

Dezentrale Systeme und Netzdienste
Institut für Telematik

Lehrstuhl
Prof. Dr. Hannes Hartenstein

Fakultät für Informatik

Diplomarbeit
2014

Analyse internationaler Nachrichtenflüsse
im Twitter-Netzwerk

Peter Michael Bolch

Mat.Nr.: 1345211

Referent:
Betreuer: Matthias Keller

Ich erkläre hiermit, dass ich die vorliegende Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, 2014

Peter Michael Bolch

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation und Hintergründe	1
1.2	Problembeschreibung	3
1.3	Fragestellungen und Anforderungen	4
1.3.1	Anforderungen	4
1.4	Gliederung der Arbeit	5
2	Grundlagen	7
2.1	Geografische Grundlagen und Begriffe	8
2.1.1	Geografische Koordinaten	8
2.1.2	Geodätisches Referenzsystem	8
2.1.3	Georeferenz	9
2.1.4	geografische Objekte	10
2.1.5	Toponyme	10
2.1.6	Georeferenzierung	10
2.1.7	Ortsverzeichnis	11
2.1.8	Geografische Position	11
2.1.9	Geografische Region	11
2.1.10	Geografische Hierarchie	11
2.2	Twitter	12
2.2.1	Geschichtliches	13
2.2.2	Was ist Twitter?	13
2.2.3	Funktionen von Twitter	15
2.2.4	Daten einer Twitter-Nachricht	18
2.2.5	geografischer Indikator	19
2.2.6	Geoinformationen in Twitter Daten	20

3	Stand der Technik	23
3.1	Kategorisierung bestehender Ansätze	23
3.1.1	ttt<sss	27
3.1.2	Probleme früherer Ansätze	29
4	Lösungsansatz	30
4.1	Überblick	30
4.2	Generelle Struktur einer Georeferenz-Basis zur Georeferenzierung	31
4.3	Geografische Indikatoren	34
4.3.1	Indikatoren mit unmittelbarem geografischen Bezug	34
4.3.2	Indikatoren mit mittelbarem geografischen Bezug	35
4.4	Toponyme als geografische Indikatoren	36
4.4.1	Probleme bei der Verwendung von Toponymen	36
4.4.2	Domänenspezifische Toponyme	39
4.4.3	Fazit	40
4.5	Der Nutzer-Standort	40
4.5.1	Fazit	45
4.6	Die Nutzer-Zeitzone	46
4.6.1	Eigenschaften der Nutzer-Zeitzone	46
4.6.2	Auflösen von Doppeldeutigkeiten	48
4.7	Verfahren zum einlernen geografischer Indikatoren am Beispiel von Twitter 48	
4.7.1	Vorverarbeitung des Nutzer-Standortes und der Nutzer-Zeitzone	49
4.7.2	Absoluten Häufigkeiten	55
4.7.3	Vorverarbeitung der geografischen Koordinaten	56
4.7.4	Erweiterte Struktur der Georeferenz-Basis	58
4.7.5	Überblick	58
4.8	Geografischer Bezug der Referenzwerte	59
4.8.1	Die absolute Häufigkeit als Hinweis auf geografischen Bezug zu Städten	60
4.8.2	Relative Häufigkeiten als Hinweis auf geografischen Bezug zu Städten	61
4.8.3	Geografischer Bezug der Referenzwerte zu Verwaltungseinheiten und Ländern	64
4.9	Verfahren zur Georeferenzierung am Beispiel von Twitter	67
4.9.1	Analyse	67

5	Implementierung	72
5.1	Komponenten der Referenzimplementierung	72
5.1.1	Architektur	72
5.1.2	Präprozessorverarbeitung - Erzeugung der N-Gramme	72
5.2	Datenbank	72
5.3	Geografie Daten	73
5.4	Data Sample	73
5.5	geonames.org	73
6	Leistungsbewertung	74
7	Schlussfolgerungen, Ausblick und Fragen	75
8	Zusammenfassung	76
9	Ideen und Notizen	77
9.1	Stakeholder analyse	77
9.2	Fragen an Matthias	77
9.2.1	Strukturell	77
9.2.2	Inhalt	77
9.3	Ideen	77
9.4	Formulierungen	78
9.4.1	unmittelbare ungesicherte geografische Indikatoren	78
9.5	Datenbasis	79
9.6	Vorteile neuer Ansatz bei Mapping auf Geografische Daten	79
	Literaturverzeichnis	80

Todo list

■ Bild Twitter-Nutzer als sensor	2
■ 1 retweet Reichweite 1000 Nutzer	17
■ Diagramm Retweet, Filterfunktion	17
■ siehe Bild ref1	17
■ siehe Bild ref2	17
■ siehe Bild ref3	17
■ siehe Bild ref4	17
■ Diagramm Antwort, Antwort Thread, Bild Antworten Button, Referenzieren .	17
■ siehe Bild	18
■ Überarbeiten subsec:Geoninformationen in Twitter Daten	20
■ Nur optional	21
■ Irgendwo auf den Umstand eingehen, dass Timezone nicht angegeben werden wird und dann der Standard gewählt wird der us central pacific time ist?	22
■ Umschreiben und woanders darauf eingehen!	22
■ Indikatoren aus [SHP ⁺ 13]	27
■ Tabelle einfügen, bereits fertig, nur noch Format anpassen (Lesbarkeit)	28
■ Requirements Tabelle einfügen	28
■ geografische Entität definieren	29

■ Schema Darstellung Eingabe->Ausgabe Sketchbook A1: Unterschrift Die Eingabe besteht aus dem Nutzer-Standort sowie der Nutzer-Zeitzone. Als Rahmenbedingungen wird die gewünschte Hierarchieebene sowie der Schwellwert für die Konfidenz angegeben. Als Ausgabe erhält man eine Georeferenz	30
■ img:georef, img:ablaufEinlernen, img:Georeferenzierung	31
■ XXX->wenn in Grundlagen soll	34
■ XXX->Annahmen raus, können nicht mehr verwendet werden	36
■ Bieberville beispiel in lernen	40
■ Quelle: http://www.zeitzonen.de/images/frontend/mod_tz_map/zeitzonen_weltkarte.gif	47
■ Bild Auswahlliste	47
■ nearestNeighbourImage with Voronoi-Region and Zuordnung ländlich, ballungszentrum	58
■ Datensätze in Grundlagen?	72
■ Eventuell was über die Geo Indexe in der Datenbank und die Nearest Neighbour Berechnungen.	72
■ in Implementierung verschieben	73
■ In Einleitung	77
■ Korrelation zwischen Lokalisierungssicherheit und tatsächlichem Match berechnen	78

1 Einleitung

1.1 Motivation und Hintergründe

Über den Mikroblogging-Dienst Twitter lassen sich in Echtzeit 140 Zeichen lange Textnachrichten veröffentlichen. Seit dem Start des Mikroblogging-Dienstes im Jahr 2006 sind die Nutzerzahlen kontinuierlich angestiegen. 2010 konnte Twitter 75 Millionen aktive Nutzer verzeichnen [CCL10]. Im Jahr 2013 wird Twitter täglich von zirka 100 Millionen Menschen weltweit aktiv genutzt. Dies berichtete Twitter 2013 in seinem Prospekt zum Börsengang [ti13]. Zur Gesamtanzahl der Nutzer-Konten gibt es von Twitter keine Informationen. Dies kann mitunter damit begründet werden, dass die Gesamtanzahl der Nutzer-Konten auch inaktive Nutzer einschliesst und somit keine Informationen über die tatsächliche Aktivität im Netzwerk liefert. Auch andere soziale Netzwerke ziehen die aktiven Nutzer als Metrik heran, des weiteren wird die Metrik vom Interactive Advertising Bureau (IAB) empfohlen. [IAB]

Die Twitter-Nutzer verfassen täglich mehr als 500 Millionen Nachrichten, sogenannte Tweets [ti13].¹ Die meisten dieser Tweets sind öffentlich zugänglich und können von allen Twitter-Nutzern uneingeschränkt betrachtet werden. Twitter bietet zusätzlich eine sogenannte Streaming-API an, welche es ermöglicht Tweets programmatisch zu empfangen.² Die Streaming-API stellt ein Echtzeit-Sample der aktuell versendeten Tweets bereit und liefert maximal 1% aller Tweets die zum aktuellen Zeitpunkt verfasst wurden [MPLC13]. Über die sogenannte Filter-API lassen sich die Tweets nach bestimmten Kriterien wie Nutzer-ID, geografischer Region oder Schlüsselwörtern filtern.³

¹Im Abschnitt Grundlagen wird der Begriff Tweet genauer untersucht, für den Moment sollen darunter die Nachrichten verstanden werden, welche von den Twitter-Nutzern verfasst werden

²API: Application Programming Interface oder auch Programmierschnittstelle

³<https://dev.twitter.com/docs/streaming-apis>

Ein Tweet besteht aus einer Reihe von Informationen. Neben dem Verfasser, ist der Tweet-Text die wichtigste Information die in einem Tweet enthalten ist. Der Tweet-Text wird vom Nutzer verfasst und abgesendet, er beinhaltet die zentrale Information eines Tweets. In den 140 Zeichen des Tweet-Textes teilen Twitter-Nutzer Informationen unterschiedlicher Ausprägung aus. Unter anderem wird über privates, Sportergebnisse, Großereignisse, persönliche Erfahrungen oder persönliche Meinungen berichtet. Auch Bilder und Web-Links können in einem Tweet-Text enthalten sein.

Mit Hilfe der Streaming-API ist es erstmals möglich, große Mengen nutzergenerierter Informationen unterschiedlichster Ausprägung direkt zu erhalten. Durch die Möglichkeiten die Twitter bietet kann theoretisch jeder Mensch Nachrichten und Informationen über das Twitter-Netzwerk verbreiten und weitergeben. Diese Masse an nutzergenerierten Informationen bietet Wissenschaftlern in verschiedenen Bereichen zahlreiche neue Möglichkeiten.

Sakaki et al interpretieren die Twitter-Kurznachrichten beispielsweise als Sensor-Daten [SOM10]. Der Twitter-Nutzer fungiert dabei als Sensor, der ein beliebiges Ereignis erfährt oder erlebt. Möglicherweise berichtet der Twitter-Nutzer im Tweet-Text über dieses Ereignis. Damit kann der Text als Sensor-Datum interpretiert werden, wenn auch erhebliches Rauschen in der Gesamtheit der Tweets zu erwarten ist. Sakaki et al zeigen aber, dass mit diesem Vorgehen, Erdbebenzentren lokalisiert oder die Trajektorie eines Typhoons vorhergesagt werden können.

Bild Twitter-Nutzer als sensor

Auch die Sozialwissenschaften und die Meinungsforschung profitieren von dem enormen Informationsfundus der durch Twitter geboten wird. Tumasjan et al. untersuchen in [TSSW11] wie sich die politische Landschaft im Twitter-Netzwerk widerspiegelt. Die Wissenschaftler haben zur Bundestagswahl 2009 100.000 Tweets analysiert und stellten fest, dass die Erwähnungen von Parteien und Politikern in Twitter, den Wahlausgang sehr genau widerspiegeln.

Die Kommunikation innerhalb des Twitter-Netzwerks kann aber auch neue Einsichten über die globale Kommunikation oder die Ausbreitung von Nachrichten liefern. Garcia-Gavilanes et al. erforschen in [GGMQ14] die Kommunikation zwischen Ländern. Es wird gezeigt, dass die globale Kommunikation innerhalb des Twitter-Netzwerks nicht nur von der geografischen Distanz abhängig ist, sondern auch von sozialen, ökonomischen und kulturellen Attributen eines Landes.

Selbst die Epidemieforschung kann von den Daten des Twitter-Netzwerks profitieren. So zeigten Szomsor et al. in [SKD11], dass die Vorhersage der Schweingrippe im Jahr 2009 durch die Analyse von Tweets eine Woche früher möglich gewesen wäre als dies mit konventionellen Frühwarnsystemen der Fall war.

Diese Erkenntnisse und Informationen sind allerdings nur gewinnbringend einzusetzen, wenn der Standort des Twitter-Nutzers bekannt ist. Die Information, dass eine Krankheit ausgebrochen ist, ist mit einer exakten Georeferenz wertvoller als ohne diese. Auch die Arbeit von Sakaki et al. ist auf eine Georeferenz angewiesen, wobei die Wissenschaftler angeben, dass die ungefähre Position für ihre Anwendung ausreichend ist. Bei der Untersuchung internationaler Kommunikation wiederum, ist es wichtig zu Wissen in welchem Land ein Tweet verfasst wurde. In diesem Fall kann die Georeferenz eine größere Region umfassen und muss nicht GPS-Genauigkeit aufweisen. Wohingegen eine detaillierte Untersuchung des politischen Klimas innerhalb Deutschlands eine Auflösung auf Bundesländer-Ebene erforderlich machen würde.

Twitter bietet seinen Nutzern die Möglichkeit ihren Standort im Nutzerprofil anzugeben. Hecht et al. stellen in [HHSC11] eine erste ausführliche Analyse der eingegebenen Standort-Daten bereit. Ab 2009 ermöglichte Twitter ein “per-tweet geo-tagging“ [CCL10]. Dadurch können Anwendungen, auf Endgeräten mit GPS, Längen- und Breitengrad des aktuellen Standorts als Georeferenz an den Tweet anhängen. Nur ca. 1,7% der Twitter-Kurznachrichten enthalten allerdings eine konkrete Georeferenz in dieser Form. ⁴

1.2 Problembeschreibung

Um gewinnbringende Informationen aus den Tweets erzeugen zu können, ist es wichtig Tweets eine Georeferenz zuordnen zu können. Die Anzahl der Twitter-Kurznachrichten die mit Hilfe von Längen- und Breitengrad unmittelbar einem geografischen Ort zugeordnet werden können ist sehr gering.

Es ist also wichtig ein Verfahren zu finden um Twitter-Nutzer oder Tweets eine Georeferenz zuzuordnen. Mit Hilfe der in einem Tweet vorhandenen Daten sollte eine möglichst

⁴Prüfung durch Datensatz XYZ was sich mit den Ergebnissen von [PCV13] und [SHP⁺13]

genaue Position bestimmt werden. Dies soll auch möglich sein, wenn keine konkrete geografische Angabe in Form von Längen- und Breitengrad vorliegt.

1.3 Fragestellungen und Anforderungen

Die folgenden Fragestellungen sollen beantwortet:

Q1 Wie kann Twitter-Nutzern eine Georeferenz zugeordnet werden?

1.3.1 Anforderungen

Das erarbeitete verfahren soll folgende Anforderungen erfüllen.

R1 Zuordnung einer Georeferenz zu einem Twitter-Nutzer. (R1)

R2 Unabhängig von kommerziellen Anbietern geografischer Informationen, oder sonstiger benötigter Daten. (R2)

R3 Das Ergebnis ist eine Georeferenz welche einer geografischen Hierarchieebene entspricht. Folgende Hierarchieebenen werden angeboten (R3):

a) Land oder Staat

b) Administrationsebene erster Ordnung ⁵

c) Administrationsebene zweiter Ordnung ⁶

d) Stadt

R4 Es soll möglich sein eine Mindestanforderung für die Konfidenz, mit welcher die Georeferenz bestimmt wurde, anzugeben.

R5 Verfahren unabhängig von Sprache und Schriftzeichen weltweit einsetzbar.

⁵in D Bundesländer, bspsw. Baden-Württemberg, Bayern usw.

⁶in D Regierungsbezirke bspsw. Regierungsbezirk Stuttgart, Regierungsbezirk Karlsruhe usw.

1.4 Gliederung der Arbeit

Abschnitt 2: Grundlagen

In diesem Abschnitt sollen die Grundlagen für die entwickelte Methode vermittelt werden. Es wird auf den Mikroblogging-Dienst Twitter eingegangen und es werden grundsätzliche Methoden und Verfahren vorgestellt welche zum Verständnis der entwickelten Methode benötigt werden. Ebenso werden häufig genutzte geografische Grundbegriffe vermittelt.

Abschnitt 3: Stand der Technik

Es werden aktuelle Ansätze betrachtet, eingeordnet und in Bezug auf die angegebenen Anforderungen untersucht. Es werden sowohl die Verfahren zur 'Analyse' und Zuordnung als auch die Verfahren zum abbilden der geografischen Einheiten untersucht und eingeordnet.

Abschnitt 4: Lösungsansatz

In diesem Kapitel wird, unter Berücksichtigung der gegebenen Anforderungen, ein Verfahren zur Lösung der Fragestellungen entwickelt. Um einen Überblick zu gewährleisten, wird das Verfahren zunächst allgemein betrachtet, danach wird jeder Verfahrensschritt dargelegt. Es wird gezeigt wie aus Tweet-Daten der Standort eines Twitter-Nutzers bestimmen werden kann. Dabei werden Methoden der Sprachverarbeitung, Statistik und geografische Hierarchien eingesetzt.

Bottom-Up:

1. NGramme aus Indikatoren erzeugen
2. Geomapping
3. Datenstruktur
4. Treffer zählen (NGramm + Geoid gleich usw.)

5. Geografische Hierarchieebene
6. Unsicherheit bei Lokalisierung messen (neuer Daten)
7. Justierung der Lokalisierungsunsicherheit auf geografischen Hierarchieebenen

Abschnitt 5: Referenzimplementierung der entwickelten Methode

Es werden ausgewählte Auszüge, Probleme und Fallstricke der Referenzimplementierung erläutert und erklärt.

Abschnitt 6: Leistungsbewertung der entwickelten Methode

In diesem Kapitel werden die Ergebnisse der Referenzimplementierung bewertet und, soweit sinnvoll, gegenüber bestehenden Ansätze einer kritischen Betrachtung unterzogen.

Abschnitt 7: Schlussfolgerungen

Unter besonderer Berücksichtigung der Ergebnisse des letzten Kapitels werden Schlussfolgerungen gezogen. Der Beitrag und nutzen der entwickelten Methode soll kritisch hinterfragt werden.

Abschnitt 8: Zusammenfassung und Ausblick

Zusammenfassung der Arbeit und kritischer Rückblick. Im Ausblick werden mögliche Verbesserungen und Ideen zur Weiterentwicklung gegeben.

2 Grundlagen

In den folgenden Abschnitten werden eine Reihe von Begriffen und Verfahren genutzt, die hier eingeführt werden sollen. Dies ermöglicht dem Leser die Beantwortung der Fragestellungen aus Abschnitt 4 nachzuvollziehen.

Zunächst werden einige geografische Grundbegriffe eingeführt. Danach wird auf Twitter eingegangen, es werden grundsätzliche Funktionen und Begriffe von Twitter eingeführt.

Zum Schluss wird die genutzte Datenbasis und der Einfluss der von Twitter genutzten Sampling-Strategie vorgestellt und erläutert.

2.1 Geografische Grundlagen und Begriffe

In diesem Kapitel sollen geografische Grundbegriffe erläutert werden. Einige geografische Begriffe werden in verschiedenen wissenschaftlichen Bereichen unterschiedlich genutzt und teilweise widersprüchlich definiert. Um Missverständnissen vorzubeugen wird hier definiert was in der vorliegenden Arbeit unter den einzelnen Begriffen zu verstehen ist. Eine Reihe von Begriffen wird selbst definiert um bestimmte Sachverhalte im Kontext dieser Arbeit klarer ausdrücken zu können.

2.1.1 Geografische Koordinaten

Geografische Koordinaten bestehen aus zwei Werten, einem Wert für den Längengrad und einem für den Breitengrad. Mit diesen zwei Werten kann eine Position auf dem Globus exakt bestimmt werden.

