ROSETTE
Language Identifier

BASIS TECHNOLOGY

www.basistech.com
info@basistech.com
+1 617-386-2090

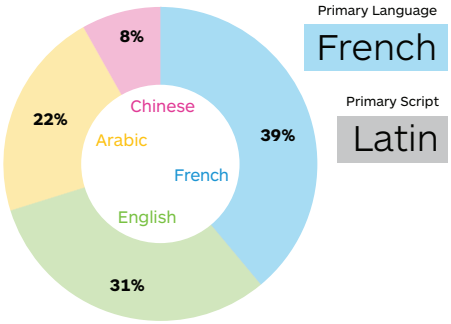
Instantly identify and triage many languages within large volumes of text.

English

即时识别和处理大量多语言文本。 Chinese

Identifiez et triez instantanément plusieurs langues à travers de nombreux textes. French

التحديد والتصنيف الفوري للعديد من اللغات ضمن كميات كبيرة من النصوص. Arabic



Identify languages and transform encodings

55 Supported Languages

Rosette® Language Identifier (RLI) analyzes text from a few words to whole documents, to detect the languages and character encoding with speed and very high accuracy. Automatic language identification is the necessary first step for applications that categorize, search, process, and store text in many languages. Individual documents may be routed to language specialists, or sent into language-specific analysis pipelines (such as Rosette Base Linguistics) to improve the quality of search results.

For applications that analyze tweets, search keywords, and other short text, RLI offers market-leading accuracy for language detection given 1-3 words (<20 bytes) up to a full sentence.

RLI achieves its incredible accuracy through the use of proprietary algorithms with information-rich language profiles derived from statistical analysis, in addition to language-specific methods for short text language detection. Basis Technology continually improves the Rosette product family with language additions, feature updates, and the latest innovations from the academic world.

KEY FEATURES

- Simple API
- Fast and scalable
- Industrial-strength support
- Easy installation
- Flexible and customizable
- Java or C++
- Unix, Linux, Mac, or Windows
- Component of the Rosette SDK

Select Customers

bing EMC² Turner StumbleUpon YAHOO! equivio

Start using RLI today
Try our free product evaluation
www.basistech.com

ROSETTE
Language Identifier

IDENTIFICATION FEATURES

- Identifies the primary or dominant language of a document
- Identifies the language scripts within the document, such as Latin and Cyrillic
- Determines the languages and their percentages within multilingual documents
- Works with texts that have been transliterated, such as Arabic chat that is written in the Latin script
- Accurate with short strings—from 1-3 words (<20 bytes) to a full sentence to enable full analysis of search queries, tweets, image captions, metadata, news headlines, email subject lines, and more.

LANGUAGE BOUNDARY LOCATOR

J'ai été surprise par cette surprise. Vice President Biden spoke about this in Munich. El carpintero prensa los bordes de la placa decorativa. Proper wound care management prevents die Geige gibt einen schoenen Laut von sich.

ENGLISH FRENCH GERMAN SPANISH

Digital text is often composed of multiple languages within the same document, presenting a challenge to computers and humans alike. RLI enriches the text with start and end markers for each language placed within multilingual documents—even if all the languages are written in the same script—such as English, French, German, or Italian. Boundaries of each writing system are also detected, such as Latin, Cyrillic, Japanese kana, or Chinese hanzi.

ENCODING CONVERSION

ISO-8859-1 → UNICODE

Although modern text encoding standards, such as XML, mandate the use of Unicode, many existing applications, documents, websites, and data streams use "legacy encodings," such as ASCII, ISO 8859-1, Shift-JIS, and many others.

Rosette accurately converts large collections of text with these legacy encodings into a single, uniform format in the Unicode standard. This converted text can then be used in any language, which eliminates data corruption and other problems due to incompatible code.

LANGUAGE AND ENCODING COMPATIBILITY

188 *Language/ Encoding Pairs*

55 *Languages with Unicode*

7 *Latin Script Variants (Transliterations)*

44 *Legacy Encodings*

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

Compatibility

Code Base: C++, Java

Platform Support: Linux, Red Hat, Windows, Mac

HEADQUARTERS
One Alewife Center
Cambridge, MA
02140

FEDERAL
2553 Dulles View Dr.
Suite 450
Herndon, VA
20171

WEST COAST
1700 Montgomery St.
San Francisco, CA
94111

EUROPE
Furzeground Way
Middlesex UB11 1BD,
UK

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