Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Pfaffenwaldring 5B
D–70569 Stuttgart

Bachelorarbeit

Soziale Netzwerke in Wikipedia

Kim Trong Truong

Studiengang:	Softwaretechnik
Prüfer/in:	Prof. Dr. Grzegorz Dogil
Betreuer/in:	Dr. Daniel Duran
Beginn am:	17.07.2017
Beendet am:	17.01.2018
CR-Nummer:	H.3.3 Information Search and Retrieval

Kurzfassung

Das Ziel dieser Arbeit ist es, zu prüfen, ob Informationen über Sprachnutzung und Sprachkenntnis aus Wikipedia gesammelt werden können. Dazu wird geklärt, welche Informationen verfügbar sind und mit welchen Methoden diese Informationen extrahiert und verarbeitet werden können. Zudem wird die Struktur und Qualität eines extrahierten Datensatzes mithilfe einer sozialen Netzwerkanalyse untersucht. Diese Arbeit veranschaulicht, auf welche Kriterien bei der Datenextraktion aus Wikipedia geachtet werden muss und welche Struktur die erhaltenen Daten haben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Listings

1. Einleitung

Aus Informationen über die Sprachkenntnisse und Sprachnutzung von einzelnen Personen oder Personengruppen lassen sich viele Rückschlüsse ableiten. So kann die Sprachverbreitung einer Sprache untersucht werden oder ihr Aussterben untersucht werden. Ebenso sind Information über beherrschte Sprachkombination interessant, weil damit Annäherung zwischen Ländern belegt werden können oder Migrationsmuster beschrieben werden können. Mit Informationen über die Sprachverbreitung können eventuell auch Rückschlüsse auf die Verbreitung von Kultur oder anderen sozialen Merkmalen abgeleitet werden. Um diese Informationen über Sprachnutzung und Sprachfähigkeiten zu erhalten, müssen Befragungen durchgeführt werden. Eine Befragung ist aufwendig und für manche Sprachen schwer zu erhalten, daher sollen in dieser Arbeit stattdessen bereits vorhandene Daten aus Wikipedia extrahiert werden. Wikipedia ist eine Online-Enzyklopädie, in der mehrere Personen an Artikeln zusammen schreiben können. Aus dieser Tätigkeit lassen sich gleich mehrere Informationen entnehmen, nämlich mit wem der Autor zusammen geschrieben hat und welche Sprache der Autor beherrscht.

In dieser Arbeit sollen Daten von Wikipedia über Sprachnutzung und Sprachfähigkeiten aus den 11 Amtssprachen Südafrikas extrahiert werden. Die 11 Amtssprachen Südafrikas sind gut geeignet, weil viele Nutzer, die in diesen Sprachen schreiben, auch eine der anderen Sprachen beherrschen könnten. Dadurch wird ein kompakter Datensatz von Nutzern und ihren beherrschten Sprachen generiert. Zudem wird überprüft, welche Arten von Informationen über Sprachnutzung und Sprachfähigkeiten aus Wikipedia entnommen werden können.

Anschließend wird die Struktur der extrahierten Daten aus Wikipedia auf Auffälligkeiten und Qualität überprüft. Dazu wird ein soziales Netzwerk erstellt, das die Struktur der Zusammenarbeit zwischen den Autoren darstellt. Anhand dieses Netzwerkes wird eine soziale Netzwerkanalyse durchgeführt, um Metriken über das soziale Netzwerk zu erhalten. Diese Metriken werden dann mit Vergleichsmetriken für andere soziale Netzwerke aus der Literatur verglichen.

2. Grundlagen

Grundlegend für das Verständnis dieser Arbeit sind Netzwerke und der Aufbau von Wikipedia. In diesem Kapitel werden Netzwerke beschrieben, die die Grundlage einer sozialen Netzwerkanalyse sind. Außerdem werden einige Konzepte von Wikipedia erklärt, deren Verständnis wichtig für die Beschaffung der Daten ist.

2.1. Netzwerke

Ein Netzwerk wird im mathematischen Kontext auch als Graph bezeichnet. Die Begriffe Netzwerk und Graph werden oft als Synonyme verwendet. In dieser Arbeit werden diese Begriffe ebenfalls als Synonyme verwendet. Ein Netzwerk besteht aus einer Menge von Knoten, die mit Kanten zueinander verbunden werden können. Die semantische Bedeutung eines Netzwerkes wird durch die Belegung der Knoten und Kanten festgelegt. So beschreibt ein Netzwerk mit Städtenamen als Knoten und Entfernungen als Kanten ein Straßennetzwerk, wohingegen ein Netzwerk mit Namen als Knoten und Freundschaftsbeziehungen als Kanten ein Freundschaftsnetzwerk darstellt. Aufgrund der unterschiedlichen Interpretationsmöglichkeiten, bei der Belegung der Knoten und Kanten, existieren verschiedene Netzwerktypen. Netzwerke besitzen Metriken, die sich mit einer Netzwerkanalyse messen lassen. Zudem wurde in den Arbeiten [New02],[New03] und [New10] beobachtet, dass sich die verschiedenen Netzwerktypen in einigen Metriken voneinander unterscheiden.

In den folgenden Abschnitten werden einige Eigenschaften von Netzwerken, grundlegende Metriken sowie verschiedene Netzwerktypen vorgestellt.

2.1.1. Eigenschaften von Netzwerken

Gerichtete und ungerichtete Netzwerke Bei gerichtete Netzwerken besitzen die Kanten zwischen den Knoten eine Richtung, diese werden grafisch mit einem Pfeil dargestellt. Für diese Kanten werden dann ein Ursprungsknoten und ein Zielknoten festgelegt. Ungerichtete Netzwerke haben dagegen Kanten, die keine Richtung besitzen.

Gewichtete und ungewichtete Netzwerke Gewichtete Netzwerke besitzen Gewichte an den Kanten. Diese Gewichte können durch verschiedene Maßeinheiten festgelegt werden. Ungewichtete Netzwerke besitzen dagegen keine Gewichte an den Kanten.

2.1.2. Grundlegende Metriken von Netzwerken

Grundlegende Metriken sind :

Knotenanzahl beschreibt, wie viele Knoten existieren.

Kantenanzahl beschreibt, wie viele Kanten existieren.

Knotengrad beschreibt, wie viele Kanten mit einem Knoten verbunden sind.

Eingangsgrad beschreibt, wie viele Kanten mit einem Knoten als Zielknoten verbunden sind.

Ausgangsgrad beschreibt, wie viele Kanten mit einem Knoten als Ursprungsknoten verbunden sind.

Pfadlänge beschreibt, wie groß die Folge von Kanten von einem Knoten zu einem anderen Knoten ist.

In Kapitel 4 werden noch einige spezifischere Metriken vorgestellt.

2.1.3. Netzwerktypen

In [New10] hat Newman vier Arten von Netzwerktypen vorgestellt, weil diese sich durch ihre Eigenarten häufig mit ähnlichen Methoden analysieren lassen. Im Folgenden werden die unterschiedlichen Netzwerktypen nach Newman vorgestellt:

Technische Netzwerke Nach Newman sind Beispiele für ein technisches Netzwerk das Internetnetzwerk, das Telefonnetz, das Transportnetzwerk, das Stromnetz oder das Vertriebsnetzwerk. Bestimmend für technische Netzwerke sind, dass meistens etwas transportiert wird. So werden zum Beispiel beim Internetnetzwerk Datenpakete transportiert. Da in einem technischen Netzwerk Informationen oder Waren transportiert werden, sollten die Pfadlängen möglichst gering sein. Ebenfalls wichtig ist es, dass jeder Knoten mit den anderen Knoten im Netzwerk verbunden ist. Außerdem sollte bei einem Ausfall eines Knotens nicht das ganze Netzwerk zusammenbrechen, hierbei spielt die Knotengradmetrik eine wichtige Rolle. So sollen Knoten die einen hohen Knotengrad besitzen entlastet werden. Ein häufiges Ziel bei der Untersuchung dieser Netzwerke ist das Optimieren der Pfadlängen zwischen den Knoten. Ein weiteres Ziel ist es ausfallresistente Strukturen zwischen den Knoten zu finden.

Soziale Netzwerke Soziale Netzwerke beschreiben soziale Beziehungen zwischen Individuen oder Gruppen. Newman zählt zu diesen Netzwerken Freundschaftsnetzwerke und Zugehörigkeits Netzwerke. Bei Zugehörigkeits Netzwerken werden für einzelne Individuen die Zugehörigkeit zu Gruppen, Klassen oder Parteien beschrieben. Bei sozialen Netzwerken gibt es eine Unterscheidung zwischen soziometrischen Netzwerken, bei der alle Individuen im Netzwerk gleich behandelt werden und den egozentrierten Netzwerken, bei denen ein Individuum im Mittelpunkt steht. Unter dem Kleine-Welt-Experiment [TM69] wurde entdeckt, dass in einem realen sozialen Netzwerk Personen kurze Bekanntschaftsketten zu anderen Personen haben. Eine soziale Netzwerkanalyse beschreibt verschieden Verfahren, mit denen ein Netzwerk untersucht wird, dazu zählen spezifische Metriken, die in Kapitel 4 beschrieben werden.

Informationsnetzwerke Zu den Informationsnetzwerken zählt Newman das World Wide Web und Zitationsnetzwerke. In Informationsnetzwerken haben Informationen, zum Beispiel in Form von Dokumenten oder Webseiten, eine Verbindung zueinander. Im World Wide Web werden Webseiten durch Hyperlinks miteinander verbunden. Die Relevanz von Seiten kann durch die PageRank-Metrik[Pag+99], die die Wichtigkeit einer Webseite anhand der Häufigkeit der Verlinkungen innerhalb anderer Seiten bewertet, festgelegt werden. Zitationsnetzwerke beschreiben die Zitationen von akademischen Publikationen untereinander. In Zitationsnetzwerken können die am meisten zitierten Arbeiten untersucht werden, um ihre Wirkung und Einfluss auf andere Arbeiten zu beurteilen.

Biologische Netzwerke Biologische Netzwerke stellen biologische oder chemische Strukturen aus der realen Welt da. Newmans Beispiele dafür sind das biochemische Netzwerk, das neuronale Netzwerk und das ökologische Netzwerk. Bei einem biochemischen Netzwerk werden chemische Prozesse beschrieben. In einem neuronalen Netzwerk werden Neuronen aus dem Nervensystem beschrieben. Bei der Analyse von neuronalen Netzwerken wird versucht Muster zu entdecken, um das Netzwerk besser zu verstehen. Zu den ökologischen Netzwerken zählt zum Beispiel die Darstellung einer Nahrungskette in einem Ökosystem. In biologischen Netzwerken wird versucht, die Struktur und Beschaffenheit des Netzwerkes zu verstehen.

2.2. Wiki

Wikis sind Webseiten auf denen Nutzer Artikel über verschiedene Themen erstellen, bearbeiten oder lesen können. Jedes Wiki dient als Nachschlagewerk für ein bestimmtes Themengebiet. Das bekannteste Wiki ist Wikipedia, das als Online-Enzyklopädie zur Verfügung steht. Weitere Wikis sind Wiktionary, das als Wörterbuch dient, und Wikimedia Commons, das eine Sammlung aus Bildern, Videos und Audiodateien ist. Derzeit hat Wikimedia Commons die größte Anzahl[1] an gesammelten Artikeln gefolgt von Wikipedia und Wiktionary [Wik].

Die Wikimedia Foundation ist für die Betreuung von mehreren Wikis zuständig. Dazu zählen unter anderem Wikipedia, Wiktionary und Wikimedia Commons. Für diese Wikis von Wikimedia existieren einige zusätzliche Dienste, wie zum Beispiel ein globaler Benutzeraccount oder APIs mit der man automatisiert auf Daten zuzugreifen kann. Viele Wikis basieren auf der Verwaltungssoftware MediaWiki, darunter auch Wikipedia, Wiktionary und Wikimedia Commons. MediaWiki ist dafür zuständig, dass Texte online bearbeitet werden können. Außerdem sorgt die MediaWiki Software dafür, dass Änderungen an den Artikeln auf den Servern von Wikimedia geloggt und gespeichert werden. Wikis, die nicht die MediaWiki Software verwenden, können sich in der Art der Speicherung der Inhalte unterscheiden.

[1] Die größte Anzahl aus den Wikis von Wikimedia

2.3. Wikipedia

Wikipedia gestattet es Nutzern gemeinsam an Artikeln zu schreiben. Die wichtigsten Konzepte werden in den nächsten Abschnitten vorgestellt. Es werden nicht alle Einzelheiten von Wikipedia erläutert, sondern nur die für diese Arbeit relevanten Konzepte.

2.3.1. Artikel Eigenschaften

Artikel Artikel können Texte oder andere Medien wie Grafiken oder Audio enthalten. Artikel werden durch ihren Titel festgelegt. Diese Titel sind einzigartig, daher können nicht zwei Artikel den gleichen Titel besitzen. Jeder Nutzer kann einen Artikel neu anlegen oder bestehende bearbeiten. Bei der Anlegung oder Bearbeitung eines Artikels muss der Nutzer angemeldet sein, ansonsten wird er als anonymer Bearbeiter geloggt. Als anonymer Bearbeiter wird die IP-Adresse des Bearbeiters gespeichert.

Artikelversion Jeder Artikel auf Wikipedia besitzt eine Versionsnummer, das die Version des Artikels festlegt. Neu erstellte Artikel haben die Versionsnummer 0. Durch jede Bearbeitung steigt die Versionsnummer. Die höchste Versionsnummer legt die aktuellste Version des Artikels fest, und damit auch die Version die auf Wikipedia angezeigt werden soll. Über jeden Artikel gibt es eine Versionsgeschichte, in der alle existierenden Versionen mit dem Namen ihres Bearbeiters aufgelistet werden. Falls eine Version von einem anonymen Nutzer bearbeitet wurde, wird stattdessen die IP-Adresse des Nutzers angezeigt.

