Abstract

Stemmer (oder Lemmatisierer) spielen im IR (Information Retrieval) eine nützliche Rolle. Es gibt allerdings wenige Lemmatisierungsalgorithmen die frei verfügbar sind, vor allem für andere Sprachen als das Englische.

Ziel dieses Projektes ist einen Stemmer für das Rumänische mit *Snowball*[1] zu implementieren und zu evaluieren.

Snowball ist eine kleine String-verarbeitende Programmiersprache, die Porter speziell zur Definition von Stemmern geschrieben hat. Diese ermöglicht eine sehr genaue Definition von Stemmern, von denen dann schnelle Programme in ANSI C oder Java generiert werden können.

Das Projekt beinhaltet eine algorithmische Beschreibung des Stemmers, eine Implementation in Snowball und einen repräsentativen Wortschatz von etwa 30 000 Wörtern, der als Teil eines Standardtests benutzt werden kann.

Wir möchten uns dabei an Porters Stemmer für das Englische orientieren, müssen aber natürlich den Stemmer an die Besonderheiten des Rumänischen anpassen. Das heißt, Begriffe wie „Wort“, „Wortgrenze“ und „Buchstabe“ müssen definiert, Stopwörter identifiziert und unregelmäßige Formen zusammengefasst werden.
Mit Hilfe von eventuellen Datenbanken o.Ä. wollen wir erreichen, dass vom „RO-Stemmer“ als Ergebnis, bei Eingabe eines Wortes, der Stamm (und/oder das abgeleitete Wort) wiedergegeben wird.
Stoppwörter, Wörter die sich nicht oder kaum ändern, werden voraussichtlich nicht gestemmt, da dies wenig Sinn macht, und da sie z.T. auch nicht gestemmt werden können.

Bei der Eingabe muss man nicht auf diakritische Zeichen verzichten, da Snowball seit Mai 2005 auch den universellen Zeichensatz UTF-8 unterstützt.

[1] Entwickelt von Martin Porter (2001); Name wurde von SNOBOL abgeleitet.
Näheres unter: http://snowball.tartarus.org/