Korpusmanagement, Annotation und Analyse

semtracks gmbh

ingwer: Installationsanleitung

semtracks gmbh

16. April 2015

ingwer ist eine Software für Korpusmanagement, Annotation und Analyse von Texten. Die Software wurde von der semtracks gmbh (`www.semtracks.com`) als Webanwendung basierend auf MySQL und PHP entwickelt und unter der GNU GPL-Lizenz veröffentlicht.

In diesem Dokument wird erklärt, wie ingwer und die dazugehörigen Komponenten installiert werden.

1 Voraussetzungen

ingwer erfordert zwingend folgende Voraussetzungen:

1. Webserver mit installiertem PHP, Perl und MySQL-Datenbank.
2. Wortarten-Tagging-Software „TreeTagger“ mit der deutschen Library: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.
3. Für die Installation des TreeTaggers ist ein Shell-Zugang zum Server (z.B. über `ssh` notwendig).

Damit können alle internen Funktionen von ingwer benutzt werden. Es besteht aber die Möglichkeit, ingwer in Verbindung mit der Corpus Workbench (CWB) `http://cwb.sourceforge.net/` zu nutzen, um quantitative Analysen machen zu können. Es ist dann möglich, in ingwer indizierte Korpora oder Teile davon für die Corpus Workbench zu exportieren und auf dem Server über das Web-Interface der CWB, CQPWeb, zu nutzen. Dafür müssen die CWB und CQPWeb zusätzlich auf dem Server installiert werden. Installationanleitungen dazu finden sich unter folgenden Adressen:

- CWB: `http://cwb.sourceforge.net/download.php`
- CQPWeb: `http://cwb.svn.sourceforge.net/viewvc/cwb/gui/cqpweb/trunk/CQPweb-setup-manual.html`

2 Installation des TreeTaggers

Der `TreeTagger` besteht aus dem eigentlichen Tagger-Programm, bereits trainierten Sprachbibliotheken, einigen Bash-Scripts zur einfacheren Verwendung des Taggers in Verbindung mit Hilfsprogrammen wie einem Tokenzier und einem Trainer-Programm, mit dem neue Sprachbibliotheken auf der Basis getaggter Korpora erstellt werden können. Wir benötigen für ingwer alles bis auf das Trainer-Programm, das aber der Einfachheit halber auch gleich mitinstalliert wird.

Alle Dateien und die Installationanleitung findet sich unter:
`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
Im Detail gehen Sie wie folgt vor:

1. Download des Taggers für das passende Betriebssystem:[1]
   - PC-Linux: → Download
   - Mac OS X Intel: → Download

   Paket nach dem Herunterladen in ein Verzeichnis `TreeTagger` o. ä. entpacken.

2. → Tagging-Scripts in das gleiche Verzeichnis herunterladen.

3. Installationsscript herunterladen (ggf. mit Rechts-Klick *Ziel speichern unter...*): `ftp://ftp.ims.uni-stuttgart.de/pub/corpora/install-tagger.sh`.

4. Die gewünschten Sprachbibliotheken (Parameter-Dateien) herunterladen:
   - Für PC/Linux, Mac Intel: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/#Linux`.
   - Für Sparc-Solaris und Mac PowerPC: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/#Solaris`.

   Laden Sie im Minimum das *German parameter file (UTF-8)* herunter. Dieses ermöglicht das Tagging der Wortarten im Deutschen. Es gibt fürs Deutsche zusätzlich auch noch die *German chunker parameter*-Datei, die ein einfaches Parsing von NPs und VPs erlaubt. ingwer arbeitet allerdings nicht damit.

   Speichern Sie die Datei, dekomprimieren Sie sie und speichern Sie sie im `TreeTagger`-Verzeichnis im Verzeichnis `lib`. Falls es `lib` nicht gibt, erstellen Sie das Verzeichnis.

5. Öffnen Sie nun die Shell und starten Sie das Installationsscript im `TreeTagger`-Verzeichnis:

   ```
   sh install-tagger.sh
   ```

6. Sie können nun die Installation testen mit:

   ```
   echo "Das ist ein Test." | cmd/tree-tagger-german-utf8
   ```

1 Für weitere Systeme vgl. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

Es sollte dann ausgegeben werden:

```
Das PDS d
ist VAFIN sein
ein ART ein
Test NN Test
. $. .
```

3 Installation von ingwer über Installer

Das Perl-Script `install.pl` kann in vielen Situationen das gesamte ingwer-Paket installieren. Es müssen lediglich im Kopf der Datei einige Einstellungen, insbesondere über die gewünschte Datenbankstruktur, sowie Pfade zu den notwendigen Programmen, konfiguriert werden. Anschließend kann der Installer mit `sudo perl install.pl` gestartet werden.

