

Rosette Language Identifier

Automatically Classify Documents and Messages by Language and Encoding

Rosette® Language Identifier (RLI) is a high-performance, cross-platform software engine which scans unidentified or untagged documents or messages to determine both the written language and character encoding scheme with very high accuracy. Identification of language and encoding are necessary prerequisites for applications which properly categorize, search, process, and store text in any written language.

RLI utilizes proprietary algorithms for statistical language analysis, together with information-rich profiles derived from large, hand-verified corpora covering all of the supported languages.

APPLICATIONS

RLI is essential for any application which processes large volumes of multilingual text, including:

- Web Search Engines
- Enterprise Search Engines
- Information Access Platforms
- E-Discovery and Digital Forensics
- Document and Media Exploitation
- Data Mining and Data Warehousing
- E-mail and Instant Messaging
- Database Migration

When used in conjunction with a transcoding engine (such as the Rosette Core Library for Unicode), RLI provides a powerful and immediate solution for converting a large, heterogeneous collection of text into a single, uniform representation based on the Unicode standard.

BENEFITS

Although modern text encoding standards such as XML mandate the use of Unicode, many existing applications, documents, and data streams utilize so-called "legacy encodings," such as ASCII, ISO 8859-1 (also known as "Latin 1"), Shift-JIS, and countless others.

The Multipurpose Internet Mail Extensions (MIME) standards define an approved list of legacy encodings to be utilized with e-mail, newsgroups, and HTML documents. However, in practice, these standards are inconsistently observed by software developers and website creators. To operate successfully in a real-world environment, multilingual applications *must* be prepared to accept documents which are

either incorrectly tagged as to their character encoding or not tagged at all.

RLI also provides a natural solution to the problem of migrating data repositories upward from older versions which support only legacy encodings to newer versions based on the Unicode standard.

TRANSLITERATED TEXT

Many languages may be written with more than one alphabet or writing system; RLI is trained to recognize some of these. For example, the Serbian language may be written in either the Cyrillic alphabet or in the Latin alphabet. As another example, RLI will recognize the following text as Arabic:

لَمَّا كَانَ الْاِعْتِرَافُ بِالْكَرَامَةِ الْمَتَأَصِّلَةِ وَالْحَقُوقِ الْمَتَسَاوِيَةِ الثَّابِتَةِ
هُوَ اَسَاسُ الْحُرِيَةِ وَالْعَدْلِ وَالسَّلَامِ فِي الْعَالَمِ، فِانِ الْجَمْعِيَّةِ الْعَامَّةِ
تَتَنَادَى بِالْاِعْلَانِ الْعَالَمِيِّ لِحَقُوقِ الْاِنْسَانِ.

RLI will also recognize the following text as Arabic:

lamma kan aliiii'tiraf bi alkaramat al-muta'assilat wa bi
alhuquq al-mutasawiyat al-thabitat hu asas al-hurriyyat
wa'al'adl wa'alssalam fi al-'alim, fa inn al-jam'iyyah
al-'ammah tunadi bi alii'lan al-'alimi lihuquq al-insan

SYSTEM SPECIFICATIONS

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.

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.

LANGUAGES AND ENCODINGS SUPPORTED

RLI supports 188 language/encoding pairs, covering 55 languages, 7 Latin script variants (transliterations), and 39 legacy encodings, plus the modern UTF-8 encoding for every language.

Albanian — ISO 8859-1, Windows-1252	Lithuanian — ISO 8859-13, Windows-1257
Arabic — ISO 8859-6, Windows-720, Windows-1256	Malay — ISO 8859-1, Windows-1252
Arabic (transliterated) — ISO 8859-1, Windows-1252, Windows-1256	Malayalam — ISCII-Malayalam
Bengali — ISCII-Bengali	Norwegian — ISO 8859-1, Windows-1252
Bulgarian — ISO 8859-5, Windows-1251, KOI8-R	Pushto — ISO 8859-6, Windows-1256
Catalan — ISO 8859-1, Windows-1252	Pushto (transliterated) — ISO 8859-1, Windows-1252
Chinese, Simplified — GB-2312, HZ-GB-2312, GB-18030, ISO 2022-CN	Persian — Windows-1256
Chinese, Traditional — Big5, Big5-HKSCS	Persian (transliterated) — ISO 8859-1, Windows-1252, Windows-1256
Croatian — Windows-1250	Polish — ISO 8859-2, Windows-1250
Czech — ISO 8859-2, Windows-1250	Portuguese — ISO 8859-1, Windows-1252
Danish — ISO 8859-1, Windows-1252	Romanian — ISO 8859-2, Windows-1250
Dutch — ISO 8859-1, Windows-1252	Russian — ISO 8859-5, Windows-1251, KOI8-R, IBM-866, Mac Cyrillic
English — ISO 8859-1, Windows-1252	Serbian — ISO 8859-5, Windows-1251
Estonian — ISO 8859-13, Windows-1257	Serbian (transliterated) — ISO 8859-2, Windows-1250
Finnish — ISO 8859-1, Windows-1252	Slovak — Windows-1250
French — ISO 8859-1	Slovenian — Windows-1250
German — ISO 8859-1, Windows-1252	Somali — ISO 8859-1, Windows-1252
Greek — ISO 8859-7, Windows-1253	Spanish — ISO 8859-1, Windows-1252
Gujarati — ISCII-Gujarati	Swedish — ISO 8859-1, Windows-1252
Hebrew — ISO 8859-8, Windows-1255	Tagalog — ISO 8859-1, Windows-1252
Hindi — ISCII-Hindi	Tamil — ISCII-Tamil
Hungarian — ISO 8859-2, Windows-1250	Telugu — ISCII-Telugu
Icelandic — ISO 8859-1, Windows-1252	Thai — Windows-874
Indonesian (Bahasa Indonesia) — ISO 8859-1, Windows-1252	Turkish — ISO 8859-9, Windows-1254
Italian — ISO 8859-1, Windows-1252	Ukrainian — ISO 8859-5, Windows-1251, KOI8-R
Japanese — EUC-JP, ISO-2022-JP, Shift-JIS, Shift-JIS-2004 (JIS X 0213)	Urdu — ISO 8859-6, Windows-1256
Kannada — ISCII-Kannada	Urdu (transliterated) — ISO 8859-1, Windows-1252
Korean — EUC-KR, ISO-2022-KR	Uzbek — ISO 8859-5, Windows-1251, KOI8-R
Kurdish — Windows-1256	Uzbek (transliterated) — Windows-1251
Kurdish (transliterated) — ISO 8859-1, Windows-1252, Windows-1256	Vietnamese — TCVN, VIQR, VISCII, VNI, VPS
Latvian — ISO 8859-13, Windows-1257	

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



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