Die Längen- und Breitengrade beschreiben ein imaginäres Netz auf dem Globus. Dabei ziehen sich die Breitengrade wie ein Gürtel um den Globus. Der Breitengrad mit dem Wert 0, also 0 Grad Breite, verläuft entlang des Äquators. Die Längengrade hingegen verlaufen vom Nord- zum Südpol. Der Längengrad mit dem Wert 0 wurde 1884 durch ein Konsortium festgelegt, da vertikal des Globus keine natürliche Marke wie der Äquator, verläuft. Die Werte liegen im IT-Umfeld meist als Fließkommazahlen vor und beschreiben jeweils einen Winkel.

2.1.2 Geodätisches Referenzsystem

Ein geodätisches Referenzsystem dient als einheitliche Grundlage zur Angabe einer Position auf dem Globus mit Hilfe geografischer Koordinaten. Dazu wird ein kartesisches Rechtssystem mit definierter Lage und Ausrichtung festgelegt. Die Lage und Ausrichtung erfolgt relativ zur Erde. Der Ursprung des Koordinatensystems liegt im Zentrum des Globus, meist im Masseschwerpunkt der Erde. Die Z-Achse zeigt meist in Richtung Nordpol und die X-Achse in Richtung 0 Grad Länge und 0 Grad Breite. Mit diesen zwei Werten ist die Lage eines kartesischen Rechtssystem eindeutig definiert. In diesem Koordinatensystem sind Referenzpunkte festgelegt. Diese Referenzpunkte werden benötigt um einen Referenzellipsoid zu verankern. Der Referenzellipsoid soll eine möglichst

genaue Approximation der Erde darstellen und diese im geodätischen Referenzsystem repräsentieren. Ein Punkt auf diesem Ellipsoid entspricht damit einem Punkt auf der Erde.

Mit diesen Komponenten kann nun ein Punkt auf dem Ellipsoid eindeutig bestimmt werden. Der Längen und Breitengrad eines Punktes auf dem Ellipsoid lässt sich folgendermaßen bestimmen: Durch den Punkt auf dem Ellipsoid und dem Ursprung des Koordinatensystems kann eine Gerade gezogen werden. Der Breitengrad ist nun der Winkel zwischen dieser Linie und der Äquatorebene. Der Längengrad ist der Winkel zwischen der X-Achse, also dem Null Meridian und demjenigen Meridian welcher durch den Punkt geht. Durch eine Projektion der Punkte auf das Ellipsoid können diese Punkte beispielsweise auf einer Karte dargestellt werden.

Heutzutage ist das Referenzsystem WGS84 weit verbreitet.

1

2.1.3 Georeferenz

Eine Georeferenz (engl. Spatial Reference) wird auch als Raumbezug bezeichnet. Ist einem Datensatz eine Lage beziehungsweise Position zugeordnet so wird diese als Georeferenz bezeichnet. Die Art wie die Georeferenz angegeben wird und deren Genauigkeit hängen von den Anforderungen ab, die an die Georeferenz gestellt werden. Beispielsweise stellen die in Kapitel 1 erwähnten Anwendungen unterschiedliche Anforderungen an die Genauigkeit der Georeferenz.

Die Georeferenz lässt sich weiter unterteilen in: ¹

Direkte Georeferenz (direkter Raumbezug) Unter direktem Raumbezug versteht man die Angabe einer konkreten Koordinate bezüglich eines geeigneten geodätischen Referenzsystems.

Indirekte Georeferenz (indirekter Raumbezug) Unter indirektem Raumbezug werden alle Angaben verstanden die eine ungenaue Position bezüglich eines beliebigen

¹Vergleiche Geoinformatik Lexikon der Universität Rostock: <http://www.geoinformatik.uni-rostock.de/lexikon.asp> und Vorlesungen zur Geo-Informatik von Prof. Dr.-Ing. Ralf Bill : <http://www.geoinformatik.uni-rostock.de/vorlesungsthem.asp>

Referenzsystems bestimmen. Ungenau ist in dem Sinne zu verstehen, dass die Angabe der Position auch eine Fläche beschreiben kann. Zusätzlich muss das gewählte Referenzsystem nicht zwingenderweise unveränderlich sein. Beispiele für die Angabe eines indirekten Raumbezugs wären Länder, Adressen, Postleitzahlen oder auch Telefonvorwahlen. Alle diese Angaben, mit Ausnahme der Adresse, definieren eine geografische Fläche. Diese Fläche ist nicht zwingenderweise klar abzugrenzen.

2.1.4 geografische Objekte

Ein geografisches Objekt, ist ein Objekt der Realwelt, dessen Position durch eine Georeferenz bestimmt werden kann. Die EN ISO 19110 Norm beschreibt ein geografisches Objekt folgendermaßen: “Geographische Objekte sind Erscheinungen der realen Welt, die einen Bezug zur Erde (Raumbezug) haben...” Es wird insbesondere nicht festgelegt ob es sich dabei um einen direkten oder einen indirekten Raumbezug handeln muss. Beispiele für geografische Objekte sind: Städte, Länder, Häuser oder auch Fahrzeuge.

2.1.5 Toponyme

Ein Toponym ist ein Name für ein geografisches Objekts. Beispiele hierfür sind: Städtenamen, Ländernamen oder Landschaftsnamen. Ein Toponym muss nicht eindeutig sein. Zu einem geografischen Objekt können mehrere unterschiedliche Toponyme existieren. Aus einem Toponym kann, eine Georeferenz abgeleitet werden.

2.1.6 Georeferenzierung

Unter Georeferenzierung versteht man die Zuordnung einer Georeferenz zu einem Datensatz. Also den Vorgang einem Datensatz, zum Beispiel einem Twitter-Nutzer eine Georeferenz zuzuordnen.¹ In diesem Sinne stellen die Datensätze geografische Objekte dar, da diese einen Bezug zur Erde haben können.

2.1.7 Ortsverzeichnis

Sehr umfangreich mal schauen. Ein Ortsverzeichnis beinhaltet eine Liste von Werten, meist Zeichenketten, welchen eine

2.1.8 Geografische Position

Unter einer geografischen Position soll hier eine Position auf dem Globus verstanden werden deren Wert durch geografische Koordinaten angegeben wird.

2.1.9 Geografische Region

Unter einer geografischen Region werden hier Flächen auf dem Globus verstanden. Diese können nicht durch einen einzelnen Punkt beschrieben werden. Flächen werden üblicherweise durch eine Menge von Punkten beschrieben. Die Punkte liegen als geografische Koordinaten vor und werden durch Geraden verbunden. Die so eingeschlossene Fläche innerhalb des Polygons beschreibt dann die geografische Region. Bundesländer oder Länder sind Beispiele für geografische Regionen. Somit entspricht der Begriff geografische Region einer indirekten Georeferenz (indirektem Raumbezug).

2.1.10 Geografische Hierarchie

In der vorliegenden Arbeit wird eine geografische Hierarchie verwendet um eine Einteilung der Erde in geografische Regionen umzusetzen. Dabei können geografische Regionen wiederum geografische Regionen oder geografische Positionen enthalten, wodurch sich eine hierarchische Gliederung ergibt. Diese Einteilung spiegelt im wesentlichen die Einteilung der Erde in Staaten und deren individuellen Verwaltungseinheiten wieder. Im vorliegenden Fall ist das Staatsgebiet, also die Fläche über die sich der Staat erstreckt, von Interesse.²

²Zur genauen Definition eines Staatsgebietes vergleiche [JJ21]

Im Gegensatz dazu könnte die Erde auch in ein Gitternetz eingeteilt werden. Die einzelnen Zellen würden dann als Referenz für eine geografische Region verwendet werden. Dieses vorgehen wird unter anderem in [SMvZ09] angewendet.

Eine Aufteilung der Erde in geografische Regionen lässt sich auf oberster Ebene mit Hilfe von Ländern und deren Grenzen umsetzen. Daraus resultiert eine intuitiv verständliche Einteilung. Die meisten Länder sind in weitere administrative Einheiten aufgeteilt. Diese geografischen Regionen werden hier als Verwaltungseinheiten bezeichnet. Es wird zwischen Verwaltungseinheiten erster und zweiter Ordnung unterschieden. Der Vatikan-Staat und das Fürstentum Monaco sind dabei Ausnahmen. Der Vatikan-Staat oder das Fürstentum Monaco werden aufgrund ihrer Größe nicht in weitere Verwaltungseinheiten unterteilt. In der untersten Ebene der Hierarchie werden schliesslich Städte dargestellt.

Wenn man als Beispiel Deutschland heranzieht, ergibt sich eine Einteilung wie in Bild 2.1 dargestellt wird.³ Die oberste Ebene beschreibt das Land worauf die zweite Ebene die Bundesländer darstellt. Auf der dritten Ebene werden die Regierungsbezirke abgebildet, worauf die Städte in der letzten Ebene folgen. Analog kann die Einteilung für die USA vorgenommen werden, woraus sich die Hierarchie Country->State->County->City ergibt.

Bis auf die letzte Ebene wird den Objekten in der Hierarchie eine geografische Region zugeordnet. Lediglich die unterste Ebene, die der Städte, wird durch eine geografische Position exakt beschrieben. Die Ausdehnung einer Stadt wird in der gegebenen Hierarchie also nicht berücksichtigt. Jede Stadt wird als konkrete geografische Position mit Koordinaten repräsentiert.

Ortsverzeichnis Oder auch Gazetteer genannt. Auch auf Geocoding API eingehen.

2.2 Twitter

In diesem Kapitel werden grundlegende Begriffe rund um das Twitter-Netzwerk erläutert. Weiter werden die Mechanismen in Twitter erläutert und an praktischen Beispielen

³Aus Platzgründen sind im Bild pro Ebene nur einige wenige geografische Objekte aufgezählt.



Abbildung 2.1: Die Hierarchieebenen exemplarisch. Bis in die Städteebene wird nur der Pfad Welt -> Deutschland -> Baden-Württemberg -> Karlsruhe -> Karlsruhe, Pforzheim, Baden-Baden dargestellt.

erklärt. Zum Schluss wird aufgezeigt welche Informationen pro Tweet übermittelt werden und welche Daten zur Lokalisierung verwendet werden können.

2.2.1 Geschichtliches

Twitter wurde 2006 von Jack Dorsey, Biz Stone, Noah Glass und Evan Williams gegründet. Ursprünglich war Twitter zur internen Kommunikation innerhalb der Firma Odeo geplant. Schnell wurde allerdings klar, dass in dem Dienst mehr Potenzial steckt und so wurde Twitter öffentlich gemacht. Seitdem erfreut sich der Dienst einer wachsenden Nutzer-Gemeinde. Die Twitter-Gründer haben von Anfang an keine exakten Nutzer-Zahlen oder die Anzahl der versendeten Twitter-Kurznachrichten bekanntgegeben. Dies geschah einerseits, weil die Gründer davon überzeugt sind, dass anhand der reinen Nutzer-Zahlen und gesendeten Twitter-Kurznachrichten nicht die "Gesundheit" des Twitter-Netzwerks nachvollzogen werden kann, andererseits werden durch diese Massnahme auch strategische Ziele verfolgt.⁴ 2013 ging Twitter an die Börse und vermeldete 100 Millionen täglich aktive Nutzer und über 500 Millionen Twitter-Kurznachrichten, die täglich über den Dienst versendet werden.

2.2.2 Was ist Twitter?

Twitter wird als Kurznachrichten-Dienst, Mikroblogging-Dienst oder auch als soziales Netzwerk bezeichnet. Twitter Geschäftsführer Kevin Thau hat 2010 auf dem Nokia-

⁴<http://www.pbs.org/mediashift/2007/05/twitter-founders-thrive-on-micro-blogging-constraints137>

World Kongress öffentlich bestritten, dass Twitter ein Soziales-Netzwerk ist. Laut Thau handelt es sich um ein Nachrichten-, Inhalts- und Informations-Netzwerk. Er begründete dies damit, dass Twitter die Art und Weise wie Nachrichten verteilt werden geändert hat und praktisch jeder zum Journalisten werden kann. Als Beispiel nennt er die Landung des Fluges 1549 auf dem Hudson River. Die Augenzeugen hätten damals keine Mails versendet um die Nachricht zu verbreiten, sondern die Nachricht via Twitter weitergegeben. Es lassen sich eine Reihe weitere Beispiele derselben Art finden. In [POM⁺13] wird ein Vergleich zwischen sogenannten Newswire Anbietern und Twitter gezogen.⁵ Es stellte sich heraus, dass über nahezu alle Nachrichten, welche in den Newswires verbreitet wurden auch im Twitter-Netzwerk berichtet wird. Nachrichten zu bestimmten, vermutlich sehr speziellen Themen oder Auslandsnachrichten wurden ausschliesslich in Twitter gefunden. Diese Erkenntnisse decken sich mit der Einschätzung von Kevin Thau. In [KLPM10] wird die Einschätzung, bei Twitter handele es sich nicht um ein soziales Netzwerk, wissenschaftlich bestätigt. Kwak et al überprüfen die in [NP03] beschriebenen Eigenschaften sozialer Netzwerke und kommen zu dem Schluss, dass Twitter diese Eigenschaften nicht erfüllt.

Die Bezeichnung Kurznachrichten-Dienst ist irreführend, da dieser mit sms (small messenger service) in Verbindung gebracht werden kann. Tatsächlich galt der sms in der Anfangsphase von Twitter als Vorbild für den Dienst. In Twitter werden Nachrichten allerdings standardmäßig allen Benutzern zur Verfügung gestellt und können eingesehen werden. Des weiteren wird eine Liste der Nachrichten, welche von einem Nutzer verfasst wurden, als Liste in umgekehrter chronologischer Reihenfolge auf dessen Profil dargestellt. Damit ähnelt das Twitter-Profil einem Blog mit Einträgen deren Länge 140 Zeichen nicht überschreiten darf. Die Darstellung als Liste, und die Funktion einen Tweet standardmäßig allen Nutzern freizugeben unterscheidet sich grundlegend von der Funktion des sms, bei dem eine Nachricht direkt an einen Empfänger gesendet wird und nicht öffentlich ist. Im sms steht die Konversation zweier Nutzer im Vordergrund, wohingegen Nachrichten im Twitter-Netzwerk einen Broadcast an alle Nutzer darstellen.

Die 140 Zeichen langen Nachrichten in Twitter werden als Tweets bezeichnet. Tweet bedeutet übersetzt Zwitschern, womit die Redenwendung "Die Spatzen zwitschern es

⁵Newswire stellt eine Art Nachrichtenaggregator dar, über welchen Nachrichten aus verschiedenen Quellen aggregiert und weitergegeben werden. In Deutschland kommt die Deutsche Presseagentur diesem Konzept am nächsten.

von den Dächern“ auch im Twitter-Netzwerk zu einer passenden Redenwendung wird. In der vorliegenden Arbeit wird Twitter deshalb als Mikroblogging-Dienst bezeichnet.

2.2.3 Funktionen von Twitter

Der Mikroblogging-Dienst Twitter bietet neben dem Profil, auf dem die Tweets des Nutzers angezeigt werden, noch eine Reihe weiterer Funktionen. Im folgenden soll das Twitter-Profil und die Timeline kurz erläutert werden. Eine der zentralen Funktionen von Twitter ist das sogenannte Folgen, womit sich Nutzer ein Netzwerk aufbauen können aus dem sie Twitter Nachrichten erhalten. Danach werden Funktionen wie das Weitergeben eines Tweets, Favorisieren und Antworten erklärt. Zum Schluss wird auf den gesendeten Tweet Inhalt eingegangen und der Netzwerk-Charakter von Twitter untersucht.



Abbildung 2.2: Die Twitter-Timeline auf einem Twitter Profil. 1: Nutzernamen und Informationen über den Nutzer. 2: Profilbild 3: Allgemeine Informationen über den Benutzer und dessen Netzwerk 4: Nutzer-Timeline: Tweets des Nutzer in umgekehrter chronologischer Reihenfolge 5: Button zum Folgen

Das Nutzer-Profil und die Nutzer-Timeline Das Nutzer-Profil kann über die Url <http://twitter.com/BENUTZERNAME> abgerufen werden und bietet neben der Nutzer-Timeline, in der die Tweets des Nutzers angezeigt werden, eine Reihe an weiteren Informationen. In Abbildung 2.2 ist in der mitte die Timeline des Benutzers dargestellt in der dei Tweets zu sehen sind. Unter dem Profilbild links sind Informationen des Nutzers aufgelistet. Diese Informationen kann der Nutzer selbst einstellen und entscheiden welche er angeben möchte.

Folgen (Following/Follower/Tweeps) Diese Funktion erlaubt es Tweets eines bestimmten Nutzers zu abonnieren. Im Twitter-Umfeld spricht man von "following" oder "folgen", wenn man die Tweets eines bestimmten Nutzers abonniert. Hat man Tweets eines bestimmten Nutzers abonniert so wird man als dessen "Follower" bezeichnet. Das englische Wort "Follower" hat sich im Twitter-Umfeld und darüber hinaus eingebürgert und wird selten übersetzt. Auch auf der Twitter Website wird "Follower" nicht ins deutsche übersetzt. In der vorliegenden Arbeit wird deshalb auch auf eine Übersetzung verzichtet.

In Abbildung 2.2 an Position 3 wird unter "Folge ich" die Anzahl der Twitter-Nutzer angezeigt denen der Beispielnutzer folgt. Neben dem Feld "Folge ich" wird unter "Follower" angezeigt wieviele Nutzer dem Beispielnutzer folgen.

Persönliche Timeline Jeder Twitter-Nutzer hat seine persönliche Timeline, auf dieser werden die Tweets derjenigen Nutzer angezeigt, denen er folgt. Die Timeline kann als Aggregation von Tweets betrachtet werden. Diese Timeline ist die zentrale Stelle, an der die Nutzer Tweets anderer Nutzer empfangen und lesen. Auch hier werden die Tweets in umgekehrter chronologischer Reihenfolge angezeigt.

Weiterleiten eines Tweets (Retweet) Unter einem Retweet versteht man das weiterleiten eines Tweets den man nicht selbst verfasst hat an die eigenen Follower. Genauer gesagt wird der Tweet übernommen und ein Hinweis hinzugefügt, dass es sich um einen sogenannten Retweet handelt, und nicht einen vom Nutzer selbst verfassten Tweet. Diese Funktion wird hauptsächlich genutzt um Nachrichten schnell zu verbreiten ohne diese

neu eingeben zu müssen. Die Weitergabe an die eigenen Follower impliziert einen gewissen Grad an Kontrolle und Filterfunktion. Der weitergebende Nutzer kontrolliert und filtert die Nachrichten die er erhält und gibt diejenigen weiter, denen er eine Gewisse Relevanz beimisst, oder von denen er erwartet, dass sie seine Follower interessieren. Mit dieser Funktion können einzelne Nutzer eine Art Filterfunktion übernehmen, welche früher Journalisten vorbehalten war. Es darf jedoch nicht vergessen werden, dass der Nutzer nur im Rahmen seiner eigenen Möglichkeiten einen Tweet verifizieren kann und Nachrichten in Twitter keinesfalls gesicherte Fakten darstellen. Auch können Nutzer durch diese Funktion zu Tweet-Aggregatoren werden, welche Tweets von mehreren Nutzern erhalten oder sammeln, aber nur relevante oder themenspezifische Tweets weitergeben.

1 retweet
Reichweite
1000 Nutzer

Diagramm
Retweet, Filterfunktion

Hashtags Hashtags werden genutzt um Tweet Nachrichten zu kategorisieren oder Metatag Informationen zu liefern. Ein Hashtag kann vom Verfasser selbst als solches ausgezeichnet werden indem ein # vor das gewünschte Wort, welches als Hashtag fungieren soll, gesetzt wird. Hashtags ermöglichen es Tweets nach Stichworten zu filtern. Anhand der Hashtags werden auch die Twitter-Trends analysiert. Twitter Trends

Antworten und direktes ansprechen eines Nutzers Twitter bietet die Möglichkeit einzelne Nutzer direkt anzusprechen. Mit Hilfe des @-Symbols kann ein Nutzer referenziert werden. Der referenzierte Nutzer, beispielsweise @alfred, wird dann benachrichtigt, dass er in einem Tweet erwähnt wurde. Der erwähnte Nutzer muss dabei nicht Follower des Verfassers sein. Eine weitere Funktion im Twitter-Netzwerk ist das Antworten auf einen Tweet. Über eine Schaltfläche wird es ermöglicht auf einen Tweet zu Antworten. Das @-Symbol und der Nutzernamen des Verfassers werden automatisch eingetragen, womit eine Benachrichtigung an den Verfasser des Ursprungstweets erfolgt. Es ist möglich, dass auf einen Antwort-Tweet wiederum geantwortet wird, wodurch ein sogenannter Thread oder Konversation entsteht. Auch ist es möglich, dass an einer solchen Konversation mehrere Twitter-Nutzer beteiligt sind. Dies ist dann der Fall, wenn im ursprünglichen Tweet, auf weitere User referenziert wurde. Aber auch wenn ein Nutzer auf eine bestehende Konversation antwortet, werden alle beteiligten Nutzer referenziert.

siehe Bild ref1

siehe Bild ref2

siehe Bild ref3

siehe Bild ref4

Diagramm
Antwort, Antwort Thread, Bild Antworten Button, Referenzieren

Favorisieren Mit dieser Funktion lässt sich ausdrücken, dass man einen Tweet interessant oder gut findet. Auch Zustimmung wird durch favorisieren ausgedrückt. Einen Tweet zu favorisieren kann aber auch bedeuten "ich habe deine Reaktion registriert", oft um einen Antwort-Thread nicht abrupt abubrechen sondern eine zustimmende Rückmeldung zu geben ohne extra einen Tweet zu verfassen.

2.2.4 Daten einer Twitter-Nachricht

Neben den direkt sichtbaren Informationen enthält ein Tweet eine Reihe weiterer Daten. Betrachtet man einen einzelnen Tweet, beispielsweise auf twitter.com, wird der Tweet-Text, der Verfasser und die Zeit, wann der Tweet verfasst wurde, mitgeteilt. Die Gesamtheit der Daten die in einem Tweet enthalten sind werden hier allgemein als Tweet-Daten bezeichnet.

siehe Bild



Abbildung 2.3: Was ist zu sehen?

Koordinaten In den Tweet-Daten können geografische Koordinaten in Form von Längen- und Breitengrad angegeben sein. Diese Koordinaten zeigen an wo sich der Verfasser befand als er den Tweet abgesetzt hat. Wenn diese Koordinaten angegeben sind hat der Nutzer explizit zugestimmt, dass die Koordinaten seines aktuellen Aufenthaltsortes dem Tweet angehängt werden. Die Bestimmung der Koordinaten und das anhängen der Koordinaten an einen Tweet werden vollautomatisch durch das Programm übernommen mit welchem der Tweet verfasst wurde. Auf Smartphones wird meist das integrierte GPS-Modul genutzt um die Koordinaten zu bestimmen. Bei der Nutzung an einem PC wird der Tweet häufig über den Browser verfasst und die Position mit Hilfe von GeoIp ermittelt.

Daten Neben den sichtbaren Daten, welche in der Timeline angezeigt werden, enthält ein Tweet eine Reihe weiterer interessanter Informationen.

2.2.5 geografischer Indikator

Unter einem geografischen Indikator wird eine Angabe verstanden, welche direkt einem Nutzer zugeordnet werden kann und die Auskunft über die geografische Position oder Region des Nutzers geben kann. Im Zuge dieser Arbeit wurden potentielle geografische Indikatoren untersucht und eine Reihe von Eigenschaften identifiziert anhand derer sich geografische Indikatoren kategorisieren lassen. Diese Eigenschaften haben Einfluss darauf, wie und ob eine Georeferenz aus dem Indikator abgeleitet werden kann. Dabei ist zu unterscheiden ob sich die Eigenschaft auf den, durch den Nutzer eingegebenen, Wert bezieht oder auf die Information die durch die Angabe geliefert werden soll.

Objektivität der Werte geografischer Indikatoren

Der Wert eines geografischen Indikators ist genau dann objektiv wenn zwei Nutzer für denselben geografischen Ort oder dieselbe geografische Region immer den selben Wert eingeben. Ein Beispiel für einen objektiven geografischen Indikator wäre eine Liste von Ländern aus der ein Nutzer wählen kann. Der Nutzer hat dabei eine Wahl, kann aber nur aus einem begrenzten Anzahl an Möglichkeiten wählen.

Geben zwei Nutzer unterschiedliche Werte ein, obwohl sie denselben Ort oder dieselbe Region beschreiben wollen, ist dieser Wert nicht objektiv.

Zuverlässigkeit der Werte geografischer Indikatoren

Ein Wert ist genau dann zuverlässig wenn er in jedem Fall die Information enthält, welche durch das Feld repräsentiert werden soll. Unzuverlässig ist der Wert, wenn er nicht in jedem Fall die Information enthält welche durch das Feld repräsentiert werden soll.

Gesicherte Werte geografischer Indikatoren

Als gesichert gilt ein Wert genau dann wenn die enthaltene Information in jedem Fall dem tatsächlichen Wert entspricht. Im Umfeld von Twitter ist diese Eigenschaft nicht stichhaltig nachzuprüfen. Alle Angaben die ein Benutzer eingibt werden nicht verifiziert und können dementsprechen auch nicht gesichert sein. Es besteht die Möglichkeit das ein Nutzer in jedem Feld Falschangaben macht. Dies gilt es im erarbeiteten Verfahren zu beachten.

unmittelbare geografische Indikatoren in Tweet-Daten

Unmittelbare geografische Indikatoren sind solche aus denen direkt eine geografische Position abgeleitet werden kann.

bzw. die Intention des Nutzers bei der Eingabe darauf abzielt eine geografische Position zu beschreiben. Mit einer gewissen Sicherheit kann ein geografischer Bezug abgeleitet werden, da die Intention des Feldes einen Standort angibt.

2.2.6 Geoinformationen in Twitter Daten

Welche Tweet-Daten können zur Georeferenzierung herangezogen werden

Nur optional

Um diese Frage zu beantworten, müssen die Tweet-Daten eingehend untersucht werden. Dabei spielt nicht nur die reine Information die den Daten entnommen werden kann eine Rolle, sondern auch wie die Daten generiert oder eingegeben wurden. Beispielsweise kann bei einem Tweet, dem ein Längen- und Breitengrad mit einer Genauigkeit von 14 Nachkommastellen zugeordnet ist, davon ausgegangen werden, dass die geografische Position der tatsächlichen geografischen Position, von welcher der Tweet abgesetzt wurde, entspricht. Es liegt hier die Vermutung nahe, dass diese Werte durch ein mobiles GPS ⁶ erfasst worden sind. Anders verhält sich dies beispielsweise beim Tweet-Text, eine Erwähnung der Stadt New York, muss nicht bedeuten, dass der Tweet aus dieser Stadt stammt. Es impliziert nicht einmal, dass der Verfasser jemals in dieser Stadt war. Im folgenden werden einige Datenfelder, welche mit jedem Tweet versandt werden, untersucht. Dabei wird die Eignung dieser Daten als geografischer Indikatoren bewertet. Währenddessen werden anhand geeigneter Beispiele die Begriffe gesicherter -, unsicherer -, mittelbarer - und unmittelbarer geografischer Indikator eingeführt.

mögliche geografische Indikatoren

Nutzer-Standort Der Nutzer-Standort ist ein unmittelbarer geografischer Indikator. Als Nutzer-Standort kann der Twitter-Nutzer eine beliebige Zeichenfolge eingeben. Es handelt sich beim Nutzer-Standort deshalb um einen unsicheren geografischen Indikator, es ist deshalb damit zu rechnen, dass unter Umständen keine geografische Position angegeben ist und andererseits keine einheitliche Angabe bezüglich des selben Standorts erwartet werden kann. Beispielsweise beschreiben die Zeichenketten “Karlsruhe, Deutschland” und “Baden-Württemberg, Karlsruhe” den selben Ort. Noch deutlicher wird dieser Umstand, wenn man alternative Namen oder umgangssprachliche Namen für Städte betrachtet. Mit “The Big Apple” und “New York, USA” oder mit “Motown” und “Detroit, MI” sind dieselben Orte gemeint. Auch die Genauigkeit bezüglich der geografischen Position ist nicht zuverlässig vorhersagbar, sehr konkrete geografische Positionen, wie die

⁶Global Positioning System

Angabe einer Stadt oder eines Stadtteils, oder aber eine geografische Region wie beispielsweise ein Land oder ein Kontinent, sind möglich.

Nutzer-Zeitzone Die Nutzer-Zeitzone stellt dagegen einen gesicherten, unmittelbaren geografischen Indikator dar. Bei der Nutzer-Zeitzone kann aus einer Liste möglicher Werte gewählt werden, womit keine Ungenauigkeiten bezüglich der Eingabe besteht und eine definierte Zeichenkette erwartet werden kann, deren geografische Region klar definiert ist. Die Nutzer-Zeitzone beschreibt allerdings in jedem Fall eine größere geografische Region, die nicht immer mit den konventionellen Ländergrenzen korrespondiert und somit eine Bestimmung der geografischen Position nahezu unmöglich macht.

Bei beiden Indikatoren besteht natürlich die Möglichkeit der Falscheingabe durch den Benutzer. Dieser Umstand wird jedoch durch die Analyse der Daten ausgemerzt.

Irgendwo auf den Umstand eingehen, dass Timezone nicht angegeben werden wird und dann der Standard gewählt wird der us central pacific time ist?

Umschreiben und woanders darauf eingehen!

3 Stand der Technik

Die Georeferenzierung von Tweets oder Twitter-Nutzern ist ein Feld an dem nach wie vor aktiv geforscht wird. Nicht zuletzt trägt auch die große Verfügbarkeit an Twitter-Daten zu dem Umstand bei, dass Twitter in den letzten Jahren Forschungsgegenstand zahlreicher Publikationen war.

In diesem Abschnitt sollen bestehende Ansätze zur Georeferenzierung im Twitter-Umfeld untersucht werden. Es werden Kriterien zur Einordnung der bestehenden Ansätze erarbeitet und erläutert. Die Arbeiten werden mit Hilfe der Kriterien schematisch eingeordnet um einen Überblick zu erhalten. Zum Schluss wird untersucht ob die Arbeiten die bereits formulierten Anforderungen aus 1.3.1 erfüllen, und wie sich die vorliegende Arbeit von den bestehenden Ansätzen abgrenzt.

3.1 Kategorisierung bestehender Ansätze

In früheren Arbeiten wurde bereits versucht, eine Einordnung der bestehenden Verfahren vorzunehmen. Es ist interessant die Kategorisierungsansätze und die verwandten Arbeiten einiger Autoren zu studieren. Es lässt sich dadurch die Entwicklung zum Thema Lokalisierung im Twitter-Umfeld beobachten. Einige Kategorisierungsansätze werden im folgenden aufgelistet und erläutert.

Sowohl in [HHSC11] als in [CCL10] beschränken sich die verwandten Arbeiten nicht auf die Lokalisierung im Twitter-Umfeld, es werden Arbeiten zur Lokalisierung von Web-Inhalten im Allgemeinen aufgelistet. Dies lässt darauf schliessen, dass sich vor den Jahren 2010/2011 nur wenige Arbeiten mit der Lokalisierung im Twitter-Umfeld beschäftigt haben.

Kategorisierung über die untersuchte Ressource

[HHSC11] nimmt deshalb eine Kategorisierung anhand der untersuchten Ressource vor. Es wird unterschieden zwischen Forschungen zur “Lokalisierung von Microblogging-Seiten und deren Inhalten“ und der “Lokalisierung von Nutzern, welche Inhalte zu Web 2.0 Seiten beisteuern“. Zusätzlich wird in dieser Arbeit das “Verhalten der Nutzer im Umgang mit der Veröffentlichung ihres aktuellen Standorts“ und die “Vorhersage privater Informationen“ betrachtet. Darauf soll hier allerdings nicht weiter eingegangen werden.

Kategorisierung über die verwendete Methode

[CCL10] klassifiziert die vorgestellten Arbeiten anhand der verwendeten Methodik. Es wird auf Arbeiten zur Lokalisierung von Webseiten, Web-Logs, Suchanfragen und Web-Nutzern verwiesen. Diese werden in die folgenden drei Kategorien eingeteilt.

“Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis (Content analysis with terms in a gazetteer)” Es wird darunter eine einfache Datenbanksuche verstanden. Es werden einzelne Wörter in einer Datenbank nachgeschlagen um diese einem konkreten geografischen Ort zuweisen zu können. Dabei kann sowohl lokal auf eine Geo-Datenbank als auch auf Internet Ressourcen zurückgegriffen werden. In der Regel durchläuft der untersuchte Text eine manuelle oder automatische Vorverarbeitung um potenziell geografische Begriffe, sogenannte Toponyme, herauszufiltern.

“Inhaltsanalyse mit probabilistischen Sprachmodellen (Content analysis with probabilistic language models)” Dabei werden Texte oder Textteile einer Twitter-Kurznachricht zu vordefinierten geografischen Regionen wie Ländern oder Städten zugeordnet. Nach einer Vorverarbeitung des Textes erfolgt eine statistische Auswertung, um danach den Text oder einzelne Textteile, wie beispielsweise Wörter, einer geografischen Region zuzuordnen. Eine unbekannter Text kann dann mit Hilfe der zuvor gelernten Zuordnung einer geografischen Region zugeordnet werden.

“Schlussfolgerungen durch soziale Verbindungen (Inference via social relations)”

es werden soziale Verbindungen, die in Netzwerken abgebildet sind, herangezogen um Rückschlüsse auf den geografischen Ort des untersuchten Inhaltes oder einer Person ziehen zu können.

Preidhorsky et al. schlagen in [PCV13] eine weitere Einteilung anhand der Methodik vor. Allerdings werden hier ausschließlich Arbeiten im Twitter-Umfeld betrachtet.

“Geocoding” Im wesentlichen entspricht dies der “Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis” aus [CCL10]. “Geocoding” wird als Begriff in vielen Fachrichtungen unterschiedlich definiert, was zu Missverständnissen führen kann. In [Gol08] wird genauer auf den Begriff des Geocoding und die Problematik eingegangen und eine Definition des Begriffs vorgeschlagen. Im vorliegenden Kontext ist es präziser und weniger missverständlich die Methodik als “Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis” zu bezeichnen, anstatt den Begriff “Geocoding” einzusetzen.

“Geografische Themenmodelle (geografic Topic Modeling)” wird definiert als die Verbindung von “Themenmodellierung” und “Standorterkennung (Location Awareness)“. Durch klassisches “Themenmodellierung“ lässt sich aus Texten eine Menge von Themen extrahieren. Durch eine Lernphase werden Wörterbücher zu den Themen erstellt. Mit Hilfe dieser Themen-Wörterbücher kann später das Thema eines Textes bestimmt werden. [BNJ12] Unter “Standorterkennung“ wird hier verstanden, dass nicht nur das Thema sondern auch eine bestimmte Region extrahiert werden kann. Dies kann durch geografischen Koordinaten in Twitter-Kurznachrichten realisiert werden. Im Unterschied zur Kategorie “Inhaltsanalyse mit probabilistischen Sprachmodellen“ aus [CCL10] wird hier jedoch keine vorgegebene geografische Region gefordert. Vielmehr ergeben sich die geografischen Regionen aus den Themenmodellen und den zugehörigen geografischen Koordinaten. Es wird damit eine kontinuierliche Region beschrieben, welche nicht zwangsweise durch Stadt-, Staaten- oder Ländergrenzen beschränkt ist.

“Statistische Klassifizierung (Statistical classifiers)” Diese Kategorie entspricht der “Inhaltsanalyse mit probabilistischen Sprachmodellen“ wobei in [CCL10] nur eine Arbeit in dieser Kategorie betrachtet wird. [PCV13] listet mehrere Arbeiten auf, die sich in diese Kategorie einordnen lassen.

“Informationen aus sozialen Verbindungen (Social Network Information)” analog zu “Schlussfolgerungen durch soziale Verbindungen“ aus [CCL10] werden soziale Verbindungen herangezogen um den Standort zu bestimmen.

Priedhorsky et al. wählen eine ähnliche Einteilung wie vormals Cheng et al. in 2010, die verwandten Arbeiten stammen allerdings aus dem Twitter-Umfeld. Dabei ist zu bemerken, dass sich die verwendeten Methoden zur Lokalisierung im Twitter-Umfeld nicht wesentlich von denen in anderen Bereichen unterscheiden. Um die Arbeiten im Twitter-Umfeld sinnvoll voneinander abgrenzen zu können muss die Kategorisierung mehr Dimensionen umfassen. Es müssen mehr Kriterien zur Kategorisierung herangezogen werden als die reine Methodik.

Mahmud et al. betrachten in [MND12] hauptsächlich Arbeiten im Twitter-Umfeld. Diese werden in die folgenden Kategorien unterteilt.

1. “Inhaltsbasierte Standortschätzung von Tweets (Content-based Location Estimation from Tweets)”
2. “Inhaltsbasierte Standortextrahierung von Tweets (Content-based Location Extraction from Tweets)”
3. “Standortschätzung ohne den Tweet Inhalt zu nutzen (Location Estimation without using Tweets Content)”

“Inhaltsbasierte Standort-Schätzung von Tweets (Content-based Location Estimation from Tweets)” hier wird die geografische Position durch eine Inhaltsanalyse der Twitter-Kurznachricht geschätzt. Die Schätzung erfolgt dabei durch probabilistische Modelle. Diese Kategorie vereint damit “Geografische Themenmodelle“, “Statistische Klassifizierung“ aus [PCV13] mit “Inhaltsanalyse mit probabilistischen Sprachmodellen“ aus [CCL10] und ist damit als genereller anzusehen, als die vorgenannten Kategorien.

“Inhaltsbasierte Standort-Extrahierung von Tweets (Content-based Location Extraction from Tweets)” die verwandten Arbeiten in dieser Kategorie versuchen direkte Hinweise auf einen geografischen Ort aus einer Twitter-Kurznachricht zu extrahieren. Diese Kategorie ähnelt dem “Geocoding“ beziehungsweise der “Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis“.

“Standortschätzung ohne den Tweet Inhalt zu nutzen (Location Estimation without using Tweets Content)” hierunter versteht der Autor alle Informationen die nicht unmittelbar im Tweet-Text enthalten sind. Dazu zählen Informationen aus dem Nutzerprofil oder Informationen über die sozialen Verbindungen des Nutzers.

[MND12] nutzt ebenfalls die Methodik um die Arbeiten zu kategorisieren. Allerdings wird hier eine generellere Einteilung vorgenommen. So wird unterteilt, ob der Standort geschätzt oder extrahiert wurde. Mahmud et al. bringen aber auch eine weitere Dimension ein. Es wird hier zusätzlich unterschieden ob das angewendete Verfahren den Tweet-Inhalt nutzt oder andere Informationen.

Dies ist sinnvoll, denn die genannten Methoden lassen sich sowohl auf den Tweet-Inhalt als auch auf andere Informationen, beispielsweise aus dem Nutzerprofil, anwenden.

Frühere Arbeiten verweisen auf ein weiteres Spektrum an Arbeiten aus anderen Bereichen, wie Lokalisierung von Flickr Bildern oder Web-Log Einträgen. Arbeiten zur Lokalisierung im Twitter-Umfeld werden hier seltener erwähnt. In späteren Arbeiten, wie in [PCV13], wird hingegen fast ausschließlich auf Arbeiten aus dem Twitter-Umfeld verwiesen. Dies spiegelt die steigende Anzahl der Arbeiten zur Lokalisierung im Twitter-Umfeld wieder. Betrachtet man die Ausarbeitungen zur Lokalisierung im Twitter-Umfeld genauer, wird allerdings schnell klar, dass die Kategorisierung der Arbeiten anhand der verwendeten Methodik, dem Umfang nicht mehr gerecht wird.

Bei genauerer Betrachtung der Arbeiten stellt man allerdings fest, dass diese Klassifizierungen dem Umfang der Arbeiten nicht gerecht wird. [HHSC11] verweist auf ähnliche Ansätze mit einem anderen Untersuchungsgegenstand. [CCL10] kategorisiert die Arbeiten anhand der Methodik, und verweist ebenso auf andere Untersuchungsgegenstände. [PCV13] verweist ausschliesslich auf Arbeiten im Twitter-Umfeld und kategorisiert diese anhand der verwendeten Methodik. Die Methodeneinteilung ist aufgrund der Begriffswahl missverständlich und kann somit zu Problemen führen.

3.1.1 ttt<sss

In [SHP⁺13] werden die folgenden Dimensionen zur Abgrenzung herangezogen.

Allerdings lassen sich noch andere Dimensionen zur Klassifizierung der Arbeiten heranziehen. Wird beispielsweise der Text einer Twitter-Kurznachricht durch eine einfache Geokodierung untersucht wird dies andere Ergebnisse liefern als eine Untersuchung auf Basis eines geografischen Themenmodells.

[HHSC11] nutzen diese Methode um eine Ground-Truth zu bestimmen indem das Userlocation-Feld in Wikipedia nachschlagen wird. Wikipedia bietet zu vielen Artikeln eine geografische Position in Form von Längen- und Breitengrad an, diese werden dann der untersuchten Twitter-Kurznachricht zugeordnet. [HGG12] nutzen die Yahoo und die Google Geocoding Api um das Userlocation-Feld eingehender zu untersuchen.

Eine weitere zu betrachtende Dimension stellt daher der konkrete Untersuchungsgegenstand in Form des Indikators dar.

Betrachtet man die Gesamtheit an arbeiten im Bereich der Lokalisierung im Twitter Netzwerk drängen sich noch mehr Dimensionen zur Klassifizierung der arbeiten auf.

1. Räumliche Indikatoren
2. Techniken
3. Fokus der Lokalisierung

-
1. Naiver Ansatz -> Geocoding mit Google Maps API V3, nur Indikatoren die geografische Namen enthalten. Prinzipiell einfache Datenbankabfrage mit ein wenig semantik. Keine Jargon Namen wie Big Apple etc.
 - a) Funktion der GMaps Api V3
 - b) Einschränkungen der GMaps Api V3
 - c) zurückgelieferte Daten der GMaps Api V3
 - d) Kurze Beschreibung wie ich die API genutzt habe
 2. aktuelle Ansätze
 - a) Verfahren mit Inhaltsanalysen
 - b) Verfahren mit Indikatoren einzelne oder mehrere

Tabelle einfügen, bereits fertig, nur noch Format anpassen (Lesbarkeit)

Requirements
Tabelle einfügen

- c) Welche Verfahren kommen beim mapping auf geografische Entitäten zum Einsatz

geografische
Entität defi-
nieren

3.1.2 Probleme früherer Ansätze

1. Genutzte API's und Indikatoren nur in bestimmten Sprachen verfügbar
2. keine Schätzung für Genauigkeit auf verschiedenen geografischen Hierarchieebenen verfügbar

4 Lösungsansatz

In diesem Kapitel wird ein Verfahren zur Georeferenzierung von Twitter-Nutzern vorgestellt. Die Fragestellungen aus Kapitel 1.3 werden, unter Berücksichtigung der Anforderungen aus Kapitel 1.3.1, beantwortet.

Im Überblick soll zunächst gezeigt werden wie die Georeferenzierung abläuft und wie dies im realisiert werden kann. Danach soll eine generelle Struktur einer Datenbasis zur Georeferenzierung vorgestellt werden. In den darauffolgenden Abschnitten wird ganz allgemein betrachtet welche Eigenschaften geografische Indikatoren haben müssen um eine Georeferenzierung zu ermöglichen.¹ Danach sollen der Nutzer-Standort und die Nutzer-Zeitzone als mögliche geografische Indikatoren aus Twitter eingehender betrachtet werden. Basierend darauf soll dann ein Verfahren zum einlernen der Datenbasis zur Georeferenzierung erarbeitet werden. Dabei wird auch die Struktur der Datenbasis weiterentwickelt. Zum Schluss wird ein Verfahren zur Georeferezzierung vorgestellt. Es wird dazu die zuvor erarbeitete Datenbasis genutzt.

4.1 Überblick

Das erarbeitete Verfahren soll es ermöglichen Twitter-Nutzern eine Georeferenz zuzuordnen. Dabei sollen als Eingabe für die Georeferenzierung geografische Indikatoren aus dem Profil eines Twitter-Nutzers verwendet werden. Als Ergebnis soll eine Georeferenz zurückgeliefert werden. Dabei hat der Anwender die Möglichkeit einen Schwellwert für die Konfidenz der Georeferenz anzugeben. Zudem kann der Nutzer die gewünschte geografische Hierarchieebene der zurückgelieferten Georeferenz angeben. In Abbildung ?? ist der generelle Ablauf dieser Georeferenzierung dargestellt.

¹Unter einem geografischen Indikator wird ein Datum verstanden welches einen geografischen Bezug aufweist, sodass eine Georeferenz abgeleitet werden kann.

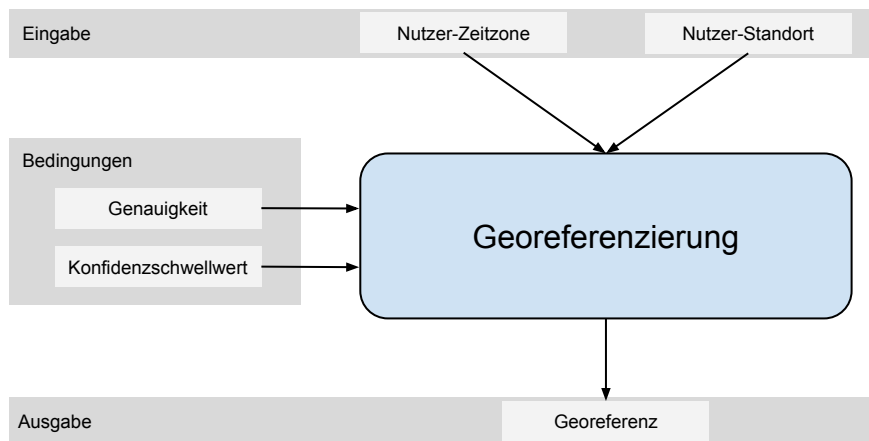


Abbildung 4.1: Ablauf Georeferenzierung

Die Georeferenzierung nach dem Schema aus Abbildung 4.1 setzt voraus, dass aus den eingegebenen geografischen Indikatoren eine Georeferenz abgeleitet werden kann. Dies ist mit einer Datenbasis, die zu einem gegebenen geografischen Indikator eine Georeferenz liefern kann möglich. Aus einer Tweet-Datensammlung soll die Zuordnung der geografischen Indikatoren zu einer Georeferenz gelernt werden. Die Tweet-Datensammlung enthält eine Menge von Tweets für welche die genutzten geografischen Indikatoren des Verfassers und eine Georeferenz in Form von Längen- und Breitengrad angegeben sind. Die gelernten Zuordnungen sollen in einer Datenbasis gespeichert werden. Diese Datenbasis wird Georeferenz-Basis genannt.

img:georef,
img:ablaufEinlern
img:Georeferenzie

4.2 Generelle Struktur einer Georeferenz-Basis zur Georeferenzierung

In diesem Kapitel soll eine generelle Struktur angegeben werden, welche die minimalen Anforderungen einer Georeferenzierung erfüllen kann. Danach wird das Schema und die nötigen Informationen, welche die Georeferenz-Basis beinhaltet, genauer betrachtet.

Nach Abbildung ?? wird bei der Georeferenzierung einem oder mehreren geografischen Indikatoren eine Georeferenz zugewiesen. In der einfachsten Variante wird lediglich ein einziger geografischer Indikator an die Georeferenzierung übergeben und genau eine Georeferenz zurückgegeben. Die Georeferenzierung muss also zu einem gegebenen geografischen Indikator eine Georeferenz bestimmen können. Dies führt zu einer ersten einfachen Struktur für die Georeferenz-Basis.

Eine erste einfache Struktur der Georeferenz-Basis

Es wird angenommen der geografische Indikator stellt immer ein eindeutiges Toponym dar. Des weiteren sind alle möglichen Toponyme sowie eine zugehörige Georeferenz bekannt. Die Georeferenz liegt als Adresse mit Straße, Hausnummer, Postleitzahl und Ortsname vor.

Jedem möglichen Toponym, welches im gegebenen geografischen Indikator vorkommen kann, soll eine Georeferenz zugeordnet werden können. Daraus ergibt sich die Anforderung, dass die Georeferenz-Basis eine Menge von Datensätzen beinhalten muss, die jeweils aus einem Toponym und einer Georeferenz bestehen. Dieser Aufbau entspricht einer Art Wörterbuch in dem Informationen zu einem gegebenen Referenzwert nachgeschlagen werden können. Im vorliegenden Fall kann also zu einem Toponym die entsprechende Georeferenz nachgeschlagen werden. Die Referenzwerte stellen dabei mögliche Werte für den geografischen Indikator dar. In Abbildung 4.1 ist ein Beispiel für eine sehr simple Struktur dargestellt.

Tabelle 4.1: Erste einfache Struktur für eine Georeferenz-Basis zur Georeferenzierung

Referenzwert	Georeferenz
Zoo-Karlsruhe	Ettlinger Straße 6 - 76137 Karlsruhe
ZKM	Lorenzstraße 19 D - 76135 Karlsruhe
Elbphilharmonie	Dammtorwall 46 - 20355 Hamburg

Wird eine Abfrage auf die Georeferenz-Basis mit den geografischen Indikatoren “Zoo-Karlsruhe“, “ZKM“ oder “Elbphilharmonie“ durchgeführt, kann nun eine Georeferenz zurückgeliefert werden. Diese simple Struktur reicht grundsätzlich aus um eine Georeferenzierungen durchführen zu können. In dem angeführten Beispiel ist die Menge der

möglichen Toponyme sehr begrenzt, aber diese kann beliebig erweitert werden. Damit können sehr mächtige Datenbanken erstellt werden.

Es sollen nun der Referenzwert und die Georeferenz genauer betrachtet werden.

Form der Georeferenz Die Form in der die Georeferenz angegeben wird ist abhängig von der Anwendung. Im Beispiel 4.1 wurden Adressen verwendet. Dazu muss das angegebene Toponym oder die Zeichenkette jedoch eine Adresse besitzen. Ein See in der Wildnis Alaskas wird keine solche Adresse aufweisen. Aber auch die geografische Position einer Stadt oder eines Landes kann nicht durch eine Adresse beschrieben werden. Die Form in der die Georeferenz angegeben wird kommt auf den jeweiligen Anwendungsfall an. Denkbar sind hier unter anderem:

- geografische Koordinaten
- vollständige Adressen
- Ländernamen
- Städtename
- Namen für Verwaltungseinheiten
- Zeitzonen
- ein Straßenname und eine Kilometerbezeichnung

Grundsätzlich sind alle Formen, welche eine direkte oder indirekte Georeferenz darstellen, denkbar. Wichtig ist nur, dass die Angabe der Georeferenz in einer Form erfolgt, welche für die gegebene Anwendung ausreichend genau ist.

Für den Straßenverkehr ist eine Angabe einer Adresse ausreichend. Für Wanderungen in unerschlossenen Gebieten hingegen sind geografische Koordinaten notwendig.

Form der Referenzwerte Der Referenzwert ergibt sich aus den möglichen Werten der Indikatoren. Die Indikatoren sind dabei zumeist Zeichenketten. Abhängig vom Indikator können diese Zeichenketten verschiedene Informationen enthalten. Im obigen Beispiel sind die Referenzwerte Toponyme. Toponyme sind Namen für geografische Objekte, es kann ihnen also unmittelbar ein geografisches Objekt zugeordnet werden. Es muss sich dabei aber nicht um Toponyme handeln. Werte die als Referenzwerte in Frage kommen sollen als geografische Indikatoren bezeichnet werden. Toponyme sind also geografische Indikatoren, da sie direkt mit einem geografischen Objekt in Verbindung gebracht werden können. Geografische Indikatoren können aber auch indirekt mit einem geografischen Objekt in Verbindung gebracht werden. Unbrauchbar sind Werte welche weder direkt noch indirekt mit einem geografischen Objekt in Verbindung gebracht werden können. Diese Werte stellen keine geografischen Indikatoren dar. Im nächsten Abschnitt werden die geografischen Indikatoren genauer betrachtet.

4.3 Geografische Indikatoren

Ein geografischer Indikator ist ein Datum welches einen geografischen Bezug aufweist. Es sind also genau die Werte, welche in irgendeiner Weise mit einem geografischen Objekt in Verbindung gebracht werden können.

Geografische Indikatoren können unmittelbaren geografischen Bezug oder mittelbaren geografischen Bezug haben.

4.3.1 Indikatoren mit unmittelbarem geografischen Bezug

Ein geografischer Indikator mit unmittelbarem geografischen Bezug beschreibt direkt eine geografische Position oder eine geografische Region. Dem geografischen Indikator kann durch die in ihm enthaltene Information direkt eine Georeferenz auf ein geografisches Objekt zugewiesen werden.

Im Beispiel in Tabelle 4.1 wurden geografische Indikatoren mit unmittelbarem geografischen Bezug verwendet. Die Referenzwerte entsprechen Toponymen, diese sind Namen für geografische Objekte. Ihnen kann also unmittelbar ein geografisches Objekt zugeordnet werden.

XXX->wenn
in Grundlagen
soll

Ein weiteres Beispiel für geografische Indikatoren mit unmittelbarem geografischen Bezug sind Zeitzone.

Bei der Zeitzone handelt es sich um einen geografischen Indikator der eine geografische Region beschreibt. Die zugehörige Georeferenz für eine Zeitzone könnten die Länder sein welche die Zeitzone umfasst. Aber auch die Kontinente oder Städte welche in dieser Zeitzone liegen könnten, je nach Anwendungsfall, von Interesse sein.

4.3.2 Indikatoren mit mittelbarem geografischen Bezug

Ein geografischer Indikator muss nicht unmittelbar einer Georeferenz zuzuordnen sein. Dies ist genau dann der Fall, wenn die im geografischen Indikator enthaltene Information nicht direkt auf ein geografisches Objekt hindeutet.

An einem Beispiel soll gezeigt werden was einen geografischen Indikator mit mittelbarem geografischen Bezug ausmacht.

Beispiel eines mittelbaren geografischen Indikators In einem Feld auf einer Website werden die Nutzer nach alternativen Begriffen für Kartoffeln gefragt. Die folgenden drei Begriffe werden eingegeben.

1. Äbierra
2. Grumbeer
3. Tüfte

Die Information die hinter jedem dieser Begriffe steckt ist die Bezeichnung eines Gemüses. Die Begriffe bezeichnen also insbesondere kein geografisches Objekt und haben somit keinen unmittelbaren geografischen Bezug. Jede dieser Bezeichnungen stammt aber aus unterschiedlichen Regionen Deutschlands, denn es handelt sich um dialektische Begriffe. Durch ihre geografisch begrenzte Verwendung können sie damit einer geografischen Region zugeordnet werden. Durch die Verwendung eines dieser Worte kann also auf eine Region Deutschlands geschlossen werden. Dadurch kann jedem Begriff eine Georeferenz zugewiesen werden.

Mit folgender Georeferenz-Basis kann eine Georeferenzierung der obigen Begriffe durchgeführt werden:

Tabelle 4.2: Kartoffeln in verschiedenen Dialekten

Referenzwert	Georeferenz
Äbbiera	Württemberg
Grumbeer	Pfalz
Tüfte	Norddeutschland

In diesem Beispiel wurde die Zuordnung zu einem geografischen Objekt über die Verwendung der Begriffe in bestimmten Regionen ermöglicht. Aber auch andere Informationen können dazu dienen einem Indikator eine Georeferenz zuzuordnen. Es kommt grundsätzlich nicht darauf an was genau mit dem Begriff bezeichnet wird. Lediglich die geografisch begrenzte Verwendung kann hier als Hinweis auf eine Georeferenz dienen.

4.4 Toponyme als geografische Indikatoren

Toponyme sind die am häufigsten auftretende Art geografischer Indikatoren. Es handelt sich dabei um geografische Indikatoren mit unmittelbarem geografischen Bezug.

Die Georeferenzierung von Toponymen kann bereits mit der Struktur aus Tabelle 4.1 umgesetzt werden. Jedes Toponym wird dabei als Referenzwert gespeichert und eine Georeferenz zugewiesen. Es wurden allerdings mehrere Annahmen gemacht, die in der Regel nicht gelten. Anhand dieser Annahmen soll auf die Probleme bei der Verwendung von Toponymen als geografische Indikatoren eingegangen werden.

XXX-
> Annahmen
raus, können
nicht mehr
verwendet
werden

4.4.1 Probleme bei der Verwendung von Toponymen

Im folgenden soll auf diese Annahmen eingegangen werden und gezeigt werden wie-so diese im Allgemeinen nicht gelten. Des weiteren werden generelle Probleme bei der Verwendung von Toponymen zur Georeferenzierung aufgezeigt.

Alle möglichen Toponyme sind bekannt

Es existieren sehr umfangreiche Datenbanken, die Toponymen Georeferenzen zuweisen. Eine dieser Datenbanken ist das frei erhältliche Ortsverzeichnis von geonames.org. Es beinhaltet ca. 8,9 Millionen Toponyme und zugehörige Georeferenzen. Zusätzlich sind weitere ca. 8 Millionen alternative Toponyme hinterlegt. In den Datensätzen sind neben einer Georeferenz in Form geografischer Koordinaten noch andere Informationen hinterlegt wie beispielsweise Länderkürzel und Einwohnerzahlen. Auch eine geografische Hierarchie wird in der Datenbank abgebildet. Grundsätzlich ist die Datenbank aber nach obiger Struktur aufgebaut.

Trotz der großen Anzahl an hinterlegten Toponymen beinhalten solche Datenbanken nicht alle möglichen Toponyme. Aufgrund der immensen Vielfalt ist es nahezu unmöglich alle Toponyme abzudecken. Im folgenden soll an einigen Beispielen die Vielfalt der Toponyme belegt werden.

Vielfalt der Toponyme Da Toponyme im allgemeinen nicht definiert sind, und neben den offiziellen Namen für Städte, Länder usw. weitere Namen vorkommen können, ist die Vielfalt nahezu unbegrenzt.

In Wikipedia sind für die Stadt Detroit, im US-Bundesstaat Michigan, folgende Spitznamen angegeben:

- The Motor City
- Motown
- Hockeytown
- Rock City
- The D

Die ersten zwei dürften weltweit einen gewissen Bekanntheitsgrad haben. Hockeytown, Rock City und The D dürften allerdings weniger bekannt sein. Tatsächlich beinhaltet die geonames.org Datenbank keinen dieser Spitznamen. Google Maps hingegen bietet bei Eingabe der oben genannten Spitznamen Detroit als Vorschlag an.

Das Problem wird noch gravierender wenn man lokal begrenzte Spitznamen einbezieht. Es ist durchaus denkbar, dass Spitznamen für einen Stadtteil nur lokal verwendet werden und über die Stadt- oder Landesgrenze hinaus nicht bekannt ist. Diese Spitznamen können nur schwer erfasst werden. Solche Spitznamen existieren auch für Länder und Verwaltungseinheiten.

Weitere Beispiele die zu einer größeren Vielfalt der Toponyme führen:

- Historische Toponyme
- Kulturell bedingte Toponyme
- Toponyme für landschaftliche Besonderheiten
- Toponyme für bestimmte Landschaften

Die Toponyme sind eindeutig geografischen Objekten zuzuweisen und umgekehrt

Auch dies ist im allgemeinen nicht der Fall. Toponyme sind oft Doppel- oder Mehrdeutig und verweisen somit auf mehrere Georeferenzen. Allerdings können zu einer Georeferenz auch mehrere eindeutige Toponyme existieren. Doppel- und Mehrdeutigkeiten können also sowohl ausgehend von der Georeferenz, als auch ausgehend vom Referenzwert auftreten.

Zu einem Referenzwert können mehrere Georeferenzen existieren Es gibt zahlreiche Städte-Namen, die in mehreren Ländern verwendet werden. Ein gutes Beispiel hierfür sind US Städte. Da die USA ein Einwanderungsland ist, übernahmen viele Einwanderer bei der Gründung neuer Städte die Namen aus der alten Heimat. So finden sich in den USA zahlreiche Städte deren Namen exakt den deutschen Städtenamen entsprechen. In Tabelle ?? sind einige Städte-Namen und die Vorkommen in den USA aufgelistet.

Tabelle 4.3: Häufige deutsche Städtenamen in den USA

Name	Anzahl in den USA
Hannover	40
Berlin	39
Hamburg	30

Soll nun eine Georeferenz-Basis angelegt werden, welche einem Stadt-Namen einen Staat zuweist müssten für Hannover 41, für Berlin 40 und für Hamburg 31 Georeferenzen hinterlegt sein. Als Ergebnis einer Georeferenzierung für den Wert “Hamburg“ würden 31 Georeferenzen in den USA und eine in Deutschland zurückgegeben werden.

Diese Mehrdeutigkeit stellt ein Problem bei der Georeferenzierung dar. Es kann keine eindeutige Entscheidung getroffen werden welche Georeferenz dem Ort zugewiesen werden soll.

Zu einer Georeferenz können mehrere Referenzwerte existieren Dies wird anhand der Städte-Spitznamen klar. Städte-Spitznamen sind oft eindeutig, allerdings kann eine Stadt mehrere dieser eindeutigen Spitznamen besitzen. Damit ergibt sich eine Mehrdeutigkeit ausgehend von der Georeferenz zu einem Toponym. Dies ist grundsätzlich auf alle geografischen Objekte übertragbar. Zu jedem geografischen Objekt existieren potenziell mehrere gültige Toponyme.

Das Problem hierbei stellt nicht die Beziehung der Georeferenz zu den Toponymen dar, sondern eher die Vielfalt der Toponyme.

4.4.2 Domänenspezifische Toponyme

Ein weiteres Problem kann die untersuchte Domäne darstellen. In sozialen Netzwerken können sich eigene Begriffe und Formulierungen etablieren welche im allgemeinen nicht bekannt sind.

Im Twitter-Umfeld haben sich in den letzten Jahren beispielsweise einige spezielle Begriffe und Formulierungen zur Verwendung in Tweet-Texten etabliert. Das im Twitter-Umfeld spezielle Toponyme verwendet werden, kann nicht gänzlich ausgeschlossen werden.

Ein Beispiel hierfür ist “Bieberville“ welches in den untersuchten Daten von Hecht et al. öfter vorkommt. “Bieberville“ wird abgeleitet von dem Pop-Star Justin Bieber. Twitter wird oft als “Bieberville“ bezeichnet, da der Pop-Star in Twitter sehr aktiv ist und deshalb viele Fans auch in Twitter aktiv sind. Unter diesem Gesichtspunkt hätte “Bieberville“ keinen geografischen Bezug. Sucht man allerdings im Internet nach “Bieberville“

stößt man auf einen Imbiß in Groß-Bieberau. “Bieberville“ kann also durchaus einen geografischen Bezug haben, wenngleich es im Twitter-Umfeld nicht als solcher benutzt wird. Ein Nutzer-Standort der in einem Land kein Toponym darstellt, kann in einem anderen durchaus ein Toponym sein.

Die Benutzung von “Bieberville“ könnte auch ein temporäres Phänomen darstellen. In den Tweet-Daten welche für diese Arbeit verwendet werden befindet sich kein Eintrag mit dem Nutzer-Standort “Bieberville“. Dies könnte ein Hinweis auf eine temporäre Verwendung des Begriffes im Nutzer-Standort sein. Leider ist es nicht möglich Nutzer nach dem Nutzer-Standort zu suchen und deshalb kann diese Aussage nicht mit Sicherheit getroffen werden. Gibt es Toponyme die tatsächlich nur temporär auftauchen wird das Problem noch größer. Dies ist durchaus denkbar, denn mit wenigen Klicks und einer kurzen Eingabe kann der Nutzer-Standort geändert werden.

Das Wissen über solche Begriffe und Formulierungen kann nur sehr schlecht erfasst werden, wenn man die entsprechende Domäne nicht untersucht.

Bieberville
beispiel in ler-
nen

4.4.3 Fazit

Es existieren sehr umfangreiche Datenbasen um Toponymen eine Georeferenz zuzuweisen. Toponyme unterliegen grundsätzlich keiner Kontrolle und sind nicht standardisiert womit der Veilfalt keine Grenzen gesetzt sind. Es ist schwer, wenn nicht sogar unmöglich, das Wissen über geografische Objekte und deren zugehörige Toponyme vollständig zu erfassen.

4.5 Der Nutzer-Standort

Der Nutzer-Standort eines Twitter-Nutzers soll als geografischer Indikator verwendet werden. Bei der Eingabe des Nutzer-Standortes wird vom Nutzer abgefragt, wo dieser sich befindet. Die Intention der Abfrage zielt also darauf ab, dass der Nutzer einen Wert eingibt der auf ein geografisches Objekt verweist, welches seinem Standort entspricht. Es ist naheliegend, dass der Nutzer seinen Standort mit Hilfe eines Toponyms angibt. Der

Nutzer-Standort wird jedoch vom Nutzer frei eingegeben und unterliegt keinerlei Kontrolle. Dieser muss also nicht zwangsweise Werte mit geografischem Bezug enthalten.

Da allerdings Toponyme zu erwarten sind, können alle in Kapitel 4.4 erwähnten Probleme auftreten. Durch die unkontrollierte Eingabe sind tatsächlich alle möglichen Toponyme denkbar. Aber auch geografische Indikatoren mit mittelbarem geografischem Bezug oder Werte ohne geografischen Bezug können im Nutzer-Standort angegeben werden.

Hier soll nun der Nutzer-Standort genauer betrachtet werden und ob dieser überhaupt zur Erzeugung geografischer Indikatoren geeignet ist.

Geografischer Bezug des Nutzer-Standorts?

Nach Hecht et al. haben in [HHSC11] den Nutzer-Standort eingehend untersucht und sind zu folgendem Ergebnis gekommen: Wenn der Nutzer überhaupt einen Standort angegeben konnte in 80% Prozent der Fälle ein geografischer Bezug nachgewiesen werden. In den restlichen 20% der Fälle konnte im Nutzer-Standort kein geografischer Bezug nachgewiesen werden. Die Untersuchung wurde manuell vorgenommen und es durften alle zur Verfügung stehenden Mittel zur Analyse der Daten verwendet werden. Hecht et al. haben deshalb ausschließlich Daten untersucht die nachweislich aus den USA stammten.

Es kann hier also festgehalten werden, dass 80% der Nutzer-Standorte als geografischer Indikator verwendet werden können. Es muss beachtet werden, dass nur Daten aus den USA betrachtet wurden.

Im Rahmen der vorliegenden Arbeit wurde der Nutzer Standort von 1000 Twitter-Nutzern ebenfalls manuell untersucht.² Dabei wurde keinerlei Einschränkung zur Herkunft gemacht. Um zu prüfen ob ein geografischer Bezug vorliegt und eine Georeferenz zuzuordnen wurden Ortsverzeichnisse von Google-Maps und Geonames.org verwendet. Diese lassen es zu auch Nutzer-Standorte die in unbekannten Sprachen und Alphabeten verfasst wurden zu untersuchen. Es konnte dabei in 76% der Fälle ein geografischer Bezug nachgewiesen werden.

²siehe ??

In den restlichen 24% der Fälle konnte kein geografischer Bezug mit Hilfe der Datenbanken nachgewiesen werden. Dies bedeutet nicht, dass grundsätzlich kein geografischer Bezug vorhanden ist. Es konnte lediglich anhand der genutzten Quellen kein geografischer Bezug hergeleitet werden. Beispielsweise wurde “Swag City“ nicht zugewiesen, denn der Spitzname für die Stadt Ann Arbor war in den Datenbanken nicht hinterlegt.

Der Nutzer-Standort kann also in vielen Fällen einen Hinweis auf die Herkunft des Twitter-Nutzers geben. Bei der durchgeführten Untersuchung wurden insbesondere nur Toponyme und damit geografische Indikatoren mit unmittelbarem geografischen Bezug identifiziert.

Genauigkeit der geografischen Angaben

In ca. 80% der Fälle kann bei einer Angabe des Nutzer-Standortes davon ausgegangen werden, dass dieser geografischen Bezug hat. Wiederum aufgrund dessen, dass der Nutzer beliebige Eingaben machen kann, ist nicht sicher wie genau ein solcher geografischer Bezug ist.

Hecht et al. analysierten ihre Daten auch darauf wie genau die Nutzer ihren Standort angeben. Dabei ist wiederum zu beachten, dass die Daten aus den USA stammen und deshalb die Verwaltungseinheiten der USA zugrunde gelegt wurden. Dabei wurden folgende Werte festgestellt:

- 64% Stadt
- 20% Staat (Administrationsebene erster Ordnung)
- ca. 8% Intrastate (Administrationsebene zweiter Ordnung)
- ca. 5% Land

Die restlichen 13% entfallen auf Interstate Regionen, Nachbarschaften und konkrete Adressen. Interstate Regionen sind Regionen die sich über mehrere Staaten hinwegziehen. Beispiele für Interstate Regionen sind “Central United States“ oder “West-Coast“. Nachbarschaften(Neighbourhoods) sind oft Stadtteile wie “Harlem“ oder “Bronx“ in New York.

In den eigenen Untersuchungen sind folgende Werte (gerundet) festgestellt worden:

- 77% Stadt
- 8% Administrationsebene erster Ordnung
- 4% Administrationsebene zweiter Ordnung
- 10 % Land

Es ist also, selbst wenn ein geografsicher Bezug besteht, nicht sicher welcher geografischen Hierarchieebene dieser zugeordnet werden kann. Die unterschiedlichen Ergebnisse können dadurch Zustande kommen, dass sich Hecht et al. auf Tweets aus den USA beschränkt haben.

Stimmt der tatsächliche Standort mit dem angegeben überein?

Diese Frage ist grundsätzlich nur schwer zu beantworten. Die einzige Möglichkeit besteht darin Tweets mit geografischen Angaben zu untersuchen. Dabei können aber Probleme auftreten. Der Nutzer-Standort gibt oft nicht den Aufenthaltsort sondern den Heimatort an. Befindet sich der Nutzer auf Reisen oder arbeitet er in einer anderen Stadt korrespondiert die geografische Koordinate nicht mit dem angegeben Ort. Dies könnte in einige Fällen gelöst werden indem mehrere Tweets desselben Nutzers über einen längeren Zeitraum untersucht werden. Um den eigentlichen Wohnort sollte dadurch eine Häufung an abgesetzten Tweets auftreten. Damit kann ein genauerer Standort bestimmt werden. Der Nutzer kann aber auch absichtlich einen falschen Standort angeben oder bei Umzug in eine andere Stadt oder ein anderes Land schlicht vergessen haben den Nutzer-Standort anzupassen.

In der durchgeführten Untersuchung wurde festgestellt, dass auf Städtebene 60% der Tweets in weniger als 24 Kilometern Entfernung von der manuell ermittelten Stadt abgesetzt wurden. 25% der Tweets sind aber mehr als 239 Kilometer von der Manuell ermittelten Stadt entfernt. Der Median liegt bei 11 Kilometern. In Tabelle 4.4 sind weitere Informationen über die Distanzen zusammengetragen.

Tabelle 4.4: Distanzen

Median	11,5
0,25 Quantil	3,1
0,50 Quantil	11,5
0,60 Quantil	24,3
0,75 Quantil	239,0

Partieller geografischer Bezug des Nutzer-Standortes

In einigen der Einträge konnte festgestellt werden, dass nur Teile der Einträge geografischen Bezug haben. Oft wollen die Nutzer noch weitere Informationen geben, diese haben oft keinen Nutzen für eine Georeferenzierung. In der folgenden Liste sind einige Einträge von Nutzer-Standorten aufgelistet.

1. 11th Dimension | California
2. between here and there - Miami

Im ersten Fall kann für Kalifornien ein geografischer Bezug festgestellt werden. 11th Dimension hat hier offensichtlich keinen geografischen Bezug. Im zweiten Fall ist die Aussage “between here and there“ nicht zu gebrauchen, Miami kann jedoch als Bezug zur Stadt Miami in Florida, USA gebracht werden.

Es können also auch nur Teile des Nutzer-Standorts für eine Georeferenzierung von Nutzen sein.

Widersprüchliche Werte im Nutzer-Standort

Es existieren auch Einträge für Nutzer-Standorte welche mehrere widersprüchliche Angaben machen. Das bedeutet es werden zwei oder mehr Werte mit geografischem Bezug angegeben.

Auch hier sollen einige Beispiele genannt werden:

- Bolton \ / Leigh
- Liverpool \ / London
- Balikesir \ / Izmir

In diesen Beispielen sind jeweils zwei Städte angegeben. Bolton und Leigh liegen 14km auseinander. Liverpool und London trennen ca. 350km. Balikesir und Izmir ca. 180km.

Es kann nun spekuliert werden wieso der Nutzer zwei Städte angibt. Ist er in einer Stadt aufgewachsen und lebt momentan in der anderen? Pendelt er zwischen den Städten um zu arbeiten?

Wie dem auch sei, es kann nicht eindeutig entschieden werden in welcher Stadt sich der Nutzer aufhält.

Geografische Hierarchien im Nutzer-Standort

Es ist auch möglich das im Nutzer-Standort teile einer geografischen Hiararchie angegeben sind. Die Angabe von Stadt in Kombination mit einem Land. In den USA wird auch oft die Stadt und der zugehörige Bundesstaat angegeben. Hier einige Beispiele:

- Los Angeles, California
- Mato Grosso, Brazil
- Bronx, New York

Im ersten Beispiel handelt es sich um die Stadt Los Angeles und den US-Bundesstaat Kalifornien. Im zweiten Beispiel wird zuerst ein brasilianischer Bundesstaat angegeben und danach das Land. Im dritten Beispiel wird ein Stadtteil von New York angegeben und dann die Stadt oder der US-Bundesstaat. Da der US-Bundesstaat denselben Namen hat wie die Stadt, und beide in einer Hierarchiebeziehung mit Bronx stehen, kann hier nicht unterschieden werden was gemeint ist.

4.5.1 Fazit

Der Nutzer-Standort eignet sich besonders gut als geografischer Indikator, allerdings sind die Eingaben nicht standardisiert und können somit nicht ohne Vorverarbeitung verwendet werden. Wie oben dargestellt kann der Nutzer-Standort nur partiellen geografischen Bezug haben oder mehrere Informationen verschiedener geografischer Hierarchieebenen beinhalten. Auch kann der Nutzer-Standort keinen geografischen Bezug haben. Dies

macht es grundsätzlich schwierig den Nutzer-Standort direkt als geografischen Indikator zu verwenden. Der Nutzer-Standort ist genaugenommen nur ein potenzieller geografischer Indikator, da nicht mit Sicherheit gesagt werden kann ob ein geografischer Bezug besteht oder nicht. Dies resultiert aus der völlig freien Eingabe durch den Nutzer.

Ist ein geografischer Bezug des Nutzer-Standortes nachzuweisen, handelt es sich bei dem Eintrag in den meisten Fällen um ein Toponym. Durch die durchgeführte Untersuchung konnte gezeigt werden, dass ca. 76% der Nutzer-Standorte mit einem geografischen Bezug, auf Toponyme zurückzuführen sind. Im Hinblick auf den geografischen Bezug der Einträge im Nutzer-Standort ergab sich kein signifikanter Unterschied zu den Untersuchungen von Hecht et al..

Zusammenfassend kann gesagt werden, dass der Nutzer-Standort häufig einen geografischen Bezug aufweist und es sich in diesen Fällen um Toponyme handelt. Aufgrund der völlig freien Eingabe durch den Nutzer ergeben sich bei der Auswertung der Nutzer-Standorte jedoch einige Probleme die gelöst werden müssen.

4.6 Die Nutzer-Zeitzone

In Kapitel 4.4 wurden auf die Probleme bei der Verwendung von Toponymen als geografische Indikatoren eingegangen. Dabei wurde die Doppel- und Mehrdeutigkeit von Toponymen betrachtet. Bei der Verwendung des Nutzer-Standortes kann dieses Problem auch auftauchen.

Um diesem Problem zu begegnen soll nun ein weiterer geografischer Indikator hinzugezogen werden, die Nutzer-Zeitzone. Zunächst sollen die generellen Eigenschaften der Nutzer-Zeitzone erläutert werden bevor erklärt wird wie die Nutzer-Zeitzone das Problem der Mehrdeutigkeit von Toponymen beheben kann.

4.6.1 Eigenschaften der Nutzer-Zeitzone

Die Nutzer-Zeitzone stellt einen unmittelbaren geografischen Indikator dar. Sie beschreibt eine eindeutige geografische Region auf dem Globus. Dabei entsprechen die Grenzen der

Region nicht unbedingt den Landesgrenzen oder den Grenzen sonstiger administrativer Verwaltungseinheiten. In Abbildung 4.2 sind die Zeitzonen der Erde dargestellt.



Abbildung 4.2: Zeitzonen der Erde.

Die Nutzer-Zeitzone kann in Twitter über eine Liste gewählt werden. Der Wert der Nutzer-Zeitzone stellt deshalb garantiert eine Zeitzone dar. Es ist dem Nutzer nicht möglich einen Wert einzugeben der nicht einer Zeitzone entspricht.

Allerdings ist nicht gesichert, dass der angegebene Wert auch der Zeitzone entspricht in der sich der Nutzer aufhält. Der Nutzer könnte eine bewusste Fehleingabe machen oder aber die Zeitzone nicht wählen womit der Standardwert “Pacific Time (US and Canada)” gewählt wird. In Abbildung ?? wurden Tweets anhand ihres Längen- und Breitengrades platziert. In der Abbildung sind Tweets aus den USA zu sehen. Die Farben entsprechen den gewählten Zeitzonen.

- Pacific Time Blau
- Eastern Time Rot
- Central Time Grün
- Mountain Time Pink

Quelle:
<http://www.zeitzone.net>

Bild Auswahl-
liste

Die Zeitzonen sind durch die Farben gut zu erkennen. Lediglich die dünn besiedelte Region der Mountain Time kann nur an einigen Ballungszentren erkannt werden. Grundsätzlich scheint der Großteil der Angaben aber korrekt zu sein.

4.6.2 Auflösen von Doppeldeutigkeiten

Mit der Nutzer-Zeitzone lassen sich Doppeldeutigkeiten auflösen.

Besteht zu einem Nutzer-Standort eine Doppeldeutigkeit, kann nicht entschieden werden welches geografische Objekt zugeordnet werden soll. Liegen die beiden geografischen Objekte allerdings in zwei unterschiedlichen Zeitzonen und der Nutzer hat diese angegeben kann die Doppeldeutigkeit aufgelöst werden indem zusätzlich die Zeitzone betrachtet wird. Somit kann dem Nutzer die korrekte Georeferenz zugewiesen werden. Voraussetzung dazu ist natürlich, dass die geografischen Objekte in zwei unterschiedlichen Zeitzonen liegen und die Nutzer-Zeitzone angegeben ist.

4.7 Verfahren zum einlernen geografischer Indikatoren am Beispiel von Twitter

Aus einer Tweet-Sammlung sollen die genutzten geografischen Indikatoren und deren Zuordnung zu einer Georeferenz automatisch gelernt werden. Es soll eine Georeferenz-Basis erzeugt werden, welche es ermöglicht dem genutzten geografischen Indikatoren eine Georeferenz zuzuweisen.

Der Vorteil besteht darin, dass eine domänenspezifische Georeferenz-Basis geschaffen wird welche potenziell mehr geografische Indikatoren zuordnen kann als ein normales Ortsverzeichnis. Es können dadurch domänenspezifische Eigenheiten berücksichtigt werden. Auch domäneninterne Begriffe sollen hierdurch gelernt werden können.

Es ist zunächst nötig die genutzten geografischen Indikatoren einer Vorverarbeitung zu unterziehen. Insbesondere aus dem Nutzer-Standort können durch eine Vorverarbeitung

weitere Informationen extrahiert werden. Des weiteren wird die in Abschnitt 4.2 vorgestellte Struktur der Georeferenz-Basis angepasst und erweitert. Dabei wird am grundsätzlichen Prinzip nichts geändert, es wird nach wie vor einem Referenzwert eine Georeferenz zugewiesen. Die Referenzwerte leiten sich dabei aus den geografischen Indikatoren ab.

4.7.1 Vorverarbeitung des Nutzer-Standortes und der Nutzer-Zeitzone

Ziel ist es aus dem Nutzer-Standort möglichst viele Informationen zu extrahieren und etwaige Probleme zu beseitigen. In der folgenden Liste sind einige Nutzer-Standorte angegeben. Anhand dieser Liste sollen die Vorverarbeitungsschritte demonstriert werden.

- | | |
|---------------------------------|---------------------------------|
| 1. Bélem-PA | 9. Los Angeles, USA |
| 2. West Sussex, England | 10. I ♥ New York |
| 3. South Florida | 11. †~ Los Angeles~† |
| 4. Pitmedden, Scotland, UK | 12. earth-sea |
| 5. Mato Grosso & Rio de Janeiro | 13. In front of the computer |
| 6. -****- | 14. 11th Dimension California |
| 7. USA \ / Los Angeles | 15. York |
| 8. Nottingham \ / London | 16. York |

Eliminierung von Sonder- und Satzzeichen

Es fällt auf, dass oft Sonder- und Satzzeichen verwendet werden. Beispielsweise als Trenner zwischen Toponymen unterschiedlicher geografischer Hierarchieebenen wie bei “West Sussex, England“, “USA \ / Los Angeles“ oder “Bélem-PA“. Das Trennzeichen wird dabei nicht einheitlich verwendet. Es kann deshalb nicht entschieden werden ob ein Satzzeichen als Trenner zweier Hierarchieebenen genutzt wird oder nicht. Bei “USA \ / Los Angeles“

wird \setminus als Trenner für Hierarchieebenen verwendet bei “Nottingham \setminus London“ werden zwei Städte angegeben. Es ist also insbesondere nicht klar welcher Zusammenhang zwischen den Toponymen, die durch ein Zeichen getrennt sind, besteht.

Bei “I ♥ New York “ werden Sonderzeichen zum ausdrücken von Emotionen genutzt. In “†~ Los Angeles~†“ werden Sonderzeichen als Dekoration genutzt. Einige Nutzer-Standorte bestehen ausschließlich aus Sonder- und Satzzeichen.

Grundsätzlich ist es aufgrund der Vielfalt schwierig definitiv zu entscheiden ob Satz- und Sonderzeichen einen Mehrwert bieten bei der Georeferenzierung. Es gibt Fälle in denen Sonder- oder Satzzeichen Bestandteil des Toponyms sind. Ein Beispiel hierfür wäre das “3. Arrondissement“ in Paris, das weglassen des Punktes würde hier aber zu keinen Problemen führen. Das hinzufügen von Satz- und Sonderzeichen kann allerdings mehr Probleme mit sich bringen als es Vorteile bringt. Es sollen deshalb in einem ersten Vorverarbeitungsschritt alle Sonder- und Satzzeichen entfernt werden.

Liste nach dem entfernen von Sonder- und Satzzeichen:

- | | |
|-------------------------------|-------------------------------|
| 1. Bélem PA | 9. Los Angeles USA |
| 2. West Sussex England | 10. I New York |
| 3. South Florida | 11. Los Angeles |
| 4. Pitmedden Scotland UK | 12. earth sea |
| 5. Mato Grosso Rio de Janeiro | 13. In front of the computer |
| 6. | 14. 11th Dimension California |
| 7. USA Los Angeles | 15. York |
| 8. Nottingham London | 16. York |

Der Wert 6 existiert nun nicht mehr, der Wert ist leer und wird somit nicht weiter betrachtet.

Der Vorteil besteht darin, dass unnötige Sonder- und Satzzeichen entfernt werden und die entstandenen Werte einfacher weiter verarbeitet werden können.

Zusammenfassen von Toponymen

In diesem Schritt sollen Toponyme, welche aus zwei oder mehr Teilen bestehen mit Hilfe eines Ortsverzeichnisses zusammengefasst werden. Dies kann nur für bekannte Toponyme durchgeführt werden. Damit bilden beispielsweise “Los“ und “Angeles“ eine Einheit. Dies soll mit einem + Zeichen gekennzeichnet werden.

Daraus resultiert:

- | | |
|-------------------------------|-------------------------------|
| 1. Bélem PA | 9. I New+York |
| 2. West+Sussex England | 10. Los+Angeles |
| 3. South Florida | 11. earth sea |
| 4. Pitmedden Scotland UK | 12. In front of the computer |
| 5. Mato+Grosso Rio+de+Janeiro | 13. 11th Dimension California |
| 6. USA Los+Angeles | 14. York |
| 7. Nottingham London | 15. York |
| 8. Los+Angeles USA | |

In diesem Schritt wird insbesondere keine Georeferenzierung vorgenommen. Dieser Schritt dient dazu, möglichst früh vorhandenes Wissen über Toponyme einzubeziehen. Im wesentlichen sollen die Werte hiermit für die nächsten Verarbeitungsschritte vorbereitet werden.

Alphanumerische Sortierung

In diesem Schritt sollen die Werte alphanumerisch sortiert werden. Dadurch ist die Reihenfolge in der die Werte angegeben werden unerheblich.

Nach der Sortierung liegt folgende Tabelle vor:

- | | |
|------------------------|--------------------------|
| 1. Bélem PA | 3. Florida South |
| 2. England West+Sussex | 4. Pitmedden Scotland UK |

- | | |
|-------------------------------|-------------------------------|
| 5. Mato+Grosso Rio+de+Janeiro | 11. earth sea |
| 6. Los+Angeles USA | 12. computer front In of the |
| 7. London Nottingham | 13. 11th California Dimension |
| 8. Los+Angeles USA | 14. York |
| 9. I New+York | 15. York |
| 10. Los+Angeles | |

Die Werte 6 und 8 sind nun gleich. Durch die Sortierung werden Werte mit gleichem Inhalt aber unterschiedlicher Reihenfolge angeglichen. Durch das zusammenfassen der Toponyme im vorherigen Schritt werden bekannte Toponyme nicht getrennt. Es wird eine unnötige Fragmentierung der Werte vermieden.

Dieser Schritt stellt allerdings einen Kompromiss dar. Es werden zwar Werte mit gleichem Inhalt und unterschiedlicher Reihenfolge angeglichen. Aber es werden auch potenzielle Toponyme, die aus mehreren Teilen bestehen, auseinandergezogen. Aus “Motor City Michigan USA“ würde “City Michigan Motor USA“ entstehen. Der Zusammenhang zwischen Motor und City wäre nicht mehr vorhanden und könnte daraus auch nicht wiedergewonnen werden.

Erzeugung von N-Grammen

Es sollen nun N-Gramme bis zum Grad 3 erzeugt werden. Dieses Vorgehen löst gleich mehrere Probleme.

Zum ersten können sowohl geografische Indikatoren als auch Werte die kein geografischen Indikator darstellen in den Nutzer-Standorten vorhanden sein. Durch die Erzeugung von N-Grammen können diese getrennt voneinander betrachtet werden. Es löst das Problem des partiellen geografischen Bezugs des Nutzer-Standortes aus 4.5 in der Hinsicht, dass die Werte für die weitere Verarbeitung getrennt betrachtet werden können.

Zum zweiten können in einem Nutzer-Standort mehrere geografische Indikatoren enthalten sein. Der Wert “Pitmedden Scotland UK“ enthält mit “UK“ das Land, mit “Scotland“ den Landesteil und mit “Pitmedden“ eine Stadt. Mit der Georeferenz aus dem zugehörigen Tweet kann nun allen drei Werten eine Georeferenz zugeordnet werden.

Aber auch zwei verschiedene Städte wie “Nottingham\London“ können in einem Wert vorkommen. Diese werden zu “Nottingham\London“, “Nottingham“ und “London“. Wird hier wiederum beiden Werten die geografische Koordinate zugeordnet können drei Fälle auftreten. Erstens: Der Nutzer war in keiner der beiden Städte als er den Tweet abgesetzt hat. Damit sind beide Zuordnungen unbrauchbar. Zweitens: Er war in einer der beiden Städte, dann ist zumindest einer der entstandenen Datensätze brauchbar. Dies löst das Problem aus 4.5 in der Hinsicht, dass die Werte für die weitere Verarbeitung getrennt betrachtet werden können.

Besteht allerdings eine Beziehung der Werte zueinander, kann diese durch Bi- und Trigramme abgebildet werden. Aus “Pitmedden Scotland UK“ wird “Pitmedden Scotland“, “Scotland UK“ und “Pitmedden Scotland UK“ erzeugt.

Dieser Schritt soll nur an einigen ausgewählten Beispielen aus der Liste erfolgen. Die Bestandteile der N-Gramme werden zur Verdeutlichung mit <> gekennzeichnet in späteren Beispielen wird dies weggelassen, da durch das zusammenfassen von Werten mit einem plus klar ist welches die Elemente des NGramms darstellen

- | | |
|----------------------------------|----------------------------|
| 1. <England><West+Sussex> | 8. <11th><California> |
| 2. <England> | 9. <California><Dimension> |
| 3. <West+Sussex> | 10. <11th> |
| 4. <Mato+Grosso><Rio+de+Janeiro> | 11. <California> |
| 5. <Rio+de+Janeiro> | 12. <Dimension> |
| 6. <Mato+Grosso> | 13. <York> |
| 7. <11th><California><Dimension> | 14. <York> |

Der Wert 7 ist ein Trigramm. Bei den Werten 1,4,8 und 9 handelt es sich um Bigramme. Die Werte von 2,3,5,6,10,11 und 12 sind Unigramme. Auch hier entsteht aufgrund des zusammenfassens bekannter Toponyme keine Fragmentierung, da zusammengehörige Toponyme gemeinsam betrachtet werden.

In diesem Schritt werden aus dem Nutzer-Standort mehrere potenzielle geografische Indikatoren erzeugt die als Referenzwerte genutzt werden könne. Bevor diese als Referenzwerte genutzt werden soll noch die Nutzer-Zeitzone hinzugefügt werden.

Hinzufügen der Zeitzone

Es soll an dieser Stelle die Zeitzone hinzugefügt werden. Da die Zeitzone eine begrenzte Anzahl an Werten darstellt und diese vom Nutzer nicht frei eingegeben werden können muss hier keine weitere Vorverarbeitung vorgenommen werden.

Jeder Wert der in der aktuellen Liste vorkommt soll einmal mit und einmal ohne Zeitzone existieren. Damit wird garantiert, dass eingelernte Referenzwerte die eine falsche Zeitzone aufweisen, trotzdem berücksichtigt werden können. Beispielsweise ist die Nutzer-Zeitzone “Pacific Time (US & Canada)” für den Nutzer-Standort “Jakarta” nicht korrekt. Aus dieser Kombination würden die Referenzwerte “Jakarta Pacific Time (US & Canada)” und “Jakarta” entstehen. Bei einer Abfrage von “Jakarta” kann dann der Referenzwert “Jakarta” genutzt werden. Würde allerdings nur die Kombination “Jakarta Pacific Time (US & Canada)” als Referenzwert existieren würde keine Georeferenz zurückgeliefert werden können.

Die Elemente der Zeitzonen werden wiederum mit einem Plus zusammengefasst um deutlich zu machen, dass es sich um ein Element eines NGramms handelt. Um die Nutzer-Zeitzone von den aus dem Nutzer-Standort generierten Elementen unterscheiden zu können, wird die Nutzer-Zeitzone kursiv geschrieben. Auch hier wird die Liste weiter eingeschränkt und es werden lediglich noch zwei Beispiele betrachtet.

- | | |
|--------------------------------------|---|
| 1. England West+Sussex | 11. California |
| 2. England | 12. Dimension |
| 3. West+Sussex | 13. 11th California Dimension <i>Pacific+Time+US+Canada</i> |
| 4. England West+Sussex <i>London</i> | 14. 11th California <i>Pacific+Time+US+Canada</i> |
| 5. England <i>London</i> | 15. California Dimension <i>Pacific+Time+US+Canada</i> |
| 6. West+Sussex <i>London</i> | 16. 11th <i>Pacific+Time+US+Canada</i> |
| 7. 11th California Dimension | 17. California <i>Pacific+Time+US+Canada</i> |
| 8. 11th California | 18. Dimension <i>Pacific+Time+US+Canada</i> |
| 9. California Dimension | |
| 10. 11th | |

19. York

21. York

20. York *London*

22. York *Eastern+Time*

Mit Hilfe der Zeitzone können nun auch die beiden letzten Einträge unterschieden werden. Zum einen “York“ in England zum anderen “York“ in den USA. Dies löst das Problem der Mehrdeutigkeiten von Toponymen. Mit Hilfe der zusätzlichen Zeitzone können geografische Objekte, zumindest wenn sie in zwei verschiedenen Zeitzonen liegen, unterschieden werden.

Fazit

Nach der Vorverarbeitung liegen eine Menge von potenziellen geografischen Indikatoren mit zugehörigen geografischen Koordinaten vor. Bei der Vorverarbeitung wurden einige Probleme des Nutzer-Standortes und bei der Verwendung von Toponymen beseitigt.

Die so erzeugten potenziellen geografischen Indikatoren können nun in der Georeferenz-Basis als Referenzwerte gespeichert werden. Mit der so erzeugten Georeferenz-Basis ist es grundsätzlich möglich eine Georeferenzierung durchzuführen. Allerdings wird jedem potenziellen geografischen Indikator eine Georeferenz zugeordnet. Das ist grundsätzlich problematisch, da Referenzwerten die keinen geografischen Bezug aufweisen eine Georeferenz zugeordnet wird. Dies kann bei der Georeferenzierung zu Fehlern führen. Es sollte entschieden werden können ob der Referenzwert einen geografischen Bezug hat oder nicht.

Des weiteren können zu einem Referenzwert beliebig viele Datensätze existieren, selbst wenn diese auf denselben Ort verweisen. Bei einer Abfrage an die Datenbank werden potenziell große Mengen an Datensätzen zurückgeliefert die alle auf dieselbe Georeferenz verweisen.

4.7.2 Absoluten Häufigkeiten

Es sollen nun absolute Häufigkeiten in der Georeferenz-Basis eingeführt werden. Die absoluten Häufigkeiten sollen angeben wie oft eine Kombination aus Referenzwert und Georeferenz in der Georeferenz-Basis vorhanden ist. Dadurch werden Duplikate in der

Georeferenz-Basis vermieden. Zusätzlich kann ermittelt werden ob ein Referenzwert an einer bestimmten Position gehäuft auftritt.

Beim abspeichern eines neuen Datensatzes soll nun zunächst geprüft werden ob dieser bereits in der Georeferenz-Basis vorhanden ist. Ist dies der Fall, so wird die absolute Häufigkeit des entsprechenden Eintrags um 1 erhöht. Ist der Datensatz noch nicht gespeichert so wird ein neuer Datensatz angelegt und die absolute Häufigkeit mit 1 initialisiert.

Die absolute Häufigkeit misst damit wie oft ein Referenzwert an einer geografischen Position vorkommt. Es ist zu erwarten, dass ein Referenzwert mit geografischem Bezug gehäuft an einer bestimmten geografischen Position oder Region auftritt. Während Referenzwerte die keinen geografischen Bezug haben sehr verteilt oder nur sehr selten auftreten.

Betrachtet man die Tweet-Sammlung kann beim Vergleich der Georeferenzen ein Problem entstehen. Die Georeferenz aus der Tweet-Sammlung besteht aus den geografischen Koordinaten des zugehörigen Tweets. Die geografische Position wird zumeist mit Hilfe von GPS-Modulen mobiler Endgeräte wie Smartphones bestimmt. Diese können eine Position oft auf wenige Meter genau bestimmen. Das bedeutet, zwei Tweets die wenige Meter voneinander abgesetzt wurden haben unter Umständen unterschiedliche Werte für den Längen- und Breitengrad. Dies kann für die Bestimmung der Häufigkeiten problematisch sein, da die Werte in der Regel nicht exakt übereinstimmen. Um dies zu umgehen sollen die geografischen Koordinaten auf Städte abgebildet werden.

4.7.3 Vorverarbeitung der geografischen Koordinaten

Die geografischen Koordinaten sollen auf Städte aufgelöst werden. Dies entspricht der untersten Ebene der geografischen Hierarchie. Insbesondere werden damit auch alle anderen geografischen Hierarchieebenen zu einem Referenzwert bestimmt.

Jeder geografischen Koordinate soll die am nächsten gelegene Stadt zugeordnet werden. Als Ergebnis liegt nun pro Referenzwert statt einer geografischen Koordinate eine Stadt vor. Dies kann als Übergang einer kontinuierlichen Darstellung durch geografische Koordinaten zu einer diskreten Darstellung durch Städte angesehen werden.

Durch die Nearest-Neighbour Zuordnung zu Städten wird der Globus in Voronoi-Regionen eingeteilt. Die Punkte zur Erzeugung der Voronoi-Regionen bilden dabei die Städte. Jeder Voronoi-Region kann also eine Stadt zugewiesen werden. Wird ein Tweet innerhalb einer Voronoi-Region abgesetzt wird er der entsprechenden Stadt zugeordnet.

In Abbildung ?? ist die Zuordnung schematisch dargestellt.

Es bleibt zu definieren was eine Stadt ist. In dicht besiedelten Gebieten können viele kleine Städte und Dörfer vorhanden sein. Damit wäre die Positionsangabe wiederum zu genau. Deshalb soll die Definition einer Stadt hier über die Einwohnerzahl stattfinden. Es wird festgelegt, dass die Einwohnerzahl über 15000 liegen muss um als Stadt zu zählen. Der Tweet wird also der nächstgelegenen Stadt mit mehr als 15000 Einwohnern zugeordnet. In Ballungsräumen sind damit mehr Städte zu erwarten, womit die Genauigkeit zunimmt. Es ist allerdings auch zu erwarten, dass dort tendenziell mehr Tweets abgesetzt werden als in ländlichen Gebieten. Womit in jeder Voronoi-Region zu einer Stadt ausreichend Tweets vorhanden sind. Damit können potenziell mehr Tweets zu einer Stadt zugeordnet werden. Wohingegen in ländlichen Gebieten ein größeres Einzugsgebiet pro Stadt zustande kommt. Es sind allerdings auch weniger Tweets zu erwarten. Durch das größere Einzugsgebiet werden dennoch ausreichend viele Tweets auf einer Stadt zugeordnet.

Die Distanzen, welche zwischen der tatsächlichen Position des Tweets und der zugeordneten Stadt liegen wurden protokolliert. Dies spiegelt den Fehler der bei einer solchen Zuordnung entsteht wieder. In Tabelle 4.5 sind die Ergebnisse dargestellt.

Tabelle 4.5: Fehlerdistanzen zwischen Tweet Ursprung und zugeordneter Stadt (in km)

Durchschnitt	7
Median	3.5
0.25 Quantil	1.7
0.75 Quantil	6.9
0.85 Quantil	10
0.95 Quantil	24.1
0.98 Quantil	44.2
Größte Distanz	3424.5
Kleinste Distanz	0

Im Median liegt die Fehlerdistanz zwischen der tatsächlichen Position und der zugeordneten Stadt bei 3,5 Kilometern. Über die Quantile können die Fehlerdistanzen noch

genauer untersucht werden. Das 0.25 Quantil sagt aus, dass 25% aller Fehlerdistanzen unter 1,7 Kilometern liegen. 95% der fehlerdistanzen liegen unter 24,1 Kilometer. Dies ist ausreichend genau.

Die größte gemessene Distanz sind 3424,5 Kilometer. Die geringste 0 Kilometer.

nearestNeighbour
with Voronoi-
Region and
Zuordnung
ländlich, bal-
lungszentrum

4.7.4 Erweiterte Struktur der Georeferenz-Basis

Hier soll nun die erweiterte Struktur der Georeferenz-Basis vorgestellt werden. Mit der absoluten Häufigkeiten ist ein neuer Wert pro Datensatz hinzugekommen. Des weiteren wurde die Georeferenz auf eine Stadt aufgelöst und damit implizit die Verwaltungseinheit erster Ordnung (Adm1), die Verwaltungseinheit zweiter Ordnung (Adm2) und das Land bestimmt. Zu einem Referenzwert sollen neben der Stadt zudem die anderen Ebenen der geografischen Hierarchie abgelegt werden. Die daraus resultierende Struktur der Georeferenz-Basis ist in 4.6 inklusive einiger Beispieleinträge dargestellt.

Tabelle 4.6: Struktur der Georeferenz-Basis mit geografischer Hierarchie

Referenzwert	Stadt	Adm2	Adm1	Land	Häufigkeit
Los+Angeles USA	Los Angeles	LA County	CA	USA	30
Los+Angeles USA	San Francisco	SF County	CA	USA	3
Los+Angeles	Los Angeles	LA County	CA	USA	70
USA	Los Angeles	LA County	CA	USA	80
Heilbronn	Heilbronn	Regierungsbezirk Stuttgart	BaWü	BRD	90

4.7.5 Überblick

Hier soll nun ein Überblick über das gesamte Verfahren zum einlernen der Georeferenz-Basis gegeben werden.

Zunächst werden die einzelnen Vorverarbeitungsschritte für den Nutzer-Standort und die Nutzer-Zeitzone durchgeführt. Parallel kann der Längen- und Breitengrad auf eine Stadt aufgelöst werden. Die dadurch entstandenen potenziellen geografischen Indikatoren und die zugehörige Georeferenz wird nun in die Georeferenz-Basis gespeichert. Dabei wird überprüft ob dieser Datensatz bereits vorhanden ist. Ist dies der Fall, wird der Häufigkeitwert des entsprechenden Datensatzes inkrementiert. Ist der Datensatz noch nicht vorhanden wird er angelegt und die absolute Häufigkeiten mit 1 initialisiert.

Die Vorverarbeitung extrahiert dabei zusätzliche Informationen aus dem Nutzer-Standort. Durch die Bereinigung der Werte im Nutzer-Standort, die alphanumerische Sortierung, das identifizieren von Toponymen mit mehreren Worten und die darauffolgende Erzeugung von N-Grammen werden zusätzliche Informationen aus jedem Nutzer-Standort gewonnen. Die Nutzer-Zeitzone wird einbezogen um Doppeldeutigkeiten auflösen zu können.

Durch die neue Datenstruktur, mit den absoluten Häufigkeiten und der abgebildeten geografischen Hierarchie, lassen sich nun tiefergehende Analysen durchführen um eine robuste Georeferenzierung zu ermöglichen. Die absoluten Häufigkeiten geben dabei an wie oft ein Referenzwert in einer bestimmten Region vorkommt.

Durch dieses Verfahren lässt sich eine Datenbasis erzeugen die domänenspezifische Eigenheiten, in Bezug auf die Verwendung spezieller Begriffe oder Formulierungen, berücksichtigt. Des weiteren werden Toponyme, die in Ortsverzeichnissen unter Umständen nicht hinterlegt sind, berücksichtigt. Auch geografische Indikatoren mit mittelbarem geografischen Bezug, zum Beispiel die Verwendung spezieller Begriffe in einer geografischen Region, können einbezogen werden.

4.8 Geografischer Bezug der Referenzwerte

Die eingelernten Referenzwerte beinhalten alle Werte aus den Nutzer-Standorten der Tweet-Sammlung. Es sind also auch Referenzwerte vorhanden die keinen geografischen Bezug haben. Es ist die Frage zu beantworten: Wie kann bestimmt werden ob ein Referenzwert geografischen Bezug hat oder nicht? Oder: Wie kann vermieden werden, dass ein Referenzwert, der keinen geografischen Bezug hat, zur Georeferenzierung genutzt wird?

Dies ist wichtig, denn durch die Referenzwerte wird in der eigentlichen Georeferenzierung einem geografischen Indikator eine Georeferenz zugewiesen. Wird einem geografischen Indikator durch einen Referenzwert ohne geografischen Bezug eine Georeferenz zugewiesen ist diese mit hoher Wahrscheinlichkeit fehlerhaft. Denn der Referenzwert kann nicht mit einem geografischen Objekt in Bezug gebracht werden. Dies wiederum führt zu schlechten und unzuverlässigen Ergebnissen. Es muss also ein Verfahren gefunden werden um zu entscheiden ob die Referenzwerte einen geografischen Bezug haben oder nicht.

Es ist zu beachten, dass die Vorverarbeitung keine Aussage zum geografischen Bezug macht, sondern vielmehr die Referenzwerte aus den Nutzer-Standorten extrahiert. Dies soll sicherstellen, dass möglichst viele Informationen aus den Nutzer-Standorten gezogen werden können und insbesondere keine Informationen verloren gehen.

Um zu entscheiden ob ein Referenzwert geografischen Bezug hat oder nicht wird die absolute Häufigkeit verwendet. Die absoluten Häufigkeiten geben an wie oft ein Referenzwert in einer bestimmten Region, zunächst in der Voronoi-Region der entsprechenden Stadt, vorkommt. Daraus kann nun ein geografischer Bezug abgeleitet werden.

4.8.1 Die absolute Häufigkeit als Hinweis auf geografischen Bezug zu Städten

Eine hohe Häufigkeitswert kann ein Hinweis auf den geografischen Bezug eines Referenzwertes darstellen. Aufgrund der Eigenschaften des Nutzer-Standorts ist anzunehmen, dass in der Voronoi-Region einer Stadt der Name der zugehörigen Stadt häufig vorkommt. Dadurch kann eine Relevanz des Referenzwertes abgeleitet werden. Tritt der Referenzwert nicht häufig auf, so ist er von nur wenigen Nutzern als Nutzer-Standort in einer Stadt angegeben worden und somit für die Stadt nicht relevant. In Abbildung ?? sind die Tweets in denen "Istanbul" im Nutzer-Standort vorkommt aufgetragen. Es ist deutlich eine Häufung um die Stadt Istanbul zu erkennen. Werden die Tweets rund um Istanbul nun auf die Stadt Istanbul abgebildet, wird die Kombination aus dem Referenzwert "Istanbul" und der Georeferenz Istanbul sehr häufig vorkommen und dementsprechend einen hohen Häufigkeitswert aufweisen. In den Nutzer-Standorten der Tweets rund um Istanbul taucht "Istanbul" tatsächlich 972 mal auf. Damit kann eine Gewisse Relevanz für den Referenzwert "Istanbul" abgeleitet werden. Es kann also ein Schwellwert für die Häufigkeit eingeführt werden um Referenzwerte mit geografischem Bezug zu identifizieren.

Allerdings garantiert die Höhe des Häufigkeitswertes noch nicht, dass ein Referenzwert einen geografischen Bezug hat. Es können auch Werte an einem bestimmten Ort häufig vorkommen, die keinen geografischen Bezug haben. Um dies erkennen zu können muss ein weiterer Wert berechnet werden.

4.8.2 Relative Häufigkeiten als Hinweis auf geografischen Bezug zu Städten

Betrachtet man die Häufigkeitswerte isoliert voneinander wird die Verteilung des Referenzwertes außer acht gelassen. Ein Referenzwert kann eine gleichmäßige Verteilung über mehrere Städte aufweisen. Das bedeutet, er wird in vielen unterschiedlichen Städten benutzt. Es ist aber trotzdem nicht ausgeschlossen, dass in bestimmten Städten Häufungen auftreten, die über einem gewählten Schwellwert für den Häufigkeitswert liegen. Betrachtet man jedoch das globale auftreten kann diese Häufung, relativ zu den anderen Häufigkeitswerten, sehr gering sein. Ist der Häufigkeitswert reativ gesehen eher gering, kann dies ein Hinweis darauf sein, dass der Referenzwert keinen geografischen Bezug hat. Denn dies besagt, dass der Referenzwert verteilt in mehreren Städten genutzt wird. Es ist also wichtig nicht nur die Höhe des Häufigkeitswertes, sondern auch die relative Häufigkeit des Referenzwertes zu berücksichtigen.

Berechnung der relativen Häufigkeiten

Um die relativen Häufigkeiten zu berechnen soll das Vorkommen eines Referenzwertes in einer Stadt, durch die Gesamtanzahl der Vorkommen des Referenzwertes geteilt werden. Damit erhält man den prozentualen Anteil der auf eine Stadt entfallenden Vorkommen eines Referenzwertes. Dies kann mit Hilfe der Häufigkeitswerte umgesetzt werden.

Sei (r_i, c_j) ein Datensatz der Georeferenz-Basis mit Referenzwert r_i und Georeferenz c_i . Des weiteren liefert $H(r_i, c_j)$ den Häufigkeitswert zu einem Referenzwert r_i und einer Georeferenz c_i .

Damit kann die relative Häufigkeit $rel_{(r_i, c_j)}$ für jede Kombination (r_i, c_j) durch die folgende Formel berechnet werden. n_c ist dabei die Anzahl aller Georeferenzen.

$$d_{r_i, c_j} = \frac{H(r_i, c_j)}{\sum_{j=0}^{n_c} H(r_i, c_j)} \quad (4.1)$$

Beispiel

“La+Plata“ tritt rund um die Stadt La Plata in Argentinien 91 mal auf. Der Referenzwert “La+Plata“ hat offensichtlich einen geografischen Bezug zu einer Stadt, denn er ist ein Stadt-Name und stellt deshalb ein Toponym dar.

Obwohl “the“ keinen offensichtlichen geografischen Bezug ist der Häufigkeitswert von 91 vorkommen in Jakarta hoch. Der Schwellwert könnte nun aufgrund der Erfahrung mit dem Referenzwert “La+Plata“ auf 90 angesetzt werden. Dann würde davon ausgegangen werden, dass der Referenzwert “the“ einen geografischen Bezug hat. Betrachtet man allerdings die Abbildung ?? fällt auf, dass Tweets mit dem Wert “the“ im Nutzer-Standort sehr verteilt auf dem Globus auftreten. Im Gegensatz dazu tritt “La+Plata“ in den Nutzer-Standorten sehr konzentriert auf. In Abbildung ?? wird die Verteilung von Tweets deren Verfasser “La+Plata“ im Nutzer-Standort enthalten dargestellt. Es ist deutlich eine Häufung um die Stadt La Plata in Argentinien zu erkennen, weltweit tritt der Referenzwert aber sehr selten auf.

Berechnet man nun die relativen Häufigkeiten kann dies abgebildet werden. In Tabelle 4.11 sind die zugeordneten Städte, der Häufigkeitswert und die berechneten relativen Häufigkeiten für die Vorkommen des Wortes “the“ aufgetragen. Die Einträge sind absteigend nach dem Wert der relativen Häufigkeit sortiert und auf 4 Einträge gekürzt. Insgesamt kam das Wort “the“ in 2824 verschiedenen Städten vor. Dabei wurde es insgesamt 5764 mal verwendet. Trotz des hohen Häufigkeitswertes in Jakarta liegt die relative Häufigkeit bei lediglich 1,6 %. Aus den relativen Häufigkeiten lässt sich nun die, durch Abbildung ?? vermutete, globale Verteilung ablesen.

Tabelle 4.7: “the“

Stadt	Häufigkeitswert	relative Häufigkeit in %
Jakarta	91	1,6
Singapore	27	0,5
Bekasi	25	0,4
Philadelphia	23	0,4

Die Verteilung erklärt sich dadurch, dass das Wort “the“ im englischen sehr häufig auftritt. Englisch ist die internationale Verkehrssprache und wird dementsprechend global und sehr häufig verwendet. Dies spiegelt sich in der Verteilung der vorkommen auf dem

Globus wieder. Die Häufung um Jakarta kann teilweise damit erklärt werden, dass sehr viele Tweets die in und um Jakarta abgesetzt werden eine Angabe von Längen- und Breitengrad aufweisen.

In Tabelle 4.8 wird dieselbe Auswertung für “La+Plata“ dargestellt. “La+Plata“ ist insgesamt 129 mal in 23 verschiedenen Städten aufgetaucht. In der Stadt La Plata, in Argentinien, kam es 91 mal vor. Dies entspricht einer relativen Häufigkeit von 70,5%.

Tabelle 4.8: “La Plata“

Stadt	Häufigkeitswert	Relative Häufigkeit in %
La Plata	91	70,5
Villa Gesell	9	7,0
Mar del Plata	5	3,9
Quilmes	3	2,3

“La+Plata“ taucht Erwartungsgemäß am häufigsten rund um die Stadt La Plata auf, da es der Name dieser Stadt ist.

Es lässt sich also anhand der relativen Häufigkeiten ein geografischer Bezug des Referenzwertes nachweisen. Damit kann man zunächst bestimmen ob ein Referenzwert geografischen Bezug hat oder nicht. Die relativen Häufigkeiten können nach dem einlernen der Referenzwerte berechnet und in der Georeferenz-Basis hinterlegt werden. Mit Hilfe eines Schwellwertes für die relativen Häufigkeiten, kann nun bestimmt werden wann ein Referenzwert geografischen Bezug hat.

Aber auch die ausschließliche Betrachtung der relativen Häufigkeiten reicht nicht aus um die geografische Relevanz nachzuweisen. Da die relativen Häufigkeiten basierend auf den Vorkommen des Referenzwertes berechnet werden sagen diese nichts über die Häufigkeit aus. Kommt ein Referenzwert zwei mal in zwei verschiedenen Städten vor, liegen die jeweiligen relativen Häufigkeiten bei 50%. Dies ist ein sehr hoher Wert, doch da der Referenzwert nur einmal vorkam ist es sehr unwahrscheinlich, dass er einen geografischen Bezug hat.

4.8.3 Geografischer Bezug der Referenzwerte zu Verwaltungseinheiten und Ländern

Beim einlernen der Georeferenz-Basis werden im Häufigkeitswert die Vorkommen pro Stadt gespeichert. Mit den daraus errechneten relativen Häufigkeiten kann nicht entschieden werden, ob für den Referenzwert eine globale Verteilung vorliegt oder ob der Referenzwert unter Umständen nur regional begrenzt, zum Beispiel in einem Land, auftritt. Für Referenzwerte die ein Land oder eine Verwaltungseinheit bezeichnen ist auf Städteebene also eine geringe relative Häufigkeit zu erwarten. Diese Referenzwerte treten in einer größeren geografischen Region auf und sind somit über mehrere Städte verteilt. Es ist zu erwarten, dass ein Ländername in den Nutzer-Standorten von Tweets aus dem gesamten Land auftritt. Durch die Unterteilung des Landes in Stadtgebiete wird der Wert sehr verteilt auf die Städte des Landes auftreten.

Soll nun statt einer Stadt eine Verwaltungseinheit oder das Land als Georeferenz bestimmt werden, kann mit diesen Häufigkeitswerten keine Aussage über den geografischen Bezug gemacht werden.

Am folgenden Beispiel soll dieser Sachverhalt erläutert werden.

In Abbildung ?? sind in grün Tweets dargestellt, welche im Nutzer-Standort “Wales“ enthalten. Wales entspricht einer Verwaltungseinheit erster Ordnung und gehört zu Grossbritannien. Die Tweets sind in der gesamten geografischen Region, über die sich Wales erstreckt, verteilt. Außerhalb von Wales tritt “Wales“ im Nutzer-Standort sehr selten auf. Die relativen Häufigkeiten in Tabelle 4.9 bestätigen eine Verteilung über eine größere geografische Region. Anhand der relativen Häufigkeiten kann aber nicht entschieden werden ob der Referenzwert global verteilt ist, oder in einer bestimmten geografischen Region, wie einem Land, auftritt. “Wales“ taucht in insgesamt 78 Städten 346 mal in Nutzer-Standorten auf.

Die relative Häufigkeit von 12,7% weist eher darauf hin, dass “Wales“ keinen geografischen Bezug aufweist. Auf Städteebene ist dies auch durchaus korrekt. Allerdings kann aus diesen Ergebnissen kein geografischer Bezug zu einer der anderen geografischen Hierarchieebenen abgeleitet werden. Das Problem ist, dass Wales keinen geografischen Bezug zu einer Stadt aufweist, wohl aber zu einer Verwaltungseinheit erster Ordnung und daher zu einer geografischen Region.

Tabelle 4.9: “Wales“

Stadt	Häufigkeitswert	Relative Häufigkeit in %
Cardiff	44	12,7
Newport	32	9,2
Carmarthen	24	6,9
Swansea	18	5,2
...

Um dieses Problem lösen zu können müssen zu einem Referenzwert die relativen Häufigkeiten für die anderen geografischen Hierarchieebenen berechnet werden. Damit kann dann die Verteilung der Referenzwerte auf diese Hierarchieebenen betrachtet werden.

Berechnung der relativen Häufigkeiten zu Verwaltungseinheiten und Ländern

Da zu jedem Referenzwert die zugehörigen Verwaltungseinheiten und Länder bekannt sind können die Häufigkeitswerte und relativen Häufigkeiten direkt aus der Georeferenz-Basis berechnet werden.

Es müssen lediglich die Häufigkeitswerte und relativen Häufigkeiten aufsummiert werden, bei denen der Wert der geografischen Hierarchieebenen übereinstimmen. Im Beispiel aus Tabelle 4.9 müssen alle Häufigkeitswerte und relativen Häufigkeiten derjenigen Städte aufsummiert werden, die in derselben Verwaltungseinheit erster Ordnung liegen. In ?? ist dies an einigen Einträgen für den Referenzwert “Wales“ exemplarisch dargestellt. Betrachtet man die Verwaltungseinheiten zu allen Städten in denen “Wales“ im Nutzer Standort vorkommt ergibt sich folgende Liste.

1. Wales 35
2. England 30
3. unterschiedliche Verwaltungseinheiten 13

Aus dieser Betrachtung alleine lässt sich noch nicht entscheiden ob der Referenzwert “Wales“ einen geografischen Bezug zu einer Verwaltungseinheit hat. Denn der Referenzwert kam sowohl in 30 Städten in England als auch in 30 Städten in Wales vor, was

keinen signifikanten Unterschied darstellt. Summiert man allerdings die Vorkommen und relativen Häufigkeiten pro Stadt auf ergibt sich daraus Tabelle 4.10.

Tabelle 4.10: “Wales“

Adm1	Häufigkeitswert	Relative Häufigkeit in %
Wales	298	86,1
England	38	11,0
National Capital Region	1	0,3
Stockholm	1	0,3
...

Die relative Häufigkeit von 86,1% weist nun deutlich darauf hin, dass der Referenzwert einen geografischen Bezug zu Wales hat. Mit diesem Vorgehen, kann der geografische Bezug eines Referenzwertes auf jeder der geografischen Hierarchieebenen untersucht werden.

Analog können die Werte für die Verwaltungseinheit zweiter Ordnung und dem Land berechnet werden. Dieses Vorgehen ermöglicht es einen geografischen Bezug eines Referenzwertes auf jeder der geografischen Hierarchieebenen zu prüfen.

Fazit

Mit der Höhe des Häufigkeitswertes besteht ein erster Hinweis darauf ob ein Referenzwert geografischen Bezug hat oder nicht. Die alleinige Betrachtung des Häufigkeitswertes lässt aber die Verteilung der Werte auf den jeweiligen geografischen Hierarchieebenen außer betracht. Mit den berechneten relativen Häufigkeiten können die Referenzwerte zusätzlich auf ihre Verteilung untersucht werden. Die relativen Häufigkeiten können nach dem Einlernen berechnet und in der Georeferenz-Basis gespeichert werden. Während der Georeferenzierung können diese Werte dann genutzt werden um die geografischen Indikatoren zu bestimmen.

4.9 Verfahren zur Georeferenzierung am Beispiel von Twitter

Im vorhergehenden Abschnitt wurde ein Verfahren zum einlernen der Georeferenz-Basis vorgestellt. Hier soll nun das Verfahren zur Georeferenzierung eines Twitter-Nutzers vorgestellt werden. Dabei dient die Georeferenz-Basis und die in ihr abgelegten Werte als Basis für die Zuweisung einer Georeferenz.

Zunächst werden aus dem Nutzer-Standort und der Nutzer-Zeitzone potenzielle geografische Indikatoren erzeugt. Dies geschieht analog zur Erzeugung von Referenzwerten beim einlernen der Georeferenz-Basis. Es werden also dieselben Vorverarbeitungsschritte für den Nutzer-Standort und die Nutzer-Zeitzone durchgeführt. Daraus resultiert eine Menge potenzieller geografischer Indikatoren. Die Werte der potenziellen geografischen Indikatoren werden nun in der Georeferenz-Basis nachgeschlagen. Dabei werden alle Datensätze deren Referenzwerte mit den potenziellen geografischen Indikatoren korrespondieren zurückgegeben. Auf diesen Datensätzen erfolgt die weitere Verarbeitung und Bestimmung der wahrscheinlichsten Georeferenz.

Es liegt nun eine Menge an Datensätzen aus der Georeferenz-Basis vor. Die Referenzwerte entsprechen dabei den potenziellen geografischen Indikatoren. Die relativen Häufigkeiten und die Häufigkeitswerte werden für jeden Referenzwert separat analysiert. Das Ziel der Analyse ist es, die wahrscheinlichste Georeferenz zu ermitteln.

In Abbildung ?? ist der gesamte Ablauf an einem Beispiel dargestellt. In den folgenden Abschnitten soll nun die Analyse genauer betrachtet werden.

4.9.1 Analyse

Dies Analyse beinhaltet zwei Schritte. Zuerst müssen diejenigen Referenzwerte gewählt werden, welche am wahrscheinlichsten einen geografischen Bezug haben. Dabei wird jeder Referenzwert separat betrachtet. In einem nächsten Schritt wird der Referenzwert gewählt der unter den verbliebenen am wahrscheinlichsten die geografische Position des Nutzers beschreibt.

Auswahl der Referenzwerte mit geografischem Bezug

Es können pro Referenzwert zunächst mehrere Datensätze vorliegen. Aus diesen sollen diejenigen gewählt werden, welche am wahrscheinlichsten einen geografischen Bezug aufweisen. Dazu werden die relativen Häufigkeiten und die Häufigkeitswerte genutzt. Für jeden Referenzwert wird derjenige Datensatz gewählt, der den größten relativen Häufigkeitswert h_{rel} über einem Schwellwert s_{rel} aufweist. Mit diesem Schwellwert lässt sich bestimmen wie verteilt der Referenzwert auftreten kann. Zusätzlich wird geprüft ob der Häufigkeitswert ebenfalls über einem Schwellwert s_{abs} liegt. Mit diesem Schwellwert lässt sich bestimmen wie häufig der Referenzwert an einer geografischen Position oder in einer geografischen Region auftreten muss.

Bestimmung der wahrscheinlichsten Georeferenz

Nun liegt wiederum eine Menge an Datensätzen vor. Jeder Referenzwert, und damit auch jeder potenzielle geografische Indikator, taucht nur noch ein mal auf.

Aus den verbliebenen Datensätzen soll nun die Georeferenz gewählt werden. Dazu werden die relativen Häufigkeiten verglichen. Es wird der Datensatz mit der höchsten relativen Häufigkeit gewählt. Die Georeferenz dieses Datensatzes wird dann dem Twitter-Nutzer zugewiesen. Damit wird der Referenzwert ausgewählt der die größte relative Häufigkeit aller untersuchten Referenzwerte aufweist.

Die Referenzwerte stellen NGramme dar, wie in der Vorverarbeitung in Unterkapitel 4.7.1 erläutert wird. Es werden hier also insbesondere auch Uni-, Bi- und Trigramme miteinander verglichen. Darauf soll nun eingegangen werden.

Vergleich der relativen Häufigkeiten zu Uni- Bi- und Trigrammen Jedes Element eines Bi- oder Trigrammes kann potenziell einen geografischen Bezug haben. Umso mehr Elemente ein NGramm beinhaltet umso spezieller kann die Beschreibung des geografischen Objekts sein. Deshalb können NGramme mit einem höheren Grad ein Objekt genauer beschreiben als NGramme mit einem niedrigeren Grad.

Allerdings können NGramme mit einem höheren Grad auch eine schlechtere Beschreibung darstellen. Beispielsweise wenn das zusätzliche Element keinen geografsichen Bezug hat.

Bei NGrammen mit einem Grad größer zwei können also zwei Fälle unterschieden werden.

1. Die Kombination aus den Elementen des NGrammes beschreibt einen Ort genauer
2. Die Kombination aus den Elementen des NGrammes beschreibt einen Ort nicht genauer

Fall1 Ein Beispiel für den ersten Fall ist der Nutzer-Standort “york“ mit der Nutzer-Zeitzone “eastern+time+us+canada“. Durch die Vorverarbeitung werden folgende potenzielle geografische Indikatoren erzeugt.

1. york
2. york *eastern+time+us+canada*

Eine Stadt Namens York existiert sowohl in Grossbritannien als auch in den USA. Fragt man nun die beiden Referenzwerte in der Georeferenz-Basis ab erhält man folgende Werte:

Tabelle 4.11: “the“

Referenzwert relative Häufigkeit in %	Stadt	Häufigkeitswert
York 48,3	York (GB)	97
york eastern+time+us+canada 63,2	York (US)	12

Die realtive Häufigkeit für york in Kombination mit der Zeitzone ist höher. Die Zeitzone gibt zusätzliche Auskunft darüber welches York gemeint ist. Die Kombination ist spezieller, kommt deshalb seltener vor und potenziell eher dort wo sie zutrifft. In diesem Fall in York in den USA.

In den meisten Fällen beschreibt einer der beiden Indikatoren eine größere geografische Region wie beispielsweise einen Bundesstaat der USA. Wird ein weiterer Wert, beispielsweise ein Städtenamen hinzugenommen, wird die Angabe des Ortes genauer. Die Wahrscheinlichkeit, dass diese Kombination ausserhalb des Ortes auftritt wird geringer.

Fall 2 Hier können wiederum 2 Fälle unterschieden werden.

1. Beide Elemente beziehen sich auf unterschiedliche geografische Objekte
2. Nur ein Element hat geografischen Bezug das andere nicht

Wenn zu einem Referenzwert mit geografischem Bezug ein Element hinzugefügt wird, welches keinen geografischen Bezug hat, beschreibt dies den Ort nicht genauer. Es ist zu erwarten, dass die Kombination der Elemente sehr selten vorkommt oder sehr verteilt ist. Ist die Kombination verteilter, so ist der relative Wert geringer als der des einzelnen Referenzwertes mit geografischem Bezug. Ist die Kombination seltener ist der Referenzwert bereits durch den Schwellwert s_{abs} aussortiert.

Wenn zu einem Referenzwert mit geografischem Bezug ein Element hinzugefügt wird, welches zwar geografischen Bezug hat, aber dieses sich auf ein anderes geografisches Objekt bezieht ist dasselbe Verhalten zu erwarten.

Wahl der Schwellwerte

Die Wahl der beiden Schwellwerte ist abhängig von den Anforderungen. Dabei ist die gewünschte geografische Hierarchie der zurückgegeben Georeferenz ein Faktor. Und die gewünschte Genauigkeit und Trefferquote.

Gewünschte Hierarchieebene der Georeferenz Umso größer die betrachtete geografische Region ist, umso weniger Möglichkeiten zur Einteilung gibt es. Auf Städteebene gibt es 23322 verschiedene geografische Regionen. Für jede Stadt mit mehr als 15000 Einwohnern existiert dabei eine Region. Diese Städte verteilen sich auf 234 verschiedene Länder. Dieselbe Menge an Referenzwerten, verteilt sich auf Länderebene also auf weniger geografische Regionen. Dadurch werden die Werte der relativen Häufigkeit und der

absoluten Häufigkeit insgesamt größer. Relativ zueinander werden allerdings nach wie vor Referenzwerte mit geografischem Bezug größer sein als Referenzwerte ohne geografischen Bezug.

Es müssen also für jeder geografsiche Hierarchieebene geeignete Schwellwerte s_{rel} und s_{abs} gefunden werden.

Genauigkeit und Trefferquote Der zweite Faktor ist die gewünschte Trefferquote und die Genauigkeit.

Umso niedriger der Schwellwert s_{rel} ist, umso größer wird die Wahrscheinlichkeit Referenzwerte zu wählen die keinen geografischen Bezug haben. Daraus resultieren mehr fehlerhafte Zuordnungen einer Georeferenz. Wodurch die Genauigkeit schlechter wird. Dadurch können allerdings mehr Georeferenzen zugeordnet werden, wodurch die Trefferquote verbessert wird. Umso höher der Schwellwert s_{rel} gewählt wird umso mehr Referenzwerte mit geografischem Bezug werden verworfen. Die Wahrscheinlichkeit, dass die gewählten Referenzwerte tatsächlich geografischen Bezug haben ist allerdings höher. Dadurch können weniger Georeferenzen zugeordnet werden. Somit sinkt die Trefferquote. Allerdings sind die zugewiesenen Georeferenzen sicherer womit die Genauigkeit steigt.

Der Schwellwert s_{abs} vermeidet, dass Referenzwerte gewählt werden die eine hohe realtive Häufigkeit aufweisen aber aufgrund ihrer geringen Vorkommen nicht relevant sind. Die Auswirkungen der Wahl des Schwellwertes s_{abs} verhalten sich Analog zum Schwellwert s_{rel} .

Fazit Die Wahl der Schwellwerte hängt zum einen von der Hierarchieebene und zum anderen von den Anforderungen an die Genauigkeit und die Trefferquote ab. In Bezug auf die geografischen Hierarchieebenen sind lediglich separate Schwellwerte für jede geografischen Hierarchieebenen zu bestimmen, da die relativen und absoluten Häufigkeiten sich insgesamt verändern. Bezüglich der Genauigkeit und der Trefferquote ist ein Kompromiss zwischen den beiden Werten einzugehen. Die Verbesserung der Trefferquote geht mit einer Verschlechterung der Genauigkeit einher und umgekehrt. Es kann also entweder ein Kompromiss gefunden werden der ein Optimum für beide Werte darstellt. Oder einer der Werte wird optimiert.

5 Implementierung

Im Rahmen dieser Diplomarbeit ist eine Referenzimplementierung des vorgestellten Verfahrens entstanden. In Auszügen soll die Referenzimplementierung hier vorgestellt werden. Hierbei sollen insbesondere Probleme bei der Umsetzung betrachtet werden, und wie diese gelöst wurden. Damit soll die Möglichkeit gegeben werden, in eigenen Implementierungen die Probleme frühzeitig zu erkennen und zu vermeiden. Des weiteren soll ein Überblick über die genutzten Datensätze und API's gegeben werden.

Datensätze in Grundlagen?

5.1 Komponenten der Referenzimplementierung

5.1.1 Architektur

Allgemeine Architektur der Referenzimplementierung

5.1.2 Präprozessorverarbeitung - Erzeugung der N-Gramme

Warum Präprozessoren -> schnelleres ändern der Vorverarbeitung.

5.2 Datenbank

5.2.1

Eventuell was über die Geo Indexe in der Datenbank und die Nearest Neighbour Berechnungen.

5.3 Geografie Daten

in Implemen-
tierung ver-
schieben

5.4 Data Sample

Beschreibung wie Daten erzeugt wurden, Zeiträume, Analysen

5.5 geonames.org

Allgemeines zu geonames.org, was ist geonames.org.

1. Woher stammen die Daten?
2. Umfang und Informationen
3. Aktualität
4. Hierarchiebeziehungen im geonames.org Datensatz

6 Leistungsbewertung

7 Schlussfolgerungen, Ausblick und Fragen

8 Zusammenfassung

9 Ideen und Notizen

9.1 Stakeholder analyse

Welche potenziellen Stakeholder profitieren von der Arbeit? Was benötigt jeder dieser Stakeholder? Bedürfnisse analysieren und Begründen.

1. Marketing Professionals
2. Statistiker allgemein
3. Sozialwissenschaftler -> Analyse von Informationsströmen

9.2 Fragen an Matthias

9.2.1 Strukturell

1. Soll ich noch auf die Messung eines Informationsflusses eingehen? Wenn ich keine Informationsflüsse untersuche hängt dieses Thema ein wenig in der Luft.
2. ???

9.2.2 Inhalt

9.3 Ideen

1. Voraussetzungen zur Anwendung des Verfahrens

In Einleitung

- a) Lerndaten mit konkreten geografischen Angaben
 - b) Indikatoren in Lerndaten, welche auch in Datensätzen ohne konkrete geografische Angaben vorkommen (hier eventuelle Diskrepanzen zwischen geogetaggtten und nicht geogetaggtten tweets + Mentalität in bestimmten Ländern)
 - c) Indikatoren mit geografischem Bezug, oder hinreichendem geografischen Bezug, Mittelbar oder unmittelbar
2. Auf Jargon Namen für Städte eingehen, wie bspsw. the big apple -> New York City
 3. Landesgrenzen-Problematik wird durch meine Lösung obsolet -> auf stakeholder eingehen
 4. Wahrscheinlichkeiten für korrekte Lokalisierung kann angegeben und justiert werden
 5. Wenn Wahrscheinlichkeiten auf best. Ebene nicht hoch genug dann verschieben auf Admin2 -> Admin1 -> Länderebene
 6. mit vorherigem werden Unsicherheiten bei Lokalisierung abgebildet (Wichtig für Informationsflüsse)
 - 7.

Korrelation
zwischen Lo-
kalisierung-
ungssicher-
heit und tat-
sächlichem
Match berech-
nen

9.4 Formulierungen

9.4.1 unmittelbare ungesicherte geografische Indikatoren

Das "userlocation" Feld in einem Tweet kann durchaus eine konkrete Lokation beinhalten, jedoch wird auch oft irgendetwas eingetragen. [HHSC11] Es kann sich dabei um beliebige Wörter oder Sätze handeln, die einzige Limitierung ist die Anzahl zur Verfügung stehender Zeichen. Nichtsdestotrotz ist es das Ziel dieses Feldes seinen eigenen Standort anzugeben. Dabei kann allerdings nicht davon ausgegangen werden, dass der eingetragene Wert nicht doch in einem Zusammenhang mit einer geografischen Lokation steht. Bezeichnungen von Städten in Umgangssprache wie beispielsweise "The Big Apple" für New York City oder Motown für Detroit, sind für einige Personen nicht unmittelbar

zuzuordnen, geben allerdings eine konkrete Lokation an. Da die Masse an Bei bzw. Spitznamen für Städte nicht überschaubar ist und auch sprachliche Probleme bestehen ist es sinnvoll alle userlocation Einträge gleich zu behandeln und diese in erster Linie als Lokationsangaben zu behandeln. Durch die Einschränkung auf eine Geolocation werden einzelne gleich lautende Einträge, welche aber nicht auf einen konkreten Ort hinweisen in einzelnen Datensätzen abgelegt.

9.5 Datenbasis

1. Welche Datenbasis wurde genutzt
 - a) Streaming API
 - b) Is the Sample good enough (Morstatter et al 13)
 - c) When is it biased? (Morstatter et al)
 - d) How does the Data sampling Startegy Impact the Discovery of Information Diffusion in Social Media (De Choudhurry, 1)
2. Lerndatensatz
3. Kontrolldatensatz
4. Manuell getaggter Datensatz
5. Google Maps getaggter Datensatz

9.6 Vorteile neuer Ansatz bei Mapping auf Geografische Daten

Notwendigkeit/Vorteile von Hierarchiebeziehungen im Mapping auf Geograohie Daten

Literaturverzeichnis

- [BNJ12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2012.
- [CCL10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 759, 2010.
- [EOSX10] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.
- [FVMF13a] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: social butterflies or frequent fliers? ... *conference on On-line social ...*, 2013.
- [FVMF13b] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: Social butterflies or frequent fliers? *CoRR*, abs/1310.2671, 2013.
- [GGMQ14] R Garcia-Gavilanes, Y Mejova, and D Quercia. Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication. ