and install the Rosette SDK or runtime package. The Rosette SDK enables advanced 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.
- **Lemmaization**—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.
- **Entity Extraction**—Locate key semantic concepts, including people, locations, and organizations, which can be used for faceted navigation.



GOING BEYOND SEARCH

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

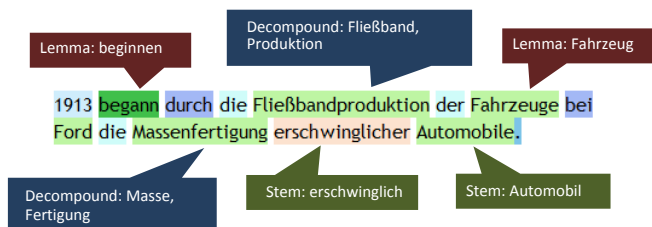
- **Language Boundary Location**—Identify the boundaries of each region of text in a distinct language so that the appropriate analyzer can be focused on each region.
- **Sentence Boundary Location**—Identify the boundaries of individual sentences within each language region.
- **Noun Phrase Extraction**—Identify the sequence of words surrounding a single noun which comprise a phrase.
- **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 language and capability.

APACHE 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
- Efficient memory usage
- Fast batch or incremental indexing
- Powerful search and ranking algorithms
- Highly scalable architecture
- Enterprise search features, like faceted navigation, hit highlighting, and replication



LANGUAGES SUPPORTED

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

Albanian	German	Polish
Arabic	Greek	Portuguese
Bulgarian	Hebrew	Romanian
Catalan	Hungarian	Russian
Chinese	Indonesian	Serbian
Croatian	Italian	Slovak
Czech	Japanese	Slovenian
Danish	Korean	Spanish
Dutch	Latvian	Swedish
English	Malay	Thai
Estonian	Norwegian	Turkish
Finnish	Pashto	Ukrainian
French	Persian	Urdu

SYSTEM PLATFORMS SUPPORTED

Software development kits (SDKs) and web services are available for the platforms listed below. Contact your sales representative for additional platform support.

AIX 6.1, PPC	Linux Ubuntu 10.x/11.x, IA32/AMD64
HP-UX 11i, IA64	MacOS
Linux CentOS 4.x/5.x, IA32/AMD64	Solaris 10, SPARC32/64, IA32/AMD64
Linux Debian 5.x, IA32/AMD64	Windows XP/Vista/7, IA32/AMD64
Linux Red Hat 4.x/5.x, IA32/AMD64	Windows Server 2003, 2008

EXPLORE FURTHER

For more information or to request an evaluation copy, please call us at 617-386-2090 or 800-697-2062, or write to info2011@basistech.com. We will be happy to assist you in evaluating the performance of our product on your data.

VISIT www.basistech.com WRITE info2011@basistech.com CALL 617-386-2090

One Alewife Center
Cambridge, MA 02140

171 Second Street
San Francisco, CA 94105

2553 Dulles View Drive
Herndon, VA 20171

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