4 Manuelle Installation von ingwer

4.1 Einrichten der Datenbank

4.1.1 Datenbank und Benutzer

ingwer benötigt eine MySQL-Datenbank und einen Benutzer, der darauf zugreifen kann. Datenbank und Benutzer können ggf. über eine Administrationsoberfläche des Servers eingerichtet werden. Legen Sie dann bitte eine Datenbank mit beliebiger Bezeichnung an und einen Benutzer, der in der Datenbank lesen, schreiben und ändern darf. Merken Sie sich Datenbankname, Benutzer und vergebenes Passwort.

Alternativ können Sie die SQL-Datei `SQLUserDB.sql` als Grundlage verwenden, um die Datenbank und den Benutzer einzurichten. Gehen Sie wie folgt vor:

1. Öffnen Sie die Datei `SQLUserDB.sql` in einem Texteditor.
2. Passen Sie folgende Werte an:
   a) `IngwerUser` (Zeilen 1 und 3): Setzen Sie einen beliebigen Benutzernamen ein.
   b) `IngwerPassword` (Zeilen 1 und 3): Setzen Sie ein beliebiges Passwort ein.
   c) `IngwerDB` (Zeile 5): Setzten Sie einen beliebigen Datenbanknamen ein.
3. Lesen Sie als MySQL-Root-Benutzer die Datei ein mit dem Shell-Befehl:
   `mysql --default-character-set=utf8 -u root -p < SQLUserDB.sql`

4.1.2 ingwer-Tabellen

Nun müssen noch die Tabellen für ingwer in der Datenbank angelegt werden. Verwenden Sie dazu die SQL-Datei `SQLIngwerDB.sql`:

1. Öffnen Sie die Datei `SQLIngwerDB.sql` in einem Texteditor.
2. Passen Sie in Zeile 1 den Wert `IngwerDB` an und ersetzen Sie ihn durch den Datenbanknamen, den Sie oben in Kapitel 4.1.1 festgelegt haben.
3. Das Einlesen der SQL-Datei erfolgt über folgende Wege:
   a) Wenn Sie eine Webadministratorenoberfläche wie „PHPMyAdmin“ o.ä. verwenden, können Sie die Datei `SQLIngwerDB.sql` darüber einlesen.
   b) Andernfalls verwenden Sie den Zugang zu MySQL über die Shell: Lesen Sie die Datei als MySQL-Root-Benutzer oder mit dem Benutzer, den Sie oben in Kapitel 4.1.1 eingerichtet haben, ein mit dem Befehl:

   ```
   mysql --default-character-set=utf8 -u [Benutzername] -p < SQLUserDB.sql
   ```

4.1.3 Anpassen der Metadaten-Felder

Die in Kapitel 4.1.2 beschriebene SQL-Datei definiert eine Reihe von Standardfeldern für die Erfassung der Metadaten. Diese Felder können den eigenen Bedürfnissen angepasst werden, indem die Tabellendefinition verändert wird. In der Datenbanktabelle `paper_meta_data_type` sind die Definition der Metadaten-Felder abgelegt. Die Tabelle hat folgende Spalten:

- `id`: Eindeutige Zahl. Beim Textimport werden die Felder Zeile für Zeile in aufsteigender Reihenfolge eingelesen, mit Ausnahme von word_count, lemma_count, content.
- `field_name`: Interne Feldbezeichnung.
- `name`: In Ingwer angezeigte Bezeichnung (z. B. bei der Suche)
- `meta_type`: Typ der Metadaten. Es gibt sechs unterschiedliche Typen, dazu später mehr.
- `table`: Tabellenname in der Datenbank, in welchem die Metadaten gespeichert werden.

Für die Typen der Metadaten müssen die entsprechenden Felder bzw. Tabellen angelegt werden:

1. `int` ist eine ganz Zahl z. B. für `word_anzahl`. Der SQL-Datentyp muss auch ganzzahliger sein.
2. `date` ist ein Datumswert. Der SQL-Datentyp muss vom Typ Date sein.
3. `string` ist eine Zeichenkette. Der SQL-Datentyp muss vom Typ varchar oder text sein.

