



ROSETTE
Entity Extractor



BASIS
TECHNOLOGY

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

Automatically find names of people, places, products, and organizations in text across many languages.

Accurate & adaptable statistical entity extraction

18 Supported Languages

Rosette® Entity Extractor (REX) delivers structure, clarity, and insight, by revealing the key information—names, places, organizations, products, and other words and phrases—lying hidden within large volumes of unstructured Big Text.

REX is the foundation for applications in eDiscovery, social media analysis, financial compliance, and government intelligence. The effectiveness of these mission-critical applications depend on REX for its accuracy, robustness, and ability to find entities across many languages.

By nature, statistically trained models are most accurate on the type of data they are trained on. Besides machine learning from a wide range of text beyond news articles, REX is unique among named entity recognition software in its adaptability. REX's field training mechanism enables you to add *your text* data to increase REX's accuracy on *your text*.

KEY FEATURES

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

Select Customers



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



How it works

STATISTICAL ENTITY EXTRACTION



Statistical modeling with advanced linguistics solves the three biggest challenges in entity extraction: finding entities which cannot be exhaustively listed, finding entities which are yet unknown, and using context to distinguish between similar entities, e.g., the place "Newton, MA" and the person "Isaac Newton".

Customized extraction

FIELD TRAINING FOR INCREASED ACCURACY

For users with text that is particularly challenging in format, style, or vocabulary, REX's unique field training capability has multiple mechanisms to adapt its statistical model to their data. Users just add a quantity of their data (unannotated or annotated), and rebuild the model for maximum accuracy.

PATTERN-MATCHING RULES



Rules expressed as regular expressions find entities which follow a pattern, such as dates, times, and email addresses. Many standard string patterns are included with REX; customers can customize by editing or adding their own rules, based on their specific needs.

CUSTOM ENTITY LISTS

Custom lists are helpful when users know that specific words or phrases in their data are almost never misspelled and always refer to the same thing (i.e., are unambiguous). REX comes with such lists for entity types like religions and nationalities.

Compatibility

Code Base



Platform Support



REX in action

The **New York Philharmonic Orchestra** will make a historic trip to **North Korea** in **February**, it has announced.

Dominique de Villepin a été nommé Premier ministre ce **mardi** en fin de matinée par **Jacques Chirac**.

The orchestra's **president** and **executive director**, **Zarin Mehta** said it would play in the capital **Pyongyang** on **February 26**. In **August**, the reclusive **communist** country's **Ministry of Culture** sent an invitation to the orchestra at **Lincoln Center** in **Manhattan**.

Person

Location

Organization

Date

Time

Title

朝鮮外務省发言人 11月1日 在平壤宣布, 朝鲜将重返六方会谈, 但前提条件是朝鲜与美国在六方会谈框架内讨论解除美国对朝鲜金融制裁问题。针对朝鲜方面的动向, 各方均表示欢迎。美联社 11月1日 报道说: “长期以来一直拒绝与平壤进行直接对话的美国总统布什认为, 各方达成一致、同意恢复六方会谈应归功于中国的斡旋。”

L'ancien **ministre de l'Intérieur**, qui n'a jamais participé à une élection, a déjeuné avec les **députés UMP** et **UDF** à l'invitation du **président de l'Assemblée nationale**, **Jean-Louis Debré**.

小澤征爾は、日本を代表する世界的な指揮者である。1973年、38歳のときに、アメリカ5大オーケストラの一つであるボストン交響楽団の音楽監督に就任した。

الخميس 5/2/1431 هـ – الموافق 21/1/2010 م (آخر تحديث) الساعة 10:01 (مكة المكرمة)، غرينتش) 7:01 (غرينتش) ناتو يفكر بمسؤول مدني لأفغانستان يخطط حلف شمال الأطلسي (ناتو) لتعيين مسؤول مدني كبير في أفغانستان، وسط دعوات لتحسين التنسيق السياسي والتنموي في البلاد وفق ما نقلته صحيفة **ول ستريت**

Predefined Entity Types

REX natively supports the following entity types. User-defined entities, such as SKU numbers, are also available.

- Person
- Location
- Organization
- Product
- Title
- Nationality
- Religion
- Credit Card Number
- Geographic Coordinate
- Money
- Generic Number
- Personal ID Number
- Phone Number
- Email Address/URL
- Distance
- Date
- Time

Available Languages

Additional languages are available through custom development.

- Arabic
- Chinese, Simplified
- Chinese, Traditional
- Dutch
- English
- French
- German
- Hebrew
- Italian
- Indonesian
- Japanese
- Korean
- Pashto
- Persian
- Portuguese
- Russian
- Spanish
- Urdu