Artikel Weiterleitung Artikel können auch als Weiterleitung dienen. Artikel mit Weiterleitungen sind Artikel, die meist keinen Inhalt haben, und auf einen anderen Artikel weiterleiten. Zum Beispiel leiten derzeit, in der deutschen Wikipedia Sprachversion, die Artikel mit dem Titel “A*” oder “A-Stern-Algorithmus” zu dem Artikel mit dem Titel A*-Algorithmus weiter.

2.3.2. Sprachversion

Wikipedia besitzt unterschiedliche Sprachversionen. Diese legen fest für welchen Sprachraum die Artikel geschrieben sind. Die Sprachversion wird durch die Subdomain festgelegt. Beispielsweise legt die Subdomain “de” in der URL https://de.wikipedia.org fest, dass die deutsche Sprachversion aufgerufen werden soll. Alle Sprachversionen sind eigenständig. Das bedeutet, dass Artikel nur in der jeweiligen Sprachversion existieren, in der sie erstellt wurden. Eine Liste mit den derzeit existierenden Sprachversionen und ihrer zugehörigen Subdomains befindet sich auf der Seite https://meta.wikimedia.org/wiki/List_of_Wikipedias.

2.3.3. Namensraum

Artikel besitzen unterschiedlich Namensräume. Diese werden durch einen Zusatz in ihrem Titel festgelegt. Zum Beispiel gibt die Seite https://de.wikipedia.org/wiki/Hilfe:Versionen einen Artikel mit dem Titel “Hilfe:Versionen” zurück, welcher sich im Hilfe-Namensraum befindet. Namensräume dienen dazu Artikel in verschiedene Artikel-Typen einzuteilen. So sind alle Artikel auf Wikipedia, die in ihrem Titel mit “Hilfe:” beginnen Hilfeseiten, welche die MediaWiki Software erklären.

Es existieren mehrere Namensräume, die auf https://en.wikipedia.org/wiki/Wikipedia:Namespace aufgelistet sind. Im nächsten Abschnitt werden einige für diese Arbeit relevanten Namensräume beschrieben:

Artikel-Namensraum In diesem Namensraum befinden sich die enzyklopädischen Artikel. Es existiert kein Zusatz vor dem Titel des Artikels.

Artikel Diskussionsseiten-Namensraum In diesem Namensraum befinden sich die Diskussionsseiten zu den enzyklopädischen Artikel. Hier kann über den Inhalt oder die Bearbeitung des zugehörigen Artikels, dessen Titel sich hinter dem Titelzusatz befinden, diskutiert werden. Der Titelzusatz für diesen Namensraum ist **Talk:**

Benutzer-Namensraum In diesem Namensraum befinden sich Artikel über die Nutzer von Wikipedia. In diesen Artikeln kann man persönliche Informationen über den Nutzer schreiben. Diese Artikeln werden meistens von dem jeweiligen Nutzer selber bearbeitet. Es ist aber auch möglich, dass andere Nutzer diese Artikel bearbeiten. Für den Besitzer ist es möglich Bearbeitungen von anderen Nutzern zu verbieten. Der Titelzusatz für diesen Namensraum ist **User:**.

Benutzer Diskussionsseiten-Namensraum In diesem Namensraum befinden sich die Diskussionsseiten zu den Artikeln über die Nutzer. Hier kann über den Artikel des zugehörigen Nutzers, dessen Name sich hinter dem Titelzusatz befindet, diskutiert werden. Dieser Namensraum dient meistens zur Kommunikation zwischen den Nutzern. Andere Nutzer können Nachrichten hinterlassen, indem sie die Benutzer Diskussionsseite editieren. Der Titelzusatz für diesen Namensraum ist **User_talk:**

Von den oben vorgestellten Titelzusätzen existieren auch Sprachversion abhängige Titelzusätze. In der deutschen Sprachversion existieren zum Beispiel stattdessen die Titelzusätze **Diskussion:**, **Benutzer:** oder **Diskussion_Nutzer:**. Die oben beschriebenen englischen Titelzusätze gelten jedoch global für alle Sprachversionen und können die deutschen Titelzusätze ersetzen. Umgekehrt gilt es nicht, das heißt, dass deutsche Titelzusätze oder die Titelzusätze andere Sprachen nur für die jeweilige Sprachversion gilt.

2.3.4. Benutzeraccount

Benutzeraccounts werden durch den Benutzernamen eindeutig unterschieden. Die Benutzernamen sind case-sensitive und deren Anfangsbuchstabe wird immer großgeschrieben. Ab dem 22. August 2008 wurde der Single-User-Login eingeführt [Sin]. Dadurch musste man sich als Nutzer nur einmal registrieren, um sich an allen Wikis von Wikimedia anzumelden. Davor musste in jedem Wiki ein neuer Benutzeraccount angelegt werden. Dies betraf ebenfalls die verschiedenen Sprachversionen von Wikipedia, da diese als eigenständige Wikis betrachtet werden. Ab Juli 2015 wurden nachträglich alle Benutzernamen, die zwischen den verschiedenen Wikis von Wikimedia uneindeutig sind, umbenannt[Sin]. Seitdem kann jeder Nutzername auf den verschiedenen Wikis eindeutig einer Person zugeordnet werden. Durch diese Änderung wurde es möglich Beiträge zwischen den Sprachversionen eindeutig einzelnen Personen zuzuordnen. Derzeit wird bei der Registrierung der Benutzeraccount nur auf der jeweiligen Sprachversion registriert. Der Nutzer wird noch nicht als Nutzer in den anderen Sprachversionen aufgelistet, jedoch ist bereits der Name des Nutzers für alle anderen Sprachversion reserviert. Erst nachdem sich der Nutzer mit

seinen bestehenden Anmeldedaten in einer anderen Sprachversion erstmalig anmeldet, wird der Nutzer auch dort aufgelistet.

3. Datenbeschaffung und Verarbeitung

Einige wissenswerten Fragen über das Sprachkenntnis und Sprachnutzung von Personen sind im Folgenden beschrieben:

- Welche Sprachen werden beherrscht?
- Wie gut wird eine Sprache beherrscht?
- Was ist die erst gelernte Sprache?
- Wie oft und wie lange wird die jeweilige Sprache verwendet?
- An welchen Orten wird die Sprache genutzt (zum Beispiel bei der Arbeit, zu Hause, auf Reisen etc.) ?

Einige dieser Fragen können mithilfe der vorliegenden Daten aus Wikipedia beantwortet werden. Ein Ziel dieser Arbeit ist es Antworten auf die oben beschriebenen Fragen für die 11 Amtssprachen Südafrikas zu erhalten. In diesem Kapitel werden Methoden vorgestellt, um die gesuchten Informationen aus Wikipedia zu erhalten. Im Abschluss dieses Kapitel werden die Ergebnisse aus der Datenbeschaffung vorgestellt und es wird kurz darauf eingegangen, welche Informationen tatsächlich aus Wikipedia generiert wurden.

3.1. Datenbeschaffung

In dieser Arbeit wurden zwei Methoden zur Datenbeschaffung betrachtet. Die erste Methode besteht darin die benötigten Informationen aus den, von der Wikimedia Foundation, bereitgestellten Speicherauszügen zu beschaffen. Die zweite Methode nutzt die MediaWiki-API, die es erlaubt spezifische Abfragen zu generieren und zu nutzen. Außerdem wird noch das Babel-System vorgestellt. Mithilfe des Babel-Systems können weitere Informationen über die Sprachfähigkeit der Nutzer aus Wikipedia gesammelt werden.

3.1.1. Wikimedia Speicherauszüge

Auf der Seite https://dumps.wikimedia.org/ findet man Speicherauszüge zu allen Artikeln aus Wikipedia und anderen Wikis von Wikimedia. In diesen Speicherauszügen sind zu jedem angelegten Artikel auch alle älteren Versionen des Artikels gespeichert. Zu jeder Version des Artikels wurden Titel,Bearbeitername, Bearbeitungszeitpunkt und der Text des Artikels gespeichert. In diesen Speicherauszügen befinden sich Artikel aus allen existierenden Namensräumen. Ein Speicherauszug wird für jedes Wiki der Wikimedia Foundation, und damit auch für jede Sprachversion

von Wikipedia, einzeln bereitgestellt. Die Speicherauszüge werden in XML-Dateien bereitgestellt. Vereinzelt gibt es diese Auszüge auch als Datenbanken im SQL-Format. Die aktuellsten Speicherauszüge werden nach angaben von Wikimedia mindestens einmal monatlich, manchmal auch zweimal im Monat, bereitgestellt. Im Folgenden wird die Struktur eines Speicherauszugs im XML-Format beschrieben:

Listing 3.1 Exemplarischer Aufbau einer pages-meta-history.xml Datei aus dem Wikimedia Speicherauszug.

```
<siteinfo>
      ... <!--siteinfo beinhaltet einige Informationen ueber die gesamte XML-->
</siteninfo>
<page>
      <title> ARTIKELNAME </title>
      <ns>0</ns>
      <id>1</id>
      <revision>
            <id>BEARBEITUNGSNUMMER</id>
            <timestamp> DATUM </timestamp>
            <contributor>
                  <username> NUTZERNAME_DES_BEARBEITERS </username>
                  <id>ID_DES_BEARBEITERS</id>
            </contributor>
            ...
            <text xml:space="preserve"> KOMPLETTER_ARTIKEL_TEXT </text>
            ...
      <revision>
            <id>BEARBEITUNGSNUMMER</id>
            <parentid>1</parentid>
            <timestamp> DATUM </timestamp>
            <contributor>
                  <username> NUTZERNAME_DES_BEARBEITERS </username>
                  <id>ID_DES_BEARBEITERS</id>
            </contributor>
            ...
            <text xml:space="preserve"> KOMPLETTER_ARTIKEL_TEXT </text>
            ...
      <revision>
      <!--Falls der Artikel weitere Bearbeitung hat, werden hier weitere
          revision-Elemente angehaengt. -->
</page>
<!--Falls weitere Artikel existieren, werden hier weitere page-Elemente angehaengt. -->
```

Mit diesen XML-Dateien lassen sich sehr viele Informationen sammeln. Ein Nachteil dieser Speicherauszüge ist die Größe, da in jeder Artikelversion zusätzlich der komplette Text gespeichert wird. Die XML-Dateien der englischen Sprachversion sind mehrere Terabyte groß.

3.1.2. Wiki-API

Die MediaWiki action API[1] ist ein Webservice, der das Abgreifen von Daten aus Wiki Projekten, die die MediaWiki Software nutzen, ermöglicht. Um Informationen zu erhalten muss eine HTTP-Get-

[1] https://www.mediawiki.org/wiki/API:Main_page

Anfrage gestellt werden. Mit der MediaWiki action Api ist es zudem möglich andere Funktionen durchzuführen, wie zum Beispiel Nutzer zu erstellen oder Artikel zu bearbeiten. Diese Funktionen werden in dieser Arbeit nicht beschrieben. Es werden nur die relevanten Abfragefunktionen für die Datenbeschaffung beschrieben. Eine HTTP-Get-Anfrage an Wikipedia ist aus mehreren Teilen aufgebaut und wird im Folgenden beschrieben.

$$https:// \underbrace{de}_{\substack{Subdomain \\ (Sprachversion)}} .\underbrace{wikipedia.org/}_{Wiki\ Projekt} \text{ w/ } \underbrace{api.php?}_{MediaWiki\ API\ Aufruf} QUERYSTRING \tag{3.1}$$

Der Query String besteht aus mehreren Parameter-Werte-Paaren, die mit dem Zeichen "&" verbunden sind. Im folgenden Abschnitt werden die für diese Arbeit wichtigen Parameter-Werte-Paare vorgestellt.

format-Paramter Dieser Parameter legt fest, welches Datenformat eine mögliche Ausgabe der Ergebnisse haben soll. Als Parameter-Werte-Paare stehen **format=xml**, **format=json**, **format=php** und **format=rawfm** zur Verfügung.

action-Paramter Dieser Parameter legt fest, welche Aktion mit der MediaWiki action API durchgeführt werden soll. Zum Beispiel kann mit **action=delete** ein Artikel gelöscht werden oder mit **action=createaccount** ein neuer Benutzeraccount erstellt werden. Mit dem Wert Parameter-Werte-Paar **action=query** werden Abfragen durchgeführt. Falls das Parameter-Werte-Paar **action=query** benutzt wird, kann die Abfrage mit den Parameter-Werte-Paaren **list-Parameter** oder **meta-Parameter** spezifiziert werden.

list-Paramter Dieser Parameter legt fest, dass eine Liste zurückgegeben werden soll. Die zurückgegebenen Informationen aus der Liste können mit den Parameter-Werte-Paare **list=allrevision** , **list=usercontrib** und **list=allredirect** weiter spezifiziert werden. Die Parameter-Werte-Paare sind wie folgt beschrieben:

list=allrevision liefert in der Liste alle durchgeführten Bearbeitungen zurück und dadurch erhält man alle Artikel mit all ihren Versionen (siehe Listing 3.2). Das Parameter-Werte-Paar **list-allrevision** wird mit folgenden Parameter-Werte-Paaren weiter spezifiziert:

arvprop=ids|timestamp|user|userid|content legt fest, welche Informationen zurückgegeben werden sollen

ids gibt die Identifikationsnummer des Artikel zurück.

timestamp gibt den Zeitpunkt der Bearbeitung zurück.

user gibt den Name des Bearbeiters zurück.

userid gibt die Identifikationsnummer des Bearbeiters zurück.

content gibt den kompletten Inhalt des bearbeiteten Artikels zurück.

Durch die Trennung mit "|" können auch mehrere, der zuvor beschriebenen Werte kombiniert werden.

arvuser="USERNAME" legt fest, dass nur Bearbeitungen der Nutzers mit dem Namen "USERNAME" zurückgegeben werden. Es kann jedoch nur eine Nutzername eingetragen werden.

arvnamespace= "WERT" kann verschiedene Werte annehmen und zwar:

***** legt fest, dass Bearbeitungen aus allen Namensräumen angezeigt werden

0 legt fest, dass nur Bearbeitungen aus dem Artikel-Namensraum angezeigt werden.

