

Rosette for Search-Based Applications

Highly Accurate Text Analysis for Search in Asian, European, and Middle Eastern Languages

Full text search is ubiquitous. We access search engines daily on the Internet, in the office, on our home computers, and on portable devices. These products make it very easy to find information, but the technology they use internally is far from simple. Inside each search engine are sophisticated algorithms known as “computational linguistics”—software which analyzes digital text to enable it to be rapidly stored, searched, and retrieved.

Since 1998, the most widely used Internet and enterprise search engines have relied on Rosette® for essential linguistic and semantic enhancement, including segmentation, lemmatization, decompounding, part-of-speech tagging, sentence boundary detection, and noun phrase extraction. With these capabilities as the foundation, our customers are setting the pace in their own markets.

“Google selected Basis Technology to provide the Asian linguistic technology needed to create the ultimate Chinese, Japanese, and Korean search engine.”

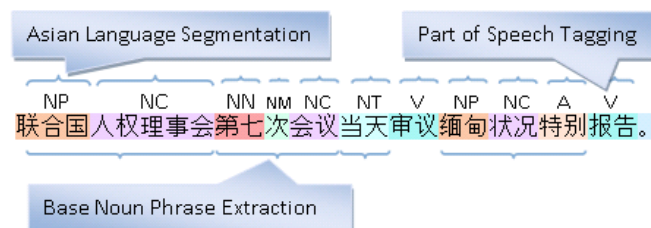
— Urs Hölzle, Fellow and Vice President, Google

The same linguistic technology which powers AOL, Ask, Google, Bing, and Yahoo is now available for search engines of all sizes and budgets. Rosette supports 19 languages, with new languages being added:

Arabic	German	Pashto
Chinese	Greek	Polish
Czech	Hungarian	Portuguese
Dutch	Italian	Russian
English	Japanese	Spanish
Farsi/Dari	Korean	Urdu
French		

THE ROSETTE SOLUTION

Rosette is designed to use a variety of different algorithms so the best approach can be applied for each language’s specific requirements. Depending on the language, a combination of lexical data, heuristic rules, and statistical models are implemented to provide the best accuracy and speed for all applications.



KEY FEATURES

Rosette provides the most advanced capabilities commercially available, whether for searching within a language or across multiple languages. Base features include:

- **Language Identification**—automatically classifies documents and messages by language and encoding.
- **Segmentation/Tokenization**—determines the boundaries of the unique lexical tokens in input data, including locating punctuation, and other special characters.
- **Lemmatization**—generates the dictionary base form for an inflected form of a verb or adjective.
- **Noun Decompounding**—divides compound nouns into sub-compounds for accurate information retrieval.
- **Part-of-Speech Identification**—tags a word’s part-of-speech such as noun, verb, or preposition.

ENHANCED SEARCH FEATURES

- **Sentence Boundary Detection**—Marks boundaries of individual sentences.
- **Base Noun Phrase Analysis**—identifies sets of words including a noun which describe a single expression.
- Ignores user-defined **stop words**.
- Supports customer-provided **dictionaries** to allow an application-specific vocabulary.
- **Language Boundary Locator**—identifies multiple language regions within a single document so individual languages can be processed and routed properly.
- **Chinese Script Converter**—processes Chinese text and converts between Simplified and Traditional forms, handling both the character variations and the word-level differences.

- **Japanese Orthographic Normalizer**—Normalizes different orthographic forms of Japanese words to a standard canonical form.

ROSETTE IN YOUR APPLICATION

Rosette is a comprehensive linguistic platform ideal for any application which must process large volumes of multilingual text, including:

- Enterprise Search Engines
- Web Search Technology
- Apache Lucene and Solr Solutions
- Information Access Platforms
- E-Discovery and Digital Forensics
- Document and Media Exploitation
- dtSearch Solutions
- Email and Instant Messaging

ROSETTE COMPONENTS

Rosette is a single API that provides access to the various linguistic capabilities described above. Search solutions typically use the following Rosette components:

- Rosette Base Linguistics (RBL)
- Rosette Language Identifier (RLI)
- Rosette Language Boundary Locator (RLBL)
- Rosette Core Library for Unicode (RCLU)

SYSTEM SPECIFICATIONS

Rosette is a portable and highly scalable software developer kit (SDK) that runs on platforms ranging from laptop PCs to multi-CPU servers processing thousands of documents per second.

A fully-documented API is provided and may be accessed from applications written in C, C++, Java, and other languages. A command-line interface is also available for testing purposes.

SDKs are available for Apple MacOS, Microsoft Windows, Sun Solaris, and multiple Linux distributions.

FOR MORE INFORMATION

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



Browse
www.basistech.com

Write
info2010@basistech.com

Call
617-386-2090 or 800-697-2062

Boston
One Alewife Center, Cambridge, MA 02140

San Francisco
171 Second Street, San Francisco, CA 94105

Washington, D.C.
13800 Coppermine Road, Herndon, VA 20171