

Rosette® アラビア語形態素解析システム

適用例

大量のアラビア語文書の解読および解析を必要とする組織・企業のために様々な特長があります。

- データベースおよびウェブの検索
- コンテンツ・フィルタリング
- ドキュメント要約・分類
- 索引／キーワード抽出
- OCR対応
- データマイニング
- 電子メールフィルタリング
- ドキュメント・オーサリング
- テキスト・マイニング

Basis Technology の Rosette® アラビア語形態素解析システム (RBL-AR: Rosette® Base Linguistics for Arabic) はポータブルで高性能なアラビア語テキスト単語分節エンジンです。ウェブページ、電子メール、データベースの文書などを対象とする、どのような種類の文書にも RBL-AR は威力を発揮します。

アラビア語の解析は決して容易なものではありません。アラビア語は一般的に、接辞と挿入辞を組み合わせることで動詞の相である、目的、活用、人称、数、性あるいは統語的属性を表現します。しかし、テキストの索引、キーワード抽出、またはその他のテキスト操作をする前に、それらを正規化しなくてはなりません。母音の不規則的な使用および欠如によるアラビア語の特有な表現も、すべて事前に正規化する必要があります。

Rosette® アラビア語形態素解析システムは、効果的な検索の妨げとなる文法的な接辞(動詞の語形変化、前置詞、代名詞など)を取り除くことによって、アラビア語の綴りを正規化し、ステミングします。また RBL-AR は、不規則複数形を単数形にするなど、複数形を変換させるために高度なコンピューター言語学アルゴリズムおよび特殊な辞書を利用しています。

特長

- 母音およびアラビア語特有な記号の削除、ハムザ(独立した子音)の統合、カシダ(アラビア語の引伸ばし記号)の削除など、綴りの正規化を行う
- 複数形を適切な単数形へと正規化する
- 接辞の削除など、基本的なステミングを行う
- ストップワードのユーザー定義が可能
- アラビア数字をラテン語の数字表記へと正規化する
- ユーザー定義辞書の使用

仕様

- 5,000 以上の不規則複数形を収録した辞書
- スタティック または ダイナミックライブラリ
- 完全な Unicode 内部アーキテクチャは、Microsoft の CP1256 および ISO8859-6、ASMO449、ASMO708 など、他のアラビア語文字コードでのトランスペアレントな入出力を行う
- 辞書をメモリーにマップし、必要なメモリー量を小さく抑え、素早い起動を実現
- C/C++API
- スレッドセーフ

対応プラットフォーム

SDK は以下のプラットフォームでご利用いただけます。

- Microsoft Windows
- Sun Solaris
- Linux
- MacOS

他のプラットフォームへの対応もご要望により承ります。

年間テクニカルサポート、アップグレード契約もあります。