4. `list` ist eine Auswahlliste. Dafür müssen zwei Tabellen folgenderweise angelegt werden:
   a) `paper2`*`Tabellenname`* mit zwei Feldern in denen die zugehörigen Schlüssel der Tabellen `paper` und *`Tabellenname`* abgespeichert werden. Die Felder sind von einem ganzzahligen SQL-Datentyp und heißen `paper_id` und *`Tabellenname`*`_id`
   b) *`Tabellenname`* mit zwei Feldern. Das erste Feld ist der eindeutige Schlüssel `id` mit ganzzahligem autoincrement Datentyp, das zweite der *`field_name`* mit varchar oder text als Datentyp.

   Beispiel:

id	field_name	name	meta_type	table
1	name	Zeitung	list	journal

Zeile aus Tabelle `paper_meta_data_type`

   a) `paper2journal` mit den zwei Feldern `paper_id` und `journal_id`
   b) `journal` mit den zwei Feldern `id` und `name`

5. `multilist` ist eine Auswahlliste mit Mehrfachauswahl. Die Tabellenstruktur entspricht der einfachen Auswahlliste.

6. `content` ist der Textkörper. Es muss genau einen content geben.

Bei den zwei Metadatenfeldern `word_count` und `lemma_count` ist der `name` und der `field_name` fest vorgegeben.

4.2 Einrichten von ingwer

Nun erfolgt noch die Installation des eigentlichen ingwer-Programms. Gehen Sie dazu wie folgt vor:

1. Passen Sie die Datei `ingwer/include/constants.php` an, indem Sie sie in einem Texteditor öffnen:
   - Zeile 6, 7 und 8: Ersetzen Sie hier `IngwerTest` durch den MySQL-Datenbanknamen, MySQL-Benutzernamen und das dazugehörige Passwort, wie Sie es in Kapitel 4.1.1 festgelegt haben.
   - In Zeilen 10 und 11 können Sie den Benutzernamen und das Passwort eintragen, falls Sie über `.htaccess` den Zugriff auf das Verzeichnis, in dem ingwer liegt, geschützt haben.
   - Zeile 37: Setzen Sie hier den absoluten Pfad zum TreeTagger, den Sie in Kapitel 2 installiert haben.
   - Zeile 38: Passen Sie hier den Pfad zum ingwer-Verzeichnis an.

- Zeilen 39 und 40: Passen Sie hier ggf. die Pfade zu `perl` und `diff` an.

2. Kopieren Sie das Verzeichnis `ingwer/` in ein Webverzeichnis Ihrer Wahl.
3. Öffnen Sie im Browser das ingwer-Verzeichnis (`http://www.ihr-server.com/ingwer/` o.ä.).
4. Viel Spaß! Eine Anleitung zur Bedienung von ingwer finden Sie in ingwer in der Rubrik „Hilfe“.

4.3 Anpassen von ingwer für die Benutzung mit der CWB

Wenn Sie ingwer in Verbindung mit der CWB und CQPWeb benutzen wollen, dann beachten Sie bitte Folgendes:

1. Installieren Sie zunächst die CWB nach folgender Anleitung: `http://cwb.sourceforge.net/download.php`
2. Installieren Sie dann CQPWeb (`http://cwb.svn.sourceforge.net/viewvc/cwb/gui/cqpweb/trunk/CQPweb-setup-manual.html`) in einem Verzeichnis `cwb` parallel zum ingwer-Verzeichnis. Die CWB-Korpora sollten in `cwbcorpora` innerhalb des ingwer-Verzeichnisses liegen.
3. Passen Sie in der ingwer-Datei `include/constants.php` folgende Zeilen an:
   - In Zeile 34: Passen Sie ggf. den Pfad zum Verzeichnis `cwbcorpora` innerhalb des ingwer-Verzeichnisses an.
   - In Zeilen 41 und 42: Passen Sie die Pfade zu `cwb-encode` und `cwb-makeall` an.

5 Lizenzbestimmungen

5.1 GNU GPL-Bestimmungen

You should have received a copy of the GNU General Public License along with this program. If not, see `http://www.gnu.org/licenses/`.

ingwer ist Freie Software: Sie können es unter den Bedingungen der GNU General Public License, wie von der Free Software Foundation, Version 3 der Lizenz oder (nach Ihrer Wahl) jeder neueren veröffentlichten Version, weiterverbreiten und/oder modifizieren.

ingwer wird in der Hoffnung, dass es nützlich sein wird, aber OHNE JEDE GEWÄHRLEISTUNG, bereitgestellt; sogar ohne die implizite Gewährleistung der MARKTFÄHIGKEIT oder EIGNUNG FÜR EINEN BESTIMMTEN ZWECK. Siehe die GNU General Public License für weitere Details.

Sie sollten eine Kopie der GNU General Public License zusammen mit diesem Programm erhalten haben. Wenn nicht, siehe `http://www.gnu.org/licenses/`.