Readme

`firstlink.py` enthält die Funktionalität, um auf einem Wikipedia-Artikel den ersten Link auf einen anderen Artikel zu erkennen und ihn als URL zurückzugeben (Funktion `get_first_article_link`). Dabei werden folgende Annahmen getroffen:

Der Link soll ***nicht*** auf

- eine ***Disambiguation***-Seite,
- eine Datei (Nationalhymne, Flagge, Logo) oder
- einen generischen Inhalt (Erläuterung zur internationalen Lautschrift, Hilfeseite)

führen.

Bei Artikel zu geografischen Orten erscheinen oben rechts die Koordinaten zu diesem Ort mit einem Link auf den Artikel zum Koordinatensystem. Dieser Link lässt sich derzeit noch nicht von anderen Artikellinks unterscheiden.

`firstlink_test.py` enthält zwei Testfälle:

1. `test_get_first_article_link` arbeitet das Dictionary `links_from_to` ab. Dieses beschreibt Beziehungen zwischen einem Quellartikel (key) und einem Zielartikel (value), d.h. der erste im Quellartikel verlinkte Artikel. Diese Beziehungen wurden durch manuelle Tests ermittelt. Der Test ist erfolgreich, wenn der erste im Quellartikel gefundene Link auf den Zielartikel verweist.
2. `test_find_random_n_away_from_target` arbeitet das Dictionary `distances`ab. Dieses beschreibt, mit wie vielen Sprüngen man von einem Ausgangsartikel auf den Artikel "Philosophy" kommt. Es gibt eine Obergrenze an auszuführenden Sprüngen (`stop_after_hops`). Dadurch kann derzeit nicht unterschieden werden, ob die Suche nach dieser Anzahl von Sprüngen erfolgreich war oder in einer Endlosschleife endete. Der Testfall prüft, ob man vom jeweiligen Quellartikel tatsächlich in der angegebenen Anzahl Sprüngen auf den Artikel "Philosophy" gelangt.

Für die Tests wird die englischsrachige Wikipedia verwendet.

WikiHopper

Dieses Programm nimmt eine Liste von Begriffen entgegen und prüft in wievielen Sprüngen man von der (deutschsprachigen) Wikipedia-Seite des Begriffs auf die auf den Artikel "Philosophie" kommt.

Dazu wird die Funktion `get_first_article_link` von `firstlink.py` aufgerufen. Anschliessend schreibt das Programm die Anzahl Sprünge in eine CSV-Datei. Bei einer ***Exeption*** wird der Buchstabe `E`, bei Überschreitung der maximalen Anzahl Sprünge der Buchstabe `X` eingetragen.

Ziel-Seite, maximale Anzahl Sprünge, Input und Output File können in der Datei `config.xml` angepasst werden.