1 legt fest, dass nur Bearbeitungen aus dem Artikel Diskussions-Namensraum angezeigt werden.

2 legt fest, dass nur Bearbeitungen aus dem Benutzer-Namensraum angezeigt werden.

3 legt fest, dass nur Bearbeitungen aus dem Benutzer Diskussions-Namensraum angezeigt werden

Eine Abfrage URL kann wie folgt aussehen:

https://de.wikipedia.org/w/api.php?action=query&format=xml&list=allrevisions&arvprop=ids|timestamp|user|userid&arvlimit=max&arvnamespace=0 [2]

Listing 3.2 Exemplarischer Aufbau einer list=allrevision Rückgabe

```
<allrevisions>
      <page pageid="ARTIKELID" ns="0" title="ARTILTITEL">
            <revisions>
                  <rev revid="BEARBEITUNGSID"
                       parentid="BEARBEITUNGSID_DER_VORHERIGEN_ARTIKLEVERSION"
                       user="BEARBEITERNAME"  userid="BEARBEITERID"
                       timestamp="BEARBEITUNGSZEITPUNKT" />
                  <!--Falls der Artikel weitere Bearbeitung hat, werden hier
                       weitere rev-Elemente angehaengt. -->
            </revisions>
      </page>
      <!--Falls weitere Artikel existieren, werden hier weitere
          page-Elemente angehaengt. -->
<\allrevision>
```

[2] Das Parameter-Werte-Paar arvlimit=max gibt an wie viele Ergebnisse zurückgegeben werden sollen. Es können Zahlen zwischen 10 und 500 eingeben werden. Falls max eingeben wird, werden 500 Ergebnisse zurückgegeben. Am Ende dieses Abschnitts wird noch beschrieben, dass Botrechte genutzt werden können, um mehr als 500 Ergebnisse zu erhalten

list=usercontrib liefert in der Liste alle durchgeführten Bearbeitungen mehrerer festgelegter Nutzer zurück und dadurch erhält man alle Artikel mit all ihren Versionen (siehe Listing 3.3). Das Parameter-Werte-Paar **list=usercontrib** wird mit folgenden Parameter-Werte-Paaren weiter spezifiziert:

ucprop=ids|timestamp|user|userid|content legt fest, welche Informationen zurückgegeben werden sollen

ids gibt die Identifikationsnummer des Artikel zurück.

title gibt den Titel des Artikels zurück.

timestamp gibt den Zeitpunkt der Bearbeitung zurück.

Durch die Trennung mit "|" können auch mehrere, der zuvor beschriebenen Werte kombiniert werden. Im Gegensatz zur list=allrevision können bei **list=usercontrib** keine Artikeltexte ausgegeben werden.

ucuser="USERNAME" legt fest, dass nur Bearbeitungen der Nutzers mit dem Namen "USERNAME" zurückgegeben werden. Es können im unterschied zur **list=allrevision&arvuser="USERNAME"** bis zu 50 Nutzer gleichzeitig angefragt werden. Die Nutzernamen müssen dazu mit "|" getrennt werden.

ucnamespace= "WERT" kann verschiedene Werte annehmen und zwar:

***** legt fest, dass Bearbeitungen aus allen Namensräumen angezeigt werden

0 legt fest, dass nur Bearbeitungen aus dem Artikel-Namensraum angezeigt werden.

1 legt fest, dass nur Bearbeitungen aus dem Artikel Diskussions-Namensraum angezeigt werden.

2 legt fest, dass nur Bearbeitungen aus dem Benutzer-Namensraum angezeigt werden.

3 legt fest, dass nur Bearbeitungen aus dem Benutzer Diskussions-Namensraum angezeigt werden

Eine Abfrage URL kann wie folgt aussehen:

https://de.wikipedia.org/w/api.php?action=query&format=xml&list=usercontribs&ucuser="USERNAME"&ucnamespace=0&ucprop=ids|title|timestamp&uclimit=max

Listing 3.3 Exemplarischer Aufbau einer list=usercontribs Rückgabe

```
<usercontribs>
      <item userid="BEARBEITERID" user="BEARBEITERNAME"
          pageid="ARTIKELID" revid="BEARBEITUNGSID"
          parentid="BEARBEITUNGSID_DER_VORHERIGEN_ARTIKLEVERSION"
          ns="0" title="ARTIKELTITEL"
          timestamp="BEARBEITUNGSZEITPUNKT" />
      <!--Falls der Nutzer weitere Bearbeitungen hat oder weitere
          Nutzer in ucuser="..." angegeben wurden, werden hier
          weitere item-Elemente angehaengt. -->
<\usercontribs>
```

list=allredirect liefert in der Liste alle Artikel mit Weiterleitung zurück (siehe Listing 3.4). Das Parameter-Werte-Paar **list-allredirects** wird mit folgenden Parameter/Werte-Paaren weiter spezifiziert:

arprops=ids|title legt fest, welche Informationen zurückgegeben werden sollen.

ids gibt die id der Seite zurück, von welcher weitergeleitet wurde.

title gibt den Titel der Seite zurück, zu welcher hingeleitet wird.

Durch die Trennung mit "|" können auch mehrere, der zuvor beschriebenen Werte kombiniert werden.

arnamespace= "WERT" kann verschiedene Werte annehmen und zwar:

***** legt fest, dass Bearbeitung aus allen Namensräumen angezeigt werden

0 legt fest, dass nur Bearbeitungen aus dem Artikel-Namensraum angezeigt werden.

1 legt fest, dass nur Bearbeitungen aus dem Artikel Diskussions-Namensraum angezeigt werden.

2 legt fest, dass nur Bearbeitungen aus dem Benutzer-Namensraum angezeigt werden.

3 legt fest, dass nur Bearbeitungen aus dem Benutzer Diskussions-Namensraum angezeigt werden

Eine Abfrage URL kann wie folgt aussehen:

https://de.wikipedia.org/w/api.php?action=query&format=xml&list=allredirects&arprop=title|ids&arnamespace=0&arlimit=max

Listing 3.4 Exemplarischer Aufbau einer list=allredirects Rückgabe

```
<allredirects>
      <r fromid="PAGEID_HERGELEITET" ns="0"
          title="PAGETITLE_HINGELEITET" />
      <r fromid="PAGEID_HERGELEITET" ns="0"
          title="PAGETITLE_HINGELEITET" />
      <!--Falls weitere Weiterleitungen existieren, werden hier
          weitere r-Elemente angehaengt. -->
<\allredirects>
```

meta-Paramter liefert allgemeine Informationen über Wikipedia zurück. Durch die Spezifizierung mit den Parameter-Werte-Paaren **meta=siteinfo** und **sisprops=statistics** werden unter anderem Informationen über die Anzahl aller Seiten[3], Artikel[4], Bearbeitungen und Nutzer gegeben.

Eine Abfrage URL kann wie folgt aussehen:

https://de.wikipedia.org/w/api.php?action=query&format=xml&meta=siteinfo&siprop=statistics

Die Anzahl der zurückgegebenen Ergebnisse ist auf 500 beschränkt. Man erhält also pro HTTP-Get-Anfrage nur 500 Ergebnisse zurück. Falls mehr als 500 Ergebnisse existieren, wird eine continueId im Ergebnis zurückgegeben. Mit dem Parameter-Werte-Paar **arvcontinue="ZURÜCKGEGEBENE CONTINUEID"**, unter dem Wert **list=allrevision**, oder **arcontinue="ZURÜCKGEGEBENE CONTINUEID"**, unter dem Wert **list=allredirect**, lassen sich die nächsten 500 Ergebnisse anzeigen. Eine Anmeldung mit einem Benutzeraccount um die MediaWiki action API zu nutzen ist nicht nötig. Es besteht jedoch die Möglichkeit sich vor der HTTP-Get-Anfrage mit Botrechten anzumelden, dadurch sind 5000 Ergebnisse satt 500 erlaubt. Um Botrechte zu erhalten muss ein Antrag eingereicht werden, daher wurde in dieser Arbeit kein Account mit Botrechten erstellt.

3.1.3. Babel

Jeder registrierter Nutzer auf Wikipedia hat eine eigene Benutzerseite. Ein Nutzer erhält in jeder Sprachversion, in der er sich bereits angemeldet hat, eine eigene Benutzerseite. Benutzerseiten dienen dazu persönliche Informationen zu veröffentlichen. Auf den Benutzerseiten können Informationen über die Sprachfähigkeiten angegeben werden, dazu wird das Babel-System[5] benutzt. Mit dem Babel-System werden die Sprachfähigkeiten mit Babel-Bausteinen angegeben. Diese können unterschiedlich aussehen und einige Beispiele sind in der Abbildung 3.1 zu finden.

[3] Seiten bezieht sich auf den von meta zurückerhaltenen Wert pages und kennzeichnet die Gesamtzahl von Artikeln aus allen Namensräumen

[4] Artikel bezieht sich auf den von meta zurückerhaltenen Wert articles und kennzeichnet nur alle Artikel im Artikel-Namensraum

[5] https://en.wikipedia.org/wiki/Wikipedia:Babel

Das Babel-System unterscheidet 5 Stufen der Sprachfähigkeit.

Muttersprache Die Sprache ist die Muttersprache

Stufe 4 Die Sprache wird fast so gut wie die Muttersprache beherrscht

Stufe 3 Der Nutzer ist in der Lage ohne Probleme Artikel zu schreiben.

Stufe 2 Der Nutzer ist in der Lage bestehende Artikel zu bearbeiten und sich an Diskussionen zu beteiligen.

Stufe 1 Der Nutzer ist in der Lage einen zu Artikel verstehen und kann einfache Fragen in dieser Sprache beantworten.

Für einige Sprachen existiert auch noch die Stufe 0, die bedeutet, dass diese Sprache nicht beherrscht wird. Das Babel-System ist freiwillig, daher kann es sein, dass viele Nutzer auf ihren Benutzerseiten keine Babel-Bausteine angegeben haben.

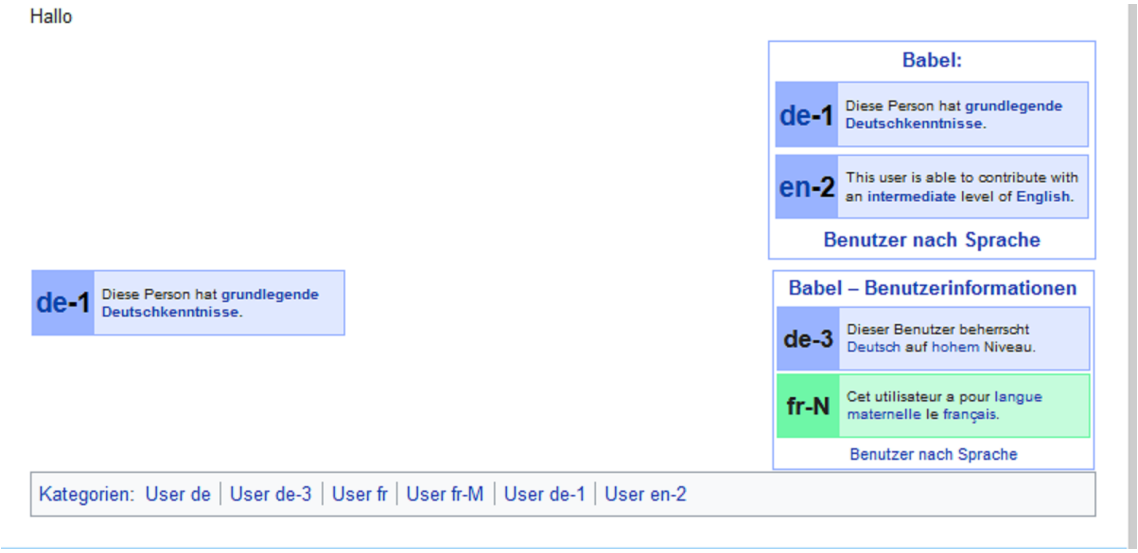

Abbildung 3.1.: Beispiel unterschiedlicher Babel-Bausteine

3.1.4. Verwandte Arbeiten zur Datenbeschaffung aus Wikipedia

In der Arbeit von Massa et al. [Mas11] wurden verschiedene Methoden zur Generierung von sozialen Netzwerkdaten aus Wikipedia untersucht. Die Forscher haben ausschließlich Artikel aus dem Benutzer Diskussionsseiten-Namensraum gesammelt. Dabei wurden nur Artikel aus der venezianischen Sprachversion benutzt. Es wurden Kommunikationsdaten zwischen den Nutzern gesammelt. Hierbei wurde der Name des Besitzers der Benutzer Diskussionsseite und die Namen der Nutzer, die auf dieser Seite eine Nachricht verfasst haben, gesammelt. Massa et al. haben drei verschiedene Methoden miteinander verglichen. Die erste Methode war das manuelle Sammeln der Nutzernamen durch Signaturen. Eine Signatur ist beispielsweise:

Es folgt meine Signatur –Arcazer (Diskussion) 14:25, 8. Jan. 2018 (CET).

Signaturen werden nicht automatisch gesetzt, sondern müssen vom Bearbeiter selber gesetzt werden. In der zweiten Methode wurde das Sammeln der Signaturen automatisiert. In der dritten Methode wurde mithilfe der Versionsgeschichte die Namen der Bearbeiter erfasst. Maasa et al. stellten unterschiede zwischen den drei Methoden fest. Nämlich in der Anzahl der erfassten Nutzer und in der Anzahl der erfassten Verbindungen zwischen den Nutzern. Die Unterschiede sind jedoch klein, falls in allen drei Methoden Bots und anonyme Nutzer ausgefiltert werden. Die Forscher sind der Meinung, dass die Methoden mit der Signaturerfassung nicht zuverlässig zwischen den verschiedenen Wikis funktionieren, da die Signaturenerfassung abhängig von den Faktoren Sprache oder Nutzernamensänderungen ist. Im Gegensatz dazu sei die Methode mit der Versionsgeschichte robuster. Nach der Meinung der Forscher konnten Bots mithilfe der Versionsgeschichte leichter entdeckt werden, da Bots keine Signaturen hinterlassen, dafür aber in der Versionsgeschichte Bearbeiternamen mit ”bot” im Namen hinterlassen.

3.1.5. Verwendete Methode der Datenbeschaffung

Für die Sammlung der Daten wurde die MediaWiki action API verwendet. Auf die Wikimedia Speicherauszüge wurde verzichtet, da die Datenmenge aus dem Speicherauszug zu groß ist. Der Speicherauszug im XML-Format der englischen Sprachversion besteht aus mehreren Terabyte. Die Verarbeitung mit diesen Daten wäre im Rahmen dieser Arbeit zu aufwendig. Außerdem wurde das Babel-System benutzt, um mehr Sprachinformationen über die Nutzer zu erhalten

3.2. Datenverarbeitung

In diesem Abschnitt wird der Vorgang der Datenverarbeitung beschrieben. Im Abschluss dieses Kapitels werden die erhaltenen Daten vorgestellt.

3.2.1. Technische System Voraussetzungen

Als System wurde ein Windows 10 Laptop mit einem Intel i3 Prozessor mit 2.40 Ghz und einem 8 GB Ram Arbeitsspeicher verwendet. Außerdem wurde die Java 64-Bit-Version verwendet.

3.2.2. Vorgang

Die Datenverarbeitung bestand aus vier Schritten, die wie folgt lauten:

Anfrage

Filtern

Nachbeschaffung

Datenvorbereitung für die soziale Netzwerkanalyse

Diese vier Schritte wurden mithilfe eines Java-Programms durchgeführt. Die Funktionsweise des Java Programms wird anhand der vier Schritte in den nächsten Abschnitten beschrieben. Der komplette Vorgang wird außerdem in der Abbildung 3.2 dargestellt.

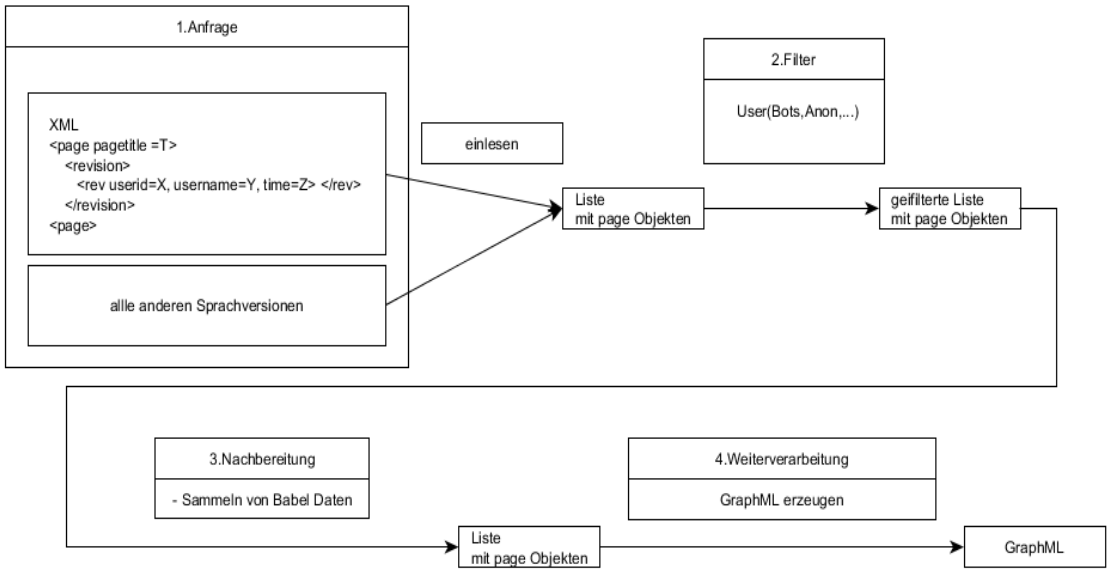

Abbildung 3.2.: Datenbeschaffung und Datenverarbeitung Vorgang

Anfrage

In diesem Schritt sollen die benötigten Sprach- und Kommunikationsinformationen aus Wikipedia extrahiert werden. Wie in Abschnitt 3.1.5 erwähnt, wurde hierfür die MediaWiki action API verwendet. Ein Ziel der Arbeit ist es Informationen aus den 11 Amtssprachen Südafrikas zu erhalten, nämlich Englisch, Sesotho, Xitsonga, Africaans, Setswana, Siswati, isiXhosa, Nord-Sotho, Süd-Ndebele, isiZulu und Tshivenda. Für neun Sprachen, also ohne Englisch und Süd-Ndebele, wurden alle Artikel und alle ihre Versionen mit der list=allrevision Funktion der MediaWiki action API angefragt. Für die Sprache Süd-Ndebele konnten keine Daten generiert werden, da zum Zeitpunkt dieser Arbeit die Süd-Ndebele Sprachversion auf Wikipedia nicht existierte. Die englische Sprache wurde gesondert behandelt, da mit der Funktion list=allrevision die Datenmenge zu groß wird. In der englischen Sprachversion wurden nur Artikel und ihre Versionen gesammelt, die von Autoren bearbeitet wurden, die auch in einer der anderen Sprachen beteiligt waren. Schrittweise wurden zuerst die neun Sprachversionen angefragt, dann die Nutzer bestimmt, und dann mithilfe der Nutzername alle relevanten Artikelversionen aus der englischen Wikipedia angefragt. Der Anfrage-Vorgang wird in Abbildung 3.3 dargestellt.

Zuständig für das Anfragen der Daten waren die beiden Methoden **grabWikiVersionHistory()** und **grabWikiVersionHistory_ByUser()** aus der Requester Klasse(siehe Anhang). Beide Methoden führen HTTP-Get-Anfragen an die MediaWiki action API durch.

Wichtig ist hierbei, dass alle eingelesen und weiterverarbeiteten Strings in UTF-8 codiert sind, da viele Artikelnamen oder Nutzernamen bei einer anderen Codierung Probleme bereiten könnten. Falls eine IDE für das Programmieren genutzt wird, muss in der IDE die Codierung auf UTF-8 umgestellt werden. Für die Rückgabe der Anfragen mit der MediaWiki action API standen wie in Abschnitt 3.1.2 beschrieben mehrere Formate zur Auswahl. In dieser Arbeit wurde als Rückgabeformat XML ausgewählt. Nach den Anfragen wurden die Ergebnisse aus den XML-Dateien ausgelesen und zugehörige Java-Objekte der Klassen Page und User generiert, um die weitere Nutzung zu vereinfachen. Die Java-Klassen Page und User sind in der Abbildung 3.4 beschrieben.

Aus der XML Rückgabe konnte für jeden Nutzer die Strings **userid** und **username** generiert werden. Für jeden Artikel konnten die Strings **pageid** und **pagetitle** generiert werden. Einzig die **username** sind über alle Sprachversionen hinweg eindeutig. Die **userid**, **pagerid** und **pagetitle** sind nur in ihrer zugehörigen Sprachversion eindeutig.

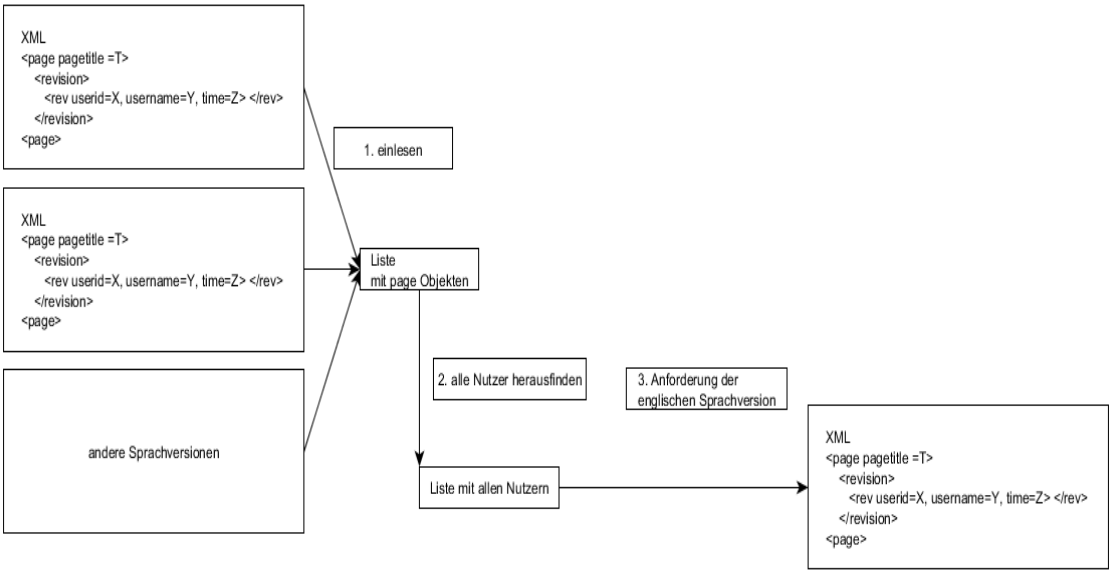

Abbildung 3.3.: Anfrage Vorgang für die englische Wikipedia

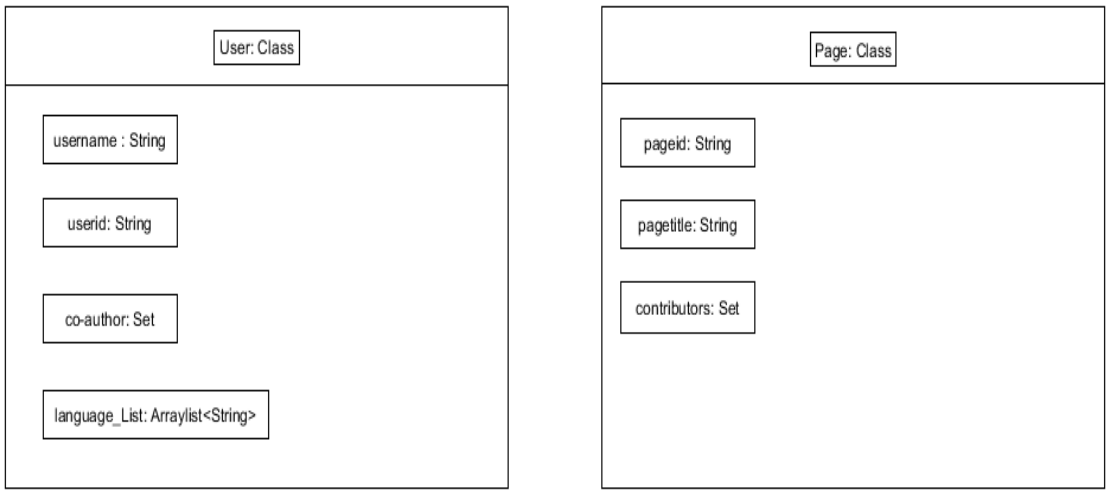

Abbildung 3.4.: Klassenbeschreibung der Kassen User und Page

Filtern

Bei der Filterung der Ergebnisse aus den Anfragen wurde darauf geachtet, dass ein **username** nicht mehrere **userids** hat oder ein **pagetitle** nicht mehrere **pageids** hat. Nutzer oder Artikel, bei denen dies eintrat, wurden entfernt. Das Filtern ist außerdem wichtig um aus der Menge der Artikeln alle Beiträge von Bots oder anonymen Nutzern zu entfernen. Bots verfälschen die Daten, da in dieser Arbeit nur das Kommunikationsverhalten von realen Personen interessant ist. Anonyme Nutzern müssen sich bei der Bearbeitung von Artikeln nicht anmelden und bei ihnen wird nur die IP-Adresse angezeigt. Anhand der IP-Adresse ist es schwer möglich eine Bearbeitung

einer konkreten Person zuzuweisen, da IP-Adressen dynamisch sein können oder unterschiedliche Personen eine IP-Adresse gemeinsam nutzen können. Beim Filtern der Bots wurde geprüft, ob sich der String "bot" unabhängig von Groß- oder Kleinschreibung im Nutzernamen befindet. Danach wurde Manuell noch nachgeprüft, ob nicht zu viele false positives ausgefiltert wurden. Die anonymen Nutzer konnten anhand der **userid** ausgefiltert werden, da anonyme Nutzer den **userid**-Wert 0 haben.

Im Programm ist die Klasse Filter (siehe Anhang) für das Filtern zuständig. Derzeit besitzt die Klasse Filter die Methoden **filterBot()** und **filterAnon()**. Die Klasse Filter lässt sich mit weiteren Filtern erweitern.

Nachbeschaffung

In diesem Schritt werden Methoden beschrieben, um Daten außerhalb der MediaWiki action API zu sammeln. Nach den Schritten Anfrage und Filtern konnten alle Namen von Nutzern, die einen Artikel in einen der neun Sprachen Sesotho, Xitsonga, Africaans, Setswana, Siswati, isiXhosa, Nord-Sotho, isiZulu und Tshivenda bearbeitet haben, extrahiert werden. Von diesen Nutzernamen wurden die zugehörigen Benutzerseiten angefragt. Dabei wurden die Wikipedia Benutzerseiten in allen neun Sprachversionen und zusätzlich in der Englischen Sprachversion angefragt. Das Ziel dieses Schrittes ist es, die Informationen der Babel-Bausteine zu bekommen. Ein Nutzer hat mehrere Möglichkeiten Babel-Bausteine anzulegen, dazu muss er einen der folgenden Bausteinen in seine Benutzerseite reinschreiben:

{{#babel: BABELCODE1| BABELCODE2}} ist eine direkte Funktion innerhalb Wikis, die die MediaWiki-Software nutzen

{{Babel|BABELCODE1| BABELCODE2}} ist eine Vorlage[6] auf Wikipedia

{{babel BABELCODE1| BABELCODE2}} ist eine Vorlage[6] auf Wikipedia

{{User BABELCODE1| BABELCODE2}} ist eine Vorlage[6] auf Wikipedia.

{{user BABELCODE1| BABELCODE2}} ist eine Vorlage[6] auf Wikipedia

Der **BABELCODE** ist nach den im Abschnitt 3.1.3 beschrieben Sprachfähigkeiten gerichtet:

xx Muttersprache

xx-4 Stufe 4

xx-3 Stufe 3

xx-2 Stufe 2

xx-1 Stufe 1

[6] Mit Vorlagen kann man auf Wikis der MediaWiki-Software kleine Textbausteine programmieren, dazu stehen einfache Konstrukte einer Programmiersprache zur Verfügung. Mit Vorlagen lassen sich z.B- Tabellen oder Diagramme erstellen.

Die **xx** beschreiben den Code der Sprache. Zum Beispiel erzeugt {{#babel: de-2| en}} einen Babel-Textblock, der angibt, dass Deutsch auf Stufe 2 beherrscht wird und Englisch als Muttersprache beherrscht wird.

Es existieren noch weiter Möglichkeiten Babel-Bausteine einzufügen, zum Beispiel durch Artikeleinbindung oder selbst erstellte Vorlagen. Bei der Artikeleinbindungsmethode wird ein neuer Artikel mit Babel-Bausteinen angelegt und dann dieser Artikel in die eigene Benutzerseite eingebunden. Die Artikeleinbindungsmethode ist daher schwer zu entdecken. Außerdem ist es noch möglich eigene Vorlagen zu definieren, welche Babel-Bausteine darstellen. Mithilfe der Wiki-API ist es möglich die {{#babel: de-1}} Variante zu erkennen. Die Wiki-API funktioniert leider nicht für die anderen Vorlagen, daher wurden in dieser Arbeit die Benutzerseiten extrahiert und darin nach dem Vorkommen von {{#babel: BABELCODE1| BABELCODE2}}, {{Babel|BABELCODE1| BABELCODE2}}, {{babel BABELCODE1| BABELCODE2}}, {{User BABELCODE1| BABELCODE2}} und {{user BABELCODE1| BABELCODE2}} gesucht.

Datenvorbereitung für die soziale Netzwerkanalyse

Aus den gesammelten Daten aus den Schritten Anfrage, Filtern und Nachbearbeitung konnte wie in der Abbildung 3.2 beschrieben eine Liste mit page-Objekten erstellt werden. Aus dieser Liste wurden für jeden Artikel all seine Bearbeiter erfasst, damit lassen sich für jeden einzelnen Nutzer, die Co-Autoren bestimmen. Aus diesen Informationen über Nutzer und Co-Autoren soll in diesem Schritt eine Datei generiert werden, die für eine soziale Netzwerkanalyse geeignet ist. In dieser Arbeit wurde dafür eine GraphML-Datei erstellt. GraphML[7] ist ein Dateiformat mit der Graphen/Netzwerke erstellt werden können. GraphML ist der Nachfolger der Graph Modelling Language(GML)[8]. In Listing 3.5 wird exemplarisch der Aufbau einer GraphML-Datei vorgestellt.

[7] http://graphml.graphdrawing.org/
[8] http://www.opengeospatial.org/standards/gml

Listing 3.5 Exemplarischer Aufbau einer GraphML-Datei

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
    http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">

     <!-- Hier koennen zusaetzliche Attribute fuer die Knoten definiert werden
         -->
     <key id="d0" for="node" attr.name="ATTRIBUTENAME_KNOTEN" attr.type="string">
          <default>"DEFAULTWERT"</default>
     </key>

     <!--Hier koennen zusaetzliche Attribute fuer die Kanten definiert werden -->
     <key id="d1" for="edge" attr.name="ATTRIBUTNAME_KANTE" attr.type="double"/>

     <!--Hier wird der Graph definiert -->
     <graph id="G" edgedefault="undirected">
          <node id="n1"> <data key="d0">green</data></node>
          <node id="n2"><data key="d0">blue</data></node>

          <edge id="e0" source="n1" target="n2"> <data key="d1">1.0</data>
              </edge>
     </graph>

</graphml>
```

In dieser Arbeit wurden Nutzernamen als id für die Knoten angegeben. Kanten wurden zwischen zwei Nutzern gebildet, wenn sie zusammen in einem Artikel editiert haben. Die Kanten zwischen den Nutzern wurden als ungerichtet festgelegt. Netzwerkvisualisierungen für die 10 Sprachversionen Sesotho, Xitsonga, Africaans, Setswana, Siswati, isiXhosa, Nord-Sotho, isiZulu, Tshivenda und Englisch sind im Anhang beigefügt.

Im Programm ist die Klasse Writer für die Generierung einer GraphML-Datei zuständig. Die Methode writeToGraphML() ist für die Erstellung einer GraphMl-Datei zuständig.

3.2.3. Erhaltene Daten

In diesem Abschnitt werden die extrahierten Informationen beschrieben und eine Beschreibung der extrahierten Daten gegeben. Aus den in den Schritten Anfrage, Filtern und Nachbearbeitung erhaltenen Datensätzen konnten folgende Informationen über die Nutzer aus Wikipedia abgeleitet werden:

- Den Namen(Benutzernamen) der Nutzer.
- Die Co-Autorennamen, mit denen der Nutzer zusammen an einen Artikel bearbeitet hat.
- Die Sprachversion, in der der Nutzer einen Artikel bearbeitet hat.
- Die Sprachfähigkeiten, die der Nutzer auf seiner Benutzerseite angeben hat.

Beschreibung der Rohdaten

Im diesem Abschnitt werden die Einzelheiten über den aus Wikipedia erhaltenen Datensatz beschrieben. Alle Seiten der Wikipedia, aus dem Artikel-Namensraum, wurden mithilfe der MediaWiki action API zwischen dem 15. November 2017 und 20. November 2017 generiert. Die Seiten stammen aus den 10 Sprachversionen Englisch, Sesotho, Xitsonga, Africaans, Setswana, Siswati, isiXhosa, Nord-Sotho, isiZulu und Tshivenda. Informationen über diesen Datensatz befindet sich in Tabelle 3.1. Informationen über die erhaltenen Daten aus Babel befinden sich in der Tabelle 3.2. Aus Babel konnte am 14.1.2018 insgesamt von 4701 verschiedenen Nutzer deren Benutzerseiten extrahiert werden, daraus konnten 2014 Babel-Bausteine entdeckt werden.

	Nutzer*	**Nutzer****	**Bots**	**Anonyme Nutzer**	**Seiten**
Africaans (af)	24124	6132	193	17799	70065
isiXhosa (xh)	753	344	85	324	1142
isZulu (zu)	1244	438	99	707	1376
Nors-Sotho (nso)	330	121	39	172	8178
Sesotho (st)	436	184	85	167	640
Setswana (tn)	649	245	85	319	798
Siswati (ss)	486	176	86	224	546
Tshivenda (ve)	456	152	79	207	371
Xitsonga (ts)	461	162	92	207	732
Englisch (en)	4282***	4282***	0	0	4254596
Total	26646	6870	208	19583	83848

*Nutzer mit Bots und Anonymen Nutzern
**Nutzer ohne Bots und Anonymen Nutzern
***Durch die in Abschnitt 3.2.2 beschriebene Anfrage Methode wurden 4300 Nutzer zurückgegeben, davon sind aber 18 Nutzer nicht in den anderen Sprachen vorgekommen. Diese 18 Nutzer wurden entfernt.

Tabelle 3.1.: Erhaltene Daten aus Wikipedia

	Stufe 0	**Stufe 1**	**Stufe 2**	**Stufe 3**	**Stufe 4**	**Muttersprache**
Africaans (af)	156	60	38	16	8	67
isiXhosa (xh)	13	0	0	0	0	0
isZulu (zu)	19	3	2	1	0	2
Nors-Sotho (nso)	6	1	0	3	0	0
Sesotho (st)	7	2	1	0	0	1
Setswana (tn)	10	3	0	2	0	1
Siswati (ss)	10	1	0	0	0	1
Tshivenda (ve)	4	0	0	0	0	0
Xitsonga (ts)	10	0	0	0	0	0
Englisch (en)	13	61	251	358	178	148

Tabelle 3.2.: Erhaltene Daten von Babel

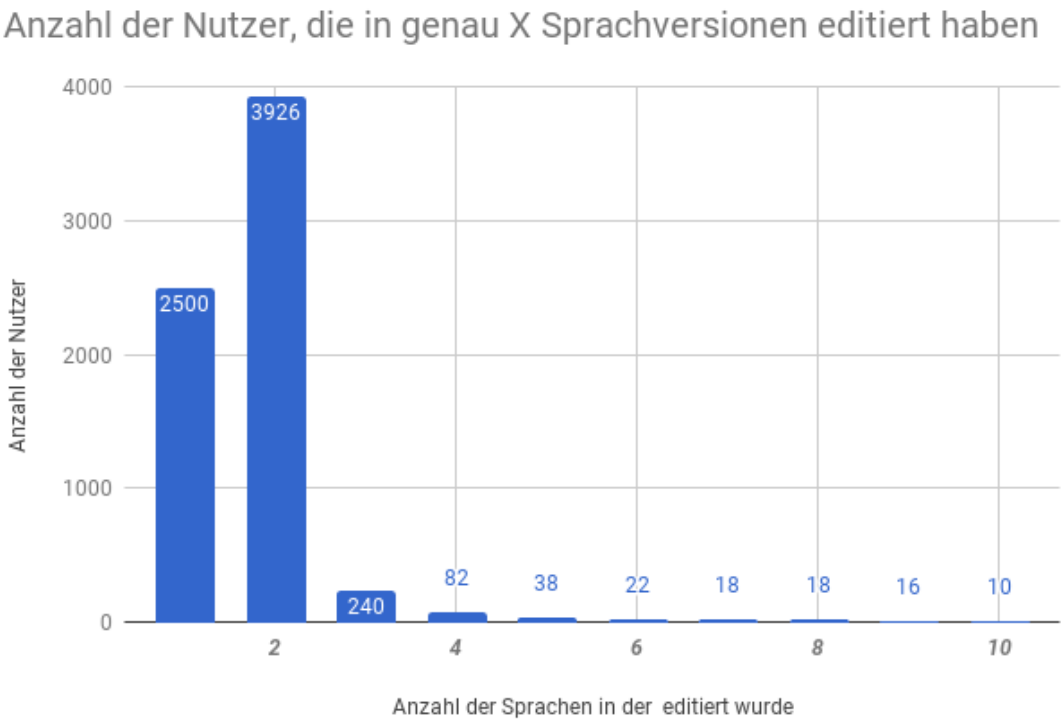

Abbildung 3.5.: Häufigkeitsverteilung in der Anzahl, der editierten Sprachversionen

In Abbildung 3.5 wird die Anzahl der Nutzer beschrieben, die in genau X Sprachversionen editiert haben. Die Sprachversionen beziehen sich nur auf die zehn beschafften Sprachversionen. Es ist durchaus möglich, dass diese Nutzer noch in anderen Sprachversionen editiert haben. Der größte Anteil der Nutzer hat in genau zwei Sprachversionen editiert. Auffallend ist, dass 44 Nutzer in acht oder mehr Sprachversionen editiert haben.

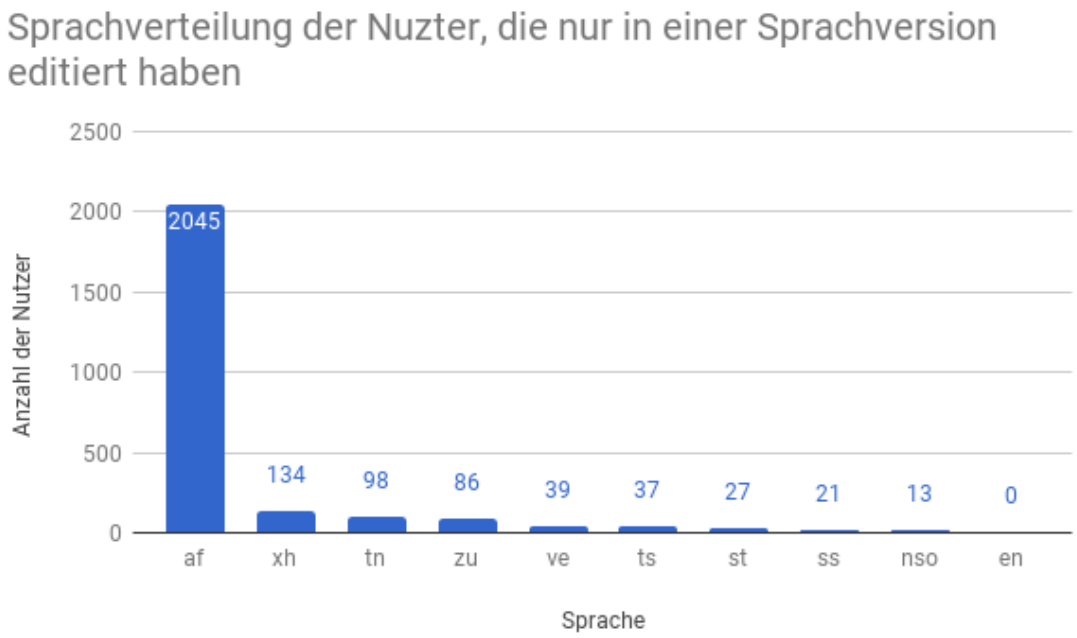

Abbildung 3.6.: Sprachverteilung der Nutzer, die in genau einer Sprachversion editiert haben

In Abbildung 3.6 wird die Anzahl der Nutzer beschrieben, die in genau einer Sprachversion editiert haben. Die Englische Sprachversion hat den Wert null, weil aufgrund der in 3.2.2 beschriebenen Anfragemethode alle Nutzer aus der englischen Sprachversionen mindestens in einer anderen Sprachversion editiert haben mussten. Die Gesamtzahl der Nutzer, die genau in genau einer Sprachversion editiert haben betrug 2500 (siehe Abb. 3.5) Die Verteilungen bezogen auf die tatsächliche Anzahl der Nutzer in der jeweiligen Sprachversion beträgt:

- 33% für Africaans(af)
- 39% für isiXhosa (xh)
- 19% isZulu (zu)
- 13% Nord-Sotho (nso)
- 14% Sesotho (st)
- 40% Setswana (tn)
- 12% Siswatis (ss)
- 25% Tshivenda (ve)
- 23% Xitonga (ts)

Diese Verteilung lässt sich in Abbildung 3.8 erkennen. In der Abbildung befinden sich die Sprachen Sesotho(ss) und Nord-Sotho (nso), welche wenige exklusive Nutzer besitzen, zentraler. Africaans (tn), isXhosa (xh) und Setswana (tn) befinden sich dagegen näher am Rand, da diese Sprachen mehr Nutzer haben, die in dieser Sprache exklusiv editiert haben.

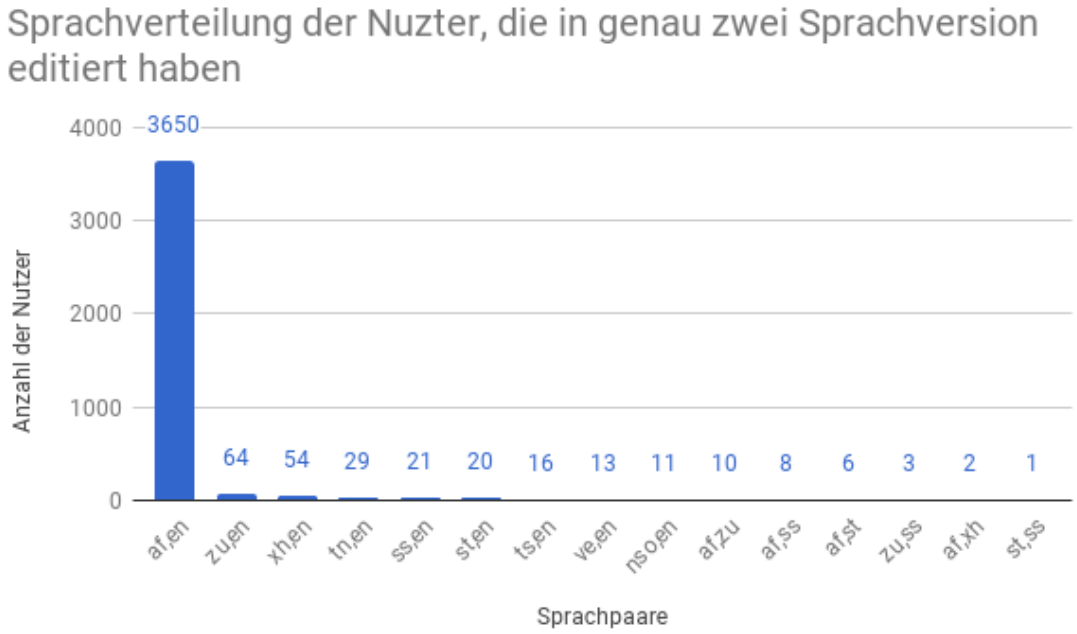

Abbildung 3.7.: Sprachpaarverteilung der Nutzer, die in genau zwei Sprachversionen editiert haben

In Abbildung 3.7 wird die Anzahl der Nutzer beschrieben, die in genau zwei Sprachversion editiert haben. Auffällig ist, dass die meisten Sprachpaare Englisch enthalten. Nur 30 von den 3926 (siehe Abb. 3.5) Nutzern, die in genau zwei Sprachversionen editiert haben, haben in einer Sprachpaarkombination ohne Englisch editiert.

In den Sprachkombinationen aus genau drei Sprachen wurde entdeckt, dass alle 240 (siehe Abb. 3.5) Nutzer auch der Englischen Sprachversion editiert haben. Die Sprachkombinationen der Nutzer, die in mehr Sprachversionen editiert haben, befindet sich im Anhang.

	Durchschnittliche Anzahl editierter Sprachen	Nutzeranzahl
Africaans (af)	$1{,}812(SD = 0.542)$	6132
isiXhosa (xh)	$3{,}201(SD = 2.187)$	344
isZulu (zu)	$3{,}527(SD = 1.825)$	438
Nors-Sotho (nso)	$5{,}083(SD = 2.528)$	121
Sesotho (st)	$4{,}679(SD = 2.474)$	184
Setswana (tn)	$3{,}502(SD = 2.488)$	245
Siswati (ss)	$4{,}528(SD = 2.460)$	176
Tshivenda (ve)	$4{,}575(SD = 2.700)$	152
Xitsonga (ts)	$4{,}475(SD = 2.700)$	162
Englisch (en)	$2{,}207(SD = 0.376)$	4282

Tabelle 3.3.: Die Tabelle beschreibt die durchschnittliche Anzahl editierter Sprachen von Nutzern aus bestimmten Sprachversionen

In Tabelle 3.3 wird die durchschnittliche Anzahl editierter Sprachen von Nutzern aus den verschiedenen Sprachversionen beschrieben. Es ist auffällig, dass Nutzer die sich in den zwei großen Sprachversionen Englisch und Africaans befinden, durchschnittlich in weniger verschiedenen Sprachversionen editiert haben. Außerdem ist in den beiden Sprachversionen die Standardabweichung viel geringer, als bei den Nutzern aus anderen Sprachversionen.

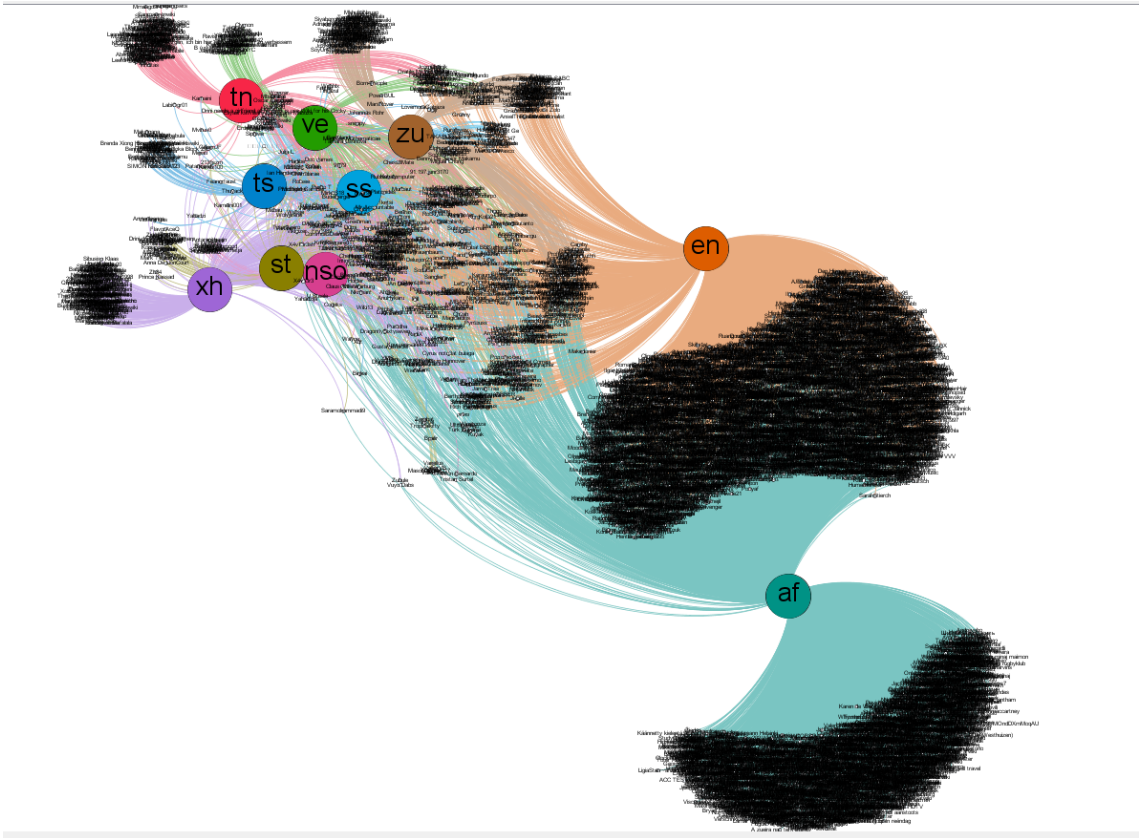

Abbildung 3.8.: Ein Netzwerk zur Beschreibung der Sprachzugehörigkeit von Nutzern. Die Namen in Schwarz stellen die verschiedenen Nutzer da. Nutzer werden mit Sprachknoten, deren sie zugehörig sind, durch Kanten verbunden. Die Kantenfarbe entspricht der Farbe der zugehörigen Sprachversion. Im Netzwerk befinden sich Nutzer, die in mehreren Sprachversionen editiert haben, zentraler. Nutzer, die in wenig Sprachversionen editiert haben befinden sich näher am Rand.

4. Soziale Netzwerkanalyse

In diesem Kapitel werden einige Netzwerkmetriken beschrieben, die häufig bei sozialen Netzwerkanalysen gemessen werden. Zu den Metriken werden auch einige Vergleichswerte von anderen sozialen Netzwerken aufgelistet. Anschließend werden die Metriken für die extrahierten Daten aus den 10 Sprachversionen berechnet. Die Metriken werden sowohl für jede Sprachversion einzeln als auch für den kombinierten Datensatz gemessen. Das Datenformat für die soziale Netzwerkanalyse wurde zuvor in Abschnitt 3.2.2 beschrieben.

4.1. Homophilie

Die Homophilie (engl. *assortative mixing*) beschreibt nach Newman [New02], die Neigungung von Individuen, bevorzugt Interaktionen mit anderen Individuen einzugehen, die ihnen ähnlich sind. Newman beschreibt drei Abhängigkeiten der Neigungen innerhalb von Netzwerken wie folgt:

Norminale Eigenschaften (engl.*enumerative characteristics*) beschreiben Abhängigkeiten, die Klassifiziert werden können. Beispiele sind Geschlecht, Herkunft oder Beruf.

Skalare Eigenschaften (engl. *scalar characteristics*) beschreiben Abhängigkeiten, die durch Zahlenwerte festgelegt werden. Beispiele sind Alter oder Gehalt.

Knotengrad beschreibt Abhängigkeit anhand des Knotengrads. Knoten mit ähnlich großen Knotengraden sind öfter miteinander verbunden

Einge Werte über Homophilie von anderen sozialen Netzwerken wurden aus [New02],[New03] und [New10, S.237] entnommen und lauten:

- r=0.208 für ein Kollaborationsnetzwerk aus Filmschauspielern
- r=0.120 für ein Co-Autorennetzwerk aus Mathematikern
- r=0.363 für ein Co-Autorennetzwerk aus Physikern
- r=0.127 für ein Co-Autorennetzwerk aus Biologen

In [New10, S.267-268] wurde zudem beschrieben, dass soziale Netzwerke häufig positive Homophiliewerte besitzen, weil Individuen sich oft in verschiedenen Gruppen voneinander abgrenzen. Homophile Netzwerke besitzen positive Werte und nicht homophile Netzwerke negative Werte. Die Homophiliewerte für die generierten Daten aus Wikipedia sind in Tabelle 4.1 beschrieben.

	Homophilie (Knotengrad)	Knotenanzahl	Kantenanzahl
Africaans (af)	−0,368	6132	109446
isiXhosa (xh)	−0,286	344	2142
isZulu (zu)	−0,307	438	3948
Nors-Sotho (nso)	−0,501	121	499
Sesotho (st)	−0,288	184	1163
Setswana (tn)	−0,274	245	1752
Siswati (ss)	−0,425	176	758
Tshivenda (ve)	−0,281	152	824
Xitsonga (ts)	−0,310	162	783
Englisch (en)	−0,417	4282	1313052
Kombiniert	−0,393	6870	1409685

Berechnungen wurden mit der assortative_degree() Funktion aus [Igra] berechnet.

Tabelle 4.1.: Homophiliewerte der Netzwerke

4.2. Clusterkoeffizient

Der Clusterkoeffizient charakterisiert die Transitivität der Beziehungen zwischen den Individuen [New10, S. 198ff]. In einem Netzwerk ist der Clusterkoeffizient Wert 1 der größte und 0 der kleinste. Bei dem Wert 1 wäre jeder Knoten mit jedem direkt verbunden. Transitivität bedeutet, dass wenn in einem Netzwerk drei Knoten A,B und C existieren, alle drei individuell miteinander verbunden sind. Ein Beispiel aus der realen Welt für die Transitivität wäre das Freundschaftsverhältnis: Falls A mit B befreundet ist und B mit C, dann ist auch A mit C befreundet. Intransitiv wäre sie, wenn eines der Individuen nicht mit einem der beiden anderen befreundet wäre.

Einge Werte über den Clusterkoeffizient von anderen sozialen Netzwerken wurden aus [New10, S.237] entnommen und lauten:

- C=0.2 für ein Kollaborationsnetzwerk aus Filmschauspielern
- C=0.15 für ein Co-Autorennetzwerk aus Mathematikern
- C=0.45 für ein Co-Autorennetzwerk aus Physikern
- C=0.088 für ein Co-Autorennetzwerk aus Biologen

Die Clusterkoeffiziente für die generierten Daten aus Wikipedia sind in Tabelle 4.2 beschrieben.

	Clusterkoeffizient	Knotenanzahl	Kantenanzahl
Africaans (af)	0,062	6132	109446
isiXhosa (xh)	0,321	344	2142
isZulu (zu)	0,330	438	3948
Nors-Sotho (nso)	0,241	121	499
Sesotho (st)	0,401	184	1163
Setswana (tn)	0,397	245	1752
Siswati (ss)	0,230	176	758
Tshivenda (ve)	0,456	152	824
Xitsonga (ts)	0,358	162	783
Englisch (en)	0,512	4282	1313052
Kombiniert	0,488	6870	1409685

Berechnungen wurden mit der transitivity() Funktion aus [Igrc] berechnet.

Tabelle 4.2.: Clusterkoeffizient der Netzwerke

4.3. Durchschnittliche Pfadlänge

Wie in 2.1.3 beschrieben, sind Pfadlängen zwischen zwei Knoten in einem sozialen Netzwerk relativ kurz.

Einge Werte über die durchschnittliche Pfadlänge von anderen sozialen Netzwerken wurden aus [New10, S.237] entnommen und lauten:

- l=3.48 für ein Kollaborationsnetzwerk aus Filmschauspielern
- l=7.57 für ein Co-Autorennetzwerk aus Mathematikern
- l=6.19 für ein Co-Autorennetzwerk aus Physikern
- l=4.92 für ein Co-Autorennetzwerk aus Biologen

Die durchschnittliche Pfadlänge für die generierten Daten aus Wikipedia sind in Tabelle 4.3 beschrieben.

	Durchschnittliche Pfadlänge	Knotenanzahl	Kantenanzahl
Africaans (af)	2,078	6132	109446
isiXhosa (xh)	2,342	344	2142
isZulu (zu)	2,167	438	3948
Nors-Sotho (nso)	2,061	121	499
Sesotho (st)	2,171	184	1163
Setswana (tn)	2,226	245	1752
Siswati (ss)	2,302	176	758
Tshivenda (ve)	2,307	152	824
Xitsonga (ts)	2,243	162	783
Englisch (en)	1,878	4282	1313052
Kombiniert	2,042	6870	1409685

Berechnungen wurden mit der mean_distance() Funktion aus [Igrb] berechnet.

Tabelle 4.3.: Durchschnittliche Pfadlänge der Netzwerke

5. Diskussion

In dieser Arbeit konnten Informationen über Sprachkenntnis und Sprachnutzung aus Wikipedia extrahiert werden. Dabei wurden nur einfache Filterungsmethoden auf die extrahierten Daten angewendet. Nämlich das Filtern von Nutzern, die die Zeichenkette “bot” im Namen haben und das Filtern von anonymen Nutzern.

Für weitere Datenextraktionen können weitere Informationen wie Bearbeitungszeitpunkte oder geschriebene Artikeltexte analysiert werden. Die wie in Abschnitt 3.1.2 beschrieben über die MediaWiki action API zur Verfügung standen. Mithilfe von Bearbeitungszeitpunkten wäre es möglich rückzuschließen, in welchen Zeiträumen der Nutzer eine Sprache verwendet hat. Daraus können Muster während der Sprachnutzung entdeckt werden. In den Mustern kann untersucht werden, ob ein Nutzer eine Sprache nur einen Tag lang verwendet hat oder über mehrere Monate lang verwendet hat. Dadurch kann besser abgeschätzt werden, ob der Nutzer eine Sprache wirklich beherrscht. Artikeltexte sollten ebenfalls analysiert werden. Hierbei sollte geprüft werden, ob der Nutzer wirklich etwas sinnvolles geschrieben hat oder nicht. Es muss aber beachtet werden, dass eine große Datenmenge verarbeitet werden muss, um die Bearbeitungszeitpunkte und Artikeltexte einzelner Nutzer zu untersuchen.

In dieser Arbeit war ursprünglich vorgesehen auch Weiterleitungsartikel (siehe Abschnitt 2.3), und die daran beteiligten Nutzer zu entfernen. Es konnte aber nicht abgeschätzt werden, wie groß die Anzahl der Weiterleitungsartikel in der englischen Sprachversion ist, daher wurden diese nicht extrahiert. Es sollte zwar selten vorkommen, dass zwei Nutzer zusammen an einem Weiterleitungsartikel gearbeitet haben, aber dieses Problem könnte negative Auswirkungen auf den generierten Datensatz haben.

Aus Babel konnten, wie in Abschnitt 3.2.3 in Tabelle 3.2 beschrieben, relativ wenige Informationen über die Sprachkenntnis erhalten werden. Für weitere Datenextraktion sollten Informationen über Sprachkenntnisse aus Babel mit in die Bewertung über die Relevanz eines Nutzers einfließen. So sollten Nutzer, die angegeben haben, die Sprache nicht zu beherrschen, ausgefiltert werden.

In Abschnitt 3.2.3 Abbildung 3.5 über die Häufigkeitsverteilung der Anzahl editierter Sprachversionen konnte entdeckt werden, dass nach dem Filtervorgang immerhin noch 44 Nutzer vorhanden waren, die in mehr als acht Sprachversionen editiert haben. Ebenfalls gab es viele Nutzer, die in mehr als fünf Sprachversionen editiert haben. Außerdem wurde in Abschnitt 3.2.3 Tabelle 3.3 beschrieben, dass kleinere Sprachversionen Nutzer hatten, die in durchschnittlich mehr Sprachversionen editiert haben. Ebenso ist die Standardabweichung für die kleineren Sprachversionen viel höher. Daher sollten Nutzer, die in vielen Sprachversionen editiert haben, auf ihr Editierverhalten überprüft werden. Für weitere Datenextraktionen sollte die Relevanz eines Nutzer anhand der Anzahl der editierten Sprachversionen überprüft werden.

Bei der sozialen Netzwerkanalyse konnte bei den Metriken Auffälligkeiten festgestellt werden. Der Homophiliewert war bei allen Netzwerken negativ. Dies bedeutet, dass sich bei den Autoren

kaum Gruppen gebildet haben. Wie in Abschnitt 4.1 beschrieben haben die meisten realen sozialen Netzwerke positive Werte, da reale soziale Netzwerke oft viele unterschiedliche Gruppen enthalten, die sich leicht voneinander abgrenzen. Um diesen Wert zu interpretieren müssen weitere Informationen über das Editierverhalten der Nutzer gesammelt und überprüft werden. Der positive Homophiliewert deutet daraufhin, dass jeder Nutzer an vielen verschiedenen Artikeln mitgewirkt hat und daher kaum Gruppenbildungen zwischen den Nutzern auftreten.

Der Clusterkoeffizient war besonders auffällig bei der der Africaans Sprachversion, dieser war viel niedriger als bei den anderen Sprachversionen. Dieser Wert sollte in einer weiteren Datenextraktion mit Daten, die anhand der zuvor beschriebenen zusätzlichen Kriterien gefiltert werden, nochmals überprüft werden.

Die durchschnittliche Pfadlänge bei allen Sprachversionen war kürzer als bei den anderen Vergleichswerten. Auch hier müssen weitere Informationen über das Editierverhalten gesammelt werden, um die durchschnittliche Pfadlänge zu interpretieren. Die niedrige durchschnittliche Pfadlänge könnte dadurch erklärt werden, dass es einzelne Nutzer gab, die auf sehr viele Seiten editiert haben und dadurch mit fast jedem anderen Nutzer vernetzt sind. Jeder Nutzer ist dann durch diesen einzelnen Nutzer mit den anderen verbunden. Schon ein einzelner Nutzer, der mit fast jedem anderen Nutzer verbunden ist, kann die durchschnittliche Pfadlänge stark beeinflussen. Daher sollten Nutzer, die sehr viele Co-Autoren besitzen auch näher untersucht werden.

Folgende Kriterien, die Auswirkungen auf den generierten Datensatz aus Wikipedia haben, konnten entdeckt werden:

- Nutzertyp (realer Nutzer, Bot oder anonymer Nutzer)
- Anzahl editierter Sprachversionen eines Nutzers
- Bearbeitungshäufigkeit eines Nutzers
- Artikeltextqualität eines Nutzers
- Sprachkenntnis aus dem Babel-System
- Anzahl Co-Autoren eines Nutzers

Generell sollte ein Filtermodell entwickelt werden, welches die Zuverlässigkeit der Informationen über einen Nutzer bestimmt. In diesem Modell sollten dann die zuvor beschriebenen Kriterien benutzt werden, um die Relevanz von den Informationen über den Nutzer zu bewerten. Dadurch kann entschieden werden, ob ein Nutzer mit in das Co-Autorennetzwerk aufgenommen werden soll oder nicht.

Abschließend lässt sich sagen, dass in dieser Arbeit einfache Methoden gefunden und beschrieben wurden, mit denen Informationen aus Wikipedia über Sprachkenntnisse und Sprachnutzung extrahiert werden können. Außerdem konnte ein Einblick auf die Strukturen dieser Daten gegeben werden, woraus sich Kriterien ableiten lassen, um bei weiteren Datenextraktionen qualitativ bessere Informationen zu erhalten.

A. Anhang

Sprachkombinationen der Nutzer, die in vier oder mehr Sprachversionen editiert haben:

Vier Sprachversionen af,zu,ss,en=7, af,zu,st,en=6, af,zu,ve,en=5, af,xh,st,en=3, af,xh,ss,en=2, af,tn,ss,ts=1

Fünf Sprachversionen af,xh,zu,st,en=3, af,zu,st,ss,en=2, af,xh,zu,ss,en=1

Sechs Sprachversionen af,xh,zu,st,ss,en=2, af,xh,st,tn,ve,en=1

Sieben Sprachversionen af,xh,nso,st,ss,ve,en=1

Acht Sprachversionen af,xh,zu,nso,st,ss,ts,en=2, af,xh,zu,nso,tn,ve,ts,en=1

Neun Sprachversionen af,xh,zu,st,tn,ss,ve,ts,en=5, af,xh,zu,nso,st,tn,ss,ve,ts=3, af,xh,zu,nso,st,tn,ss,ts,en=2, af,xh,zu,nso,st,tn,ve,ts,en=1

Zehn Sprachversionen af,xh,zu,nso,st,tn,ss,ve,ts,en=10

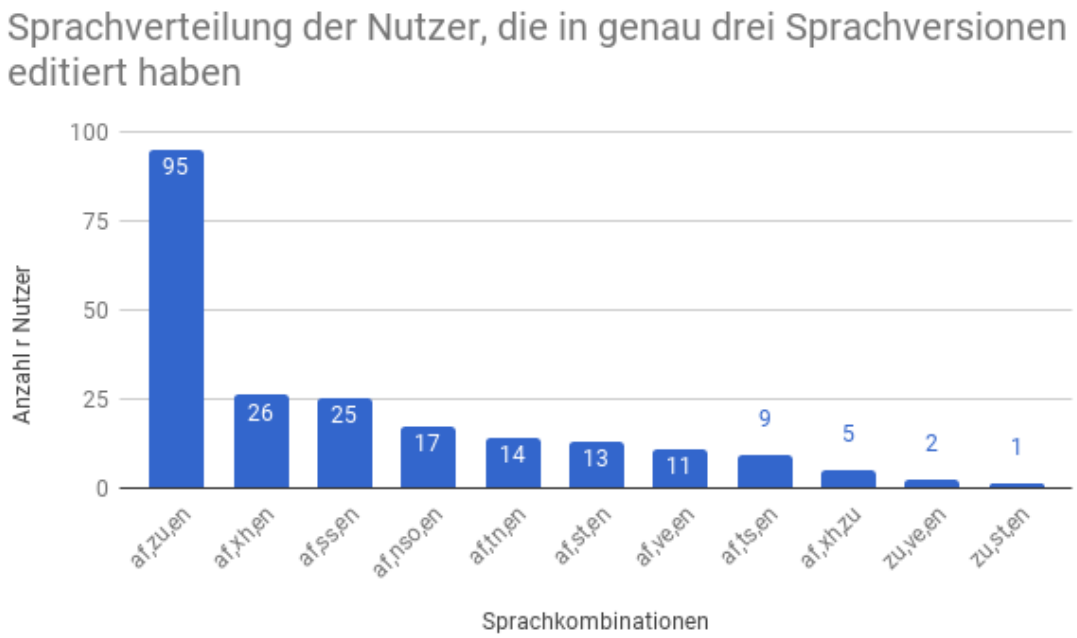

Abbildung A.1.: Sprachkombinationen der Nutzer, die in genau drei Sprachversionen editiert haben

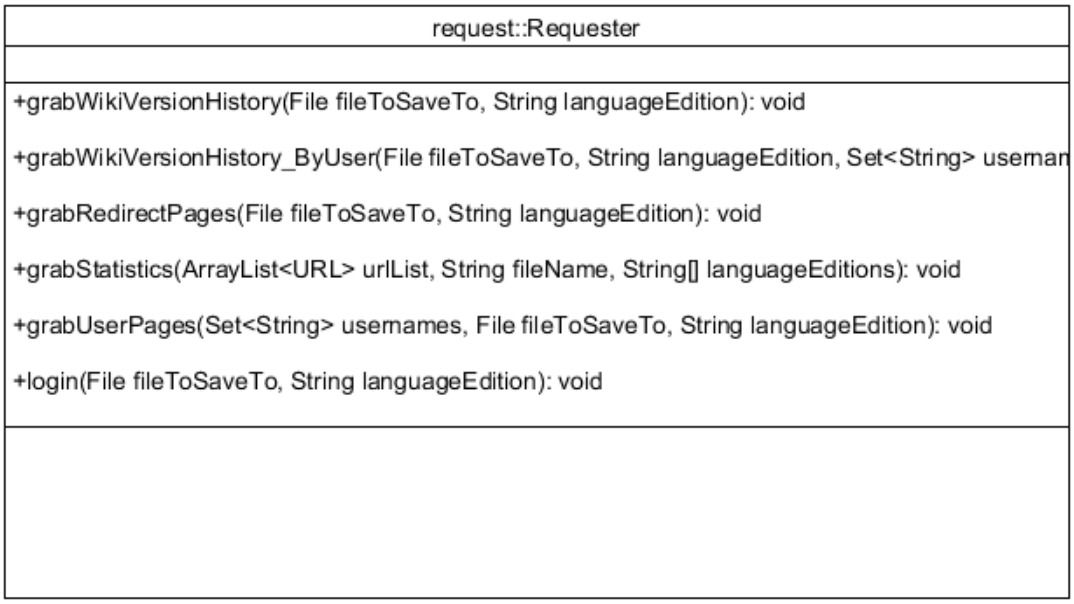

Abbildung A.2.: Request Klasse

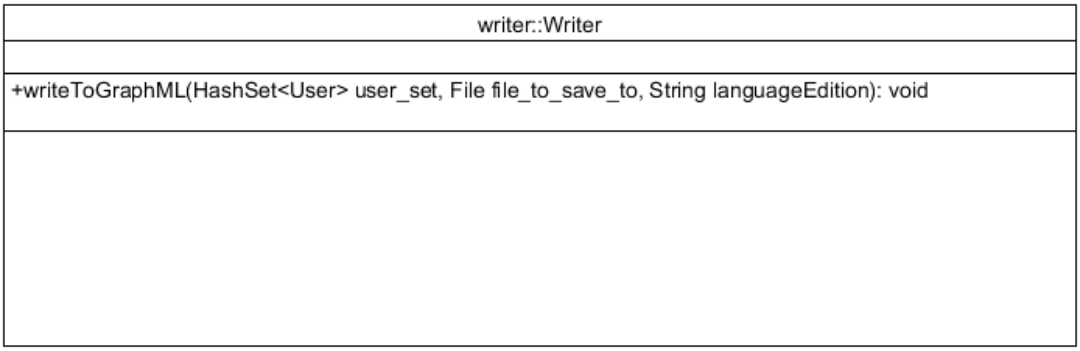

Abbildung A.3.: Filter und Writer Klasse

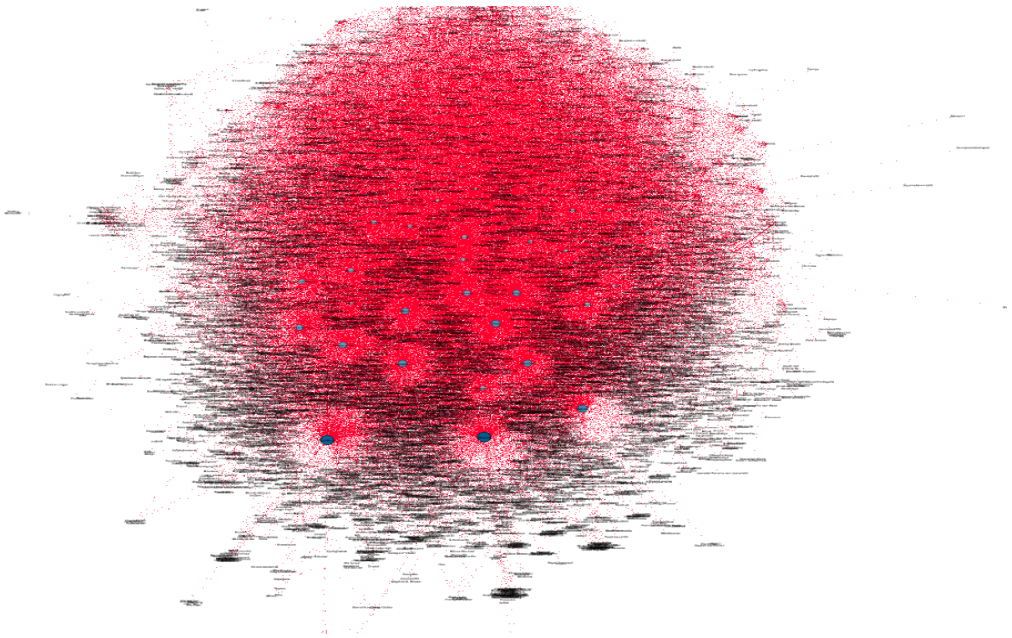

Abbildung A.4.: Africaans Netzwerk

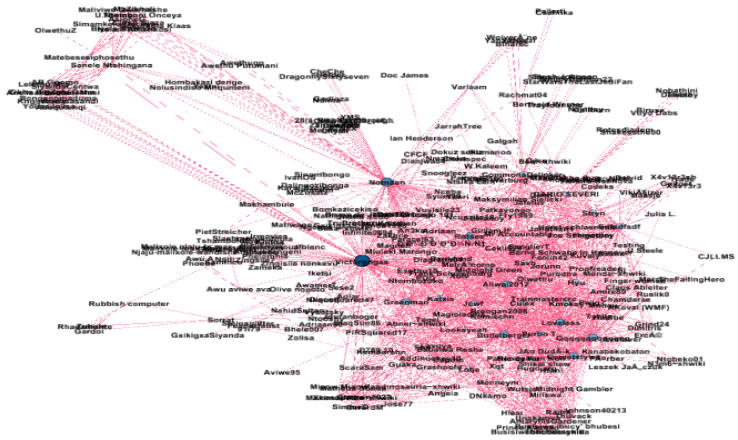

Abbildung A.5.: isiXhosa Netzwerk

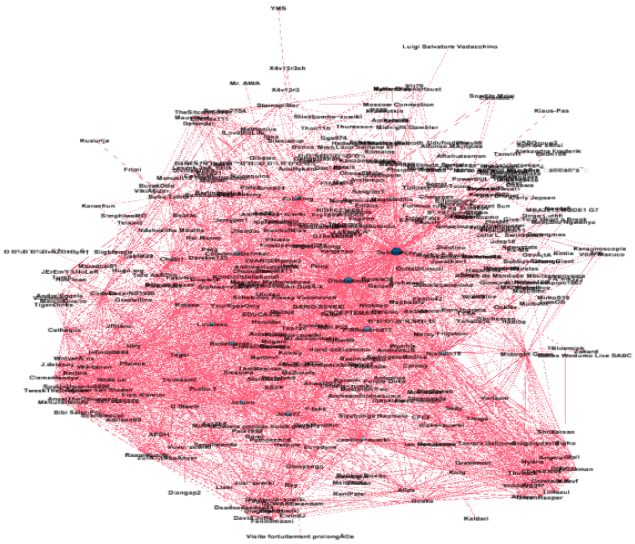

Abbildung A.6.: isZulu Netzwerk

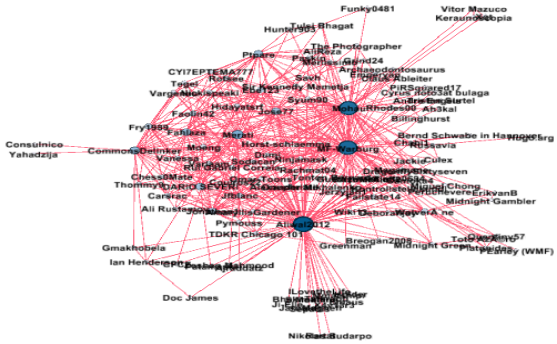

Abbildung A.7.: Nord-Sotho Netzwerk

Abbildung A.8.: Sesotho Netzwerk

Abbildung A.9.: Setswana Netzwerk

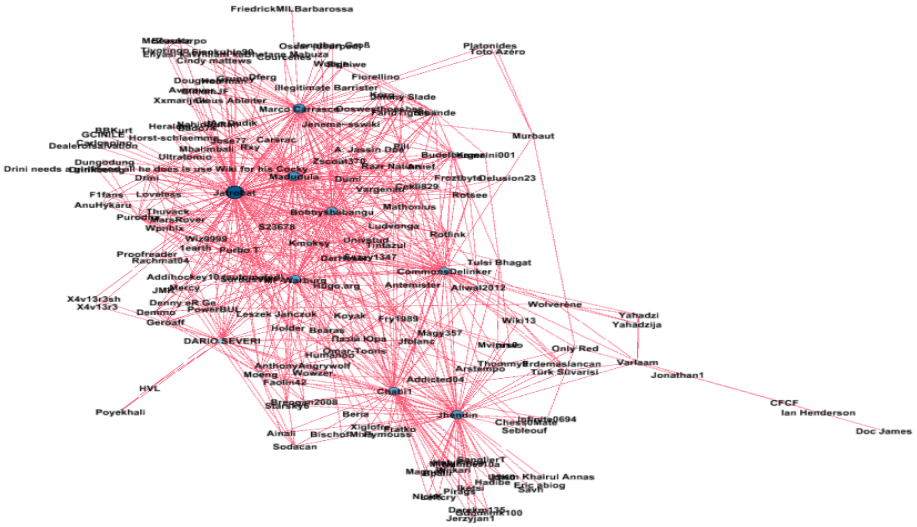

Abbildung A.10.: Siswati Netzwerk

Abbildung A.11.: Tshivenda Netzwerk

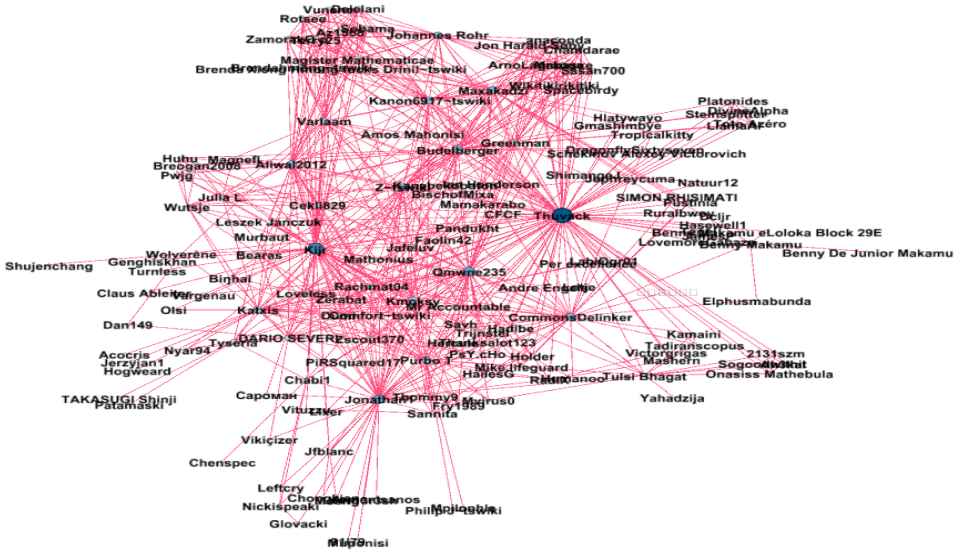

Abbildung A.12.: Xitsonga Netzwerk

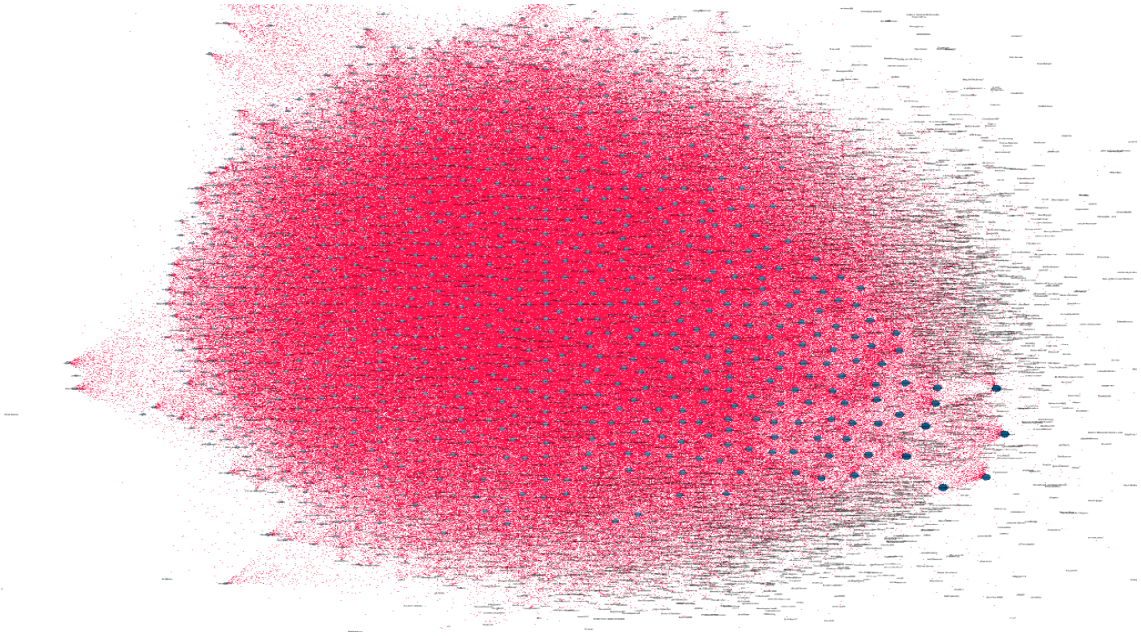

Abbildung A.13.: Englisches Netzwerk

Literaturverzeichnis

[Igra] *igraph Assortativity coefficient.* URL: http://igraph.org/r/doc/assortativity.html (besucht am 12. 01. 2018) (zitiert auf S. 40).

[Igrb] *igraph Shortest (directed or undirected) paths between vertices.* URL: http://igraph.org/r/doc/distances.html (besucht am 12. 01. 2018) (zitiert auf S. 42).

[Igrc] *igraph Transitivity.* URL: http://igraph.org/r/doc/transitivity.html (besucht am 12. 01. 2018) (zitiert auf S. 41).

[Mas11] P. Massa. „Social networks of wikipedia“. In: *Proceedings of the 22nd ACM conference on Hypertext and hypermedia.* ACM. 2011, S. 221–230 (zitiert auf S. 25).

[New02] M. E. Newman. „Assortative mixing in networks“. In: *Physical review letters* 89.20 (2002), S. 208701 (zitiert auf S. 11, 39).

[New03] M. E. Newman. „Mixing patterns in networks“. In: *Physical Review E* 67.2 (2003), S. 026126 (zitiert auf S. 11, 39).

[New10] M. Newman. *Networks: an introduction.* Oxford university press, 2010 (zitiert auf S. 11, 12, 39–41).

[Pag+99] L. Page, S. Brin, R. Motwani, T. Winograd. *The PageRank citation ranking: Bringing order to the web.* Techn. Ber. Stanford InfoLab, 1999 (zitiert auf S. 13).

[Sin] *Single-User Login for all wikis.* 2009. URL: https://blog.wikimedia.org/2015/04/14/single-user-login-for-all-wikis (besucht am 12. 01. 2018) (zitiert auf S. 15).

[TM69] J. Travers, S. Milgram. „An experimental study of the small world problem“. In: *Sociometry* (1969), S. 425–443 (zitiert auf S. 12).

[Wik] *List of Wikimedia projects by size.* URL: https://meta.wikimedia.org/wiki/List_of_Wikimedia_projects_by_size (besucht am 12. 01. 2018) (zitiert auf S. 13).

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift