Extraktion von geographischen Bezeichnungen aus Webseiten

Betreuer

Christian Siefkes: siefkes@mi.fu-berlin.de

Ziele des Projekts

In dem Projekt soll ein System zur Erkennung und Extraktion von geographischen Bezeichnungen (Namen von Orten und Ländern, Postleitzahlen u.a. Adresselemente, evt. Namen von Flüssen u.ä.) in HTML-Seiten entwickelt werden. Die gefundenen Bezeichnungen sollen nach Typ erfasst und in einer Datenbank gespeichert werden, um so Suchanfragen anhand geographischer Kriterien zu ermöglichen („alle Webseiten, die sich mit Berlin beschäftigen“).

Zu beachten ist dabei insbesondere:

- Das System darf nicht auf vorgegebene Listen von Bezeichnungen beschränkt sein, sondern muss auch unbekannte Namen erkennen und einordnen können.
- Mehrdeutige Worte sollen nur dann extrahiert werden, wenn sie ein geographisches Objekt bezeichnen (Berlin, Essen als Städte), nicht in anderen Bedeutungen (die Person Irving Berlin, Essen als Aktivität).

Beide Anforderungen können im Einzelfall natürlich nur mit einer gewissen Wahrscheinlichkeit umgesetzt werden. Die Umsetzung kann auf dem statistischen Informationsextraktions-System *TiES* (`www.inf.fu-berlin.de/inst/ag-db/software/ties/`) und anderen geeigneten Bibliotheken und Tools aufbauen.

Das System sollte wahlweise für englische und/oder deutschsprachige Webseiten funktionieren.

Durchführung

Das Projekt soll von einer kleinen Gruppe durchgeführt werden. Start und Ende des Projektes können mit dem Betreuer flexibel vereinbart werden, genauso wie die regelmäßigen Projekttreffen.

Zu Begin des Projekts müssen geeignete Ressourcen (Listen von Ortsnamen etc.) gesucht und und aufbereitet werden (vgl. z.B. Getty Thesaurus <`www.getty.edu/research/conducting_research/vocabularies/tgn/`>, Wikipedia <`www.wikipedia.org/`>). Danach erfolgt Planung, Design und Implementierung des Systems. Zum Abschluss ist die Effektivität des Verfahrens zu evaluieren (Precision und Recall).

Die Implementierung sollte in Java erfolgen. Software und Dokumentation sollten unter einer freien Lizenz (GNU LGPL o.ä.) online veröffentlicht werden.

Maximale Teilnehmerzahl: 3

Voraussetzungen

Vordiplom, gute Kenntnisse von Algorithmen und Programmierung, Interesse an der Verarbeitung von Texten in natürlicher Sprache und an Information Retrieval.

Im Anschluss an das Projekt können Studien/Bachelor/Diplomarbeiten vergeben werden.