Wikidefine

René Brückner, Christian Frommert

February 28, 2017

Wikidefine extrahiert Informationen aus Wikipedia und bildet zu jedem Artikel eine kurze Definition bestehend aus einer gegebenen Anzahl von Sätzen.

1 Einrichtung

1.1 Voraussetzungen

Wikidefine setzt eine installierte Java Runtime Environment (JRE) in der Version 8.0 voraus. Auerdem muss entweder `git` installiert sein, falls es gecloned werden sollte, oder es muss über einen Webbrowser von der Github Repository-Website heruntergeladen werden. Darüber hinaus wird maven vorausgesetzt.

Für den File Dump Extractor benötigt man einen Wikipedia Dump. Diesen erhält man von https://dumps.wikimedia.org . Von den verschiedenen Varianten wird der Dump benötigt, der alle Seiten (pages) enthält, inklusive Inhalt. Für die englische Variante ist das beispielsweise enwiki-20170220-pages-articles.xml.bz2 , für deutsch: dewiki-20170220-pages-articles.xml.bz2 .

1.2 Installation

Zuerst das Git Repository clonen:

```
git clone https://github.com/tm16wiki/wikidefine.git
```

Dann mit maven die Abhängigkeiten herunterladen und Wikidefine kompilieren:

```
cd wikidefine
mvn install
```

Anschließend Wikidefine in der Shell starten:

```
java -jar target/WikiDefine-*.jar
```

oder mit der GUI:

```
java -jar target/WikiDefine-*.jar -gui
```

2 Benutzung

Man kann Wikidefine in der Shell oder mithilfe der GUI verwenden.

2.1 Shell-Benutzung

2.1.1 Konfiguration

Wikidefine muss nach dem Start zuerst konfiguriert werden. Die erste Konfiguration muss zudem "default" genannt werden, damit diese beim nächsten Start automatisch geladen wird. Nun wird die Sprache festgelegt (aktuell de oder en). Danach muss der Pfad zum Wikipedia Dump (XML) angegeben werden. Nun werden noch die Datenbankinformationen benötigt. Unterstützt werden SQLite PostgreSQL und MySQL. Falls man einen Dateipfad zu einer nicht existierenden SQLite Datenbank angibt wird automatisch eine neue SQLite Datenbank in diesem Pfad angelegt, sofern man Schreibrechte besitzt.

2.1.2 Befehle - File Dumper

Um den File Dumper Subprozess zu starten gibt man '`fd`' ein. Nun kann man den File Dump Extractor weiter konfigurieren:

Option	Beschreibung	Befehl	Standardwert	Beispiel
Set Threads	Anzahl der zu benutzenden Threads	st	4	st 4
Maximale Definitionen	Maximale Anzahl von zu extrahierenden Defintionen	sm	Integer. MAX_VALUE	sm 5000
Dateipfad	Dateipfad zum Wikipedia Dump im XML Format	sp	in der Hauptkonfiguration festgelegter Dateipfad	sp /home/user/wikidefine/dewikidump.xml
Datenbankpfad	Pfad zur Datenbank	-	Datenbankpfad, der in der Hauptkonfiguration festgelegt wurde	Hier nicht änderbar (bitte in der Hauptkonfiguration ändern)
Datenbankexport	Spezifiziert, ob die Definitionen in der Datenbank gespeichert werden sollen	se	true	se
Debugging	Prozess zeigt akzeptierte und abgelehnte Definitionen während der Laufzeit an	sv	true	sv
Statistik	Zeigt am Ende des Prozesses die Laufzeit, die Anzahl der vorgefilterten Definitionen, die Anzahl der akzeptierten und abgelehnten Definitionen an	ss	false	ss

Nach erfolgter Konfiguration startet man den Prozess mit dem Befehl `run`.
Zum Beenden des Subprozesses gibt man `exit` ein.

2.1.3 Befehle - Web Definition

Um den Web Definition Subprozess zu starten gibt man '`wd`' ein. Nun kann man mit dem Befehl `sl` die Sprache festlegen, zum Beispiel `sl de` oder `sl en`.
Um eine Definition zu crawlen gibt man nun `d "Wikipedia Titel"` ein. Beispiel: `d "Festplattenlaufwerk"`. Die Titel sind gleichzusetzen mit der URL (hier im Beispiel: https://de.wikipedia.org/wiki/**Festplattenlaufwerk**)

Zum Beenden des Subprozesses gibt man `exit` ein.

2.2 GUI

Um Wikidefine in der GUI zu starten gibt man den Parameter `-gui` beim Start an:

```
java -jar target/WikiDefine-*.jar
```

2.2.1 GUI - Konfiguration

Im ersten Eingabefeld wird der Pfad zum Wikipedia Dump im XML Format angegeben. Durch einen Klick auf Open öffnet sich ein Dateiauswahldialog. Darunter kann man die maximale Anzahl von Definitionen angeben, die Statistik, Debug-Informationen und Datenbank-Export an- und ausschalten. Auf der rechten Seite kann man die Anzahl der zu benutzenden Threads angeben sowie die gewünschte Sprache. Nachdem man die Informationen zur Datenbank angegeben hat kann man mit dem Play-Button auf der rechten Seite den Extraktionsprozess starten. Am unteren Rand der GUI sieht man dann eine Statusbar, die Auskunft über den Fortschritt des Prozesses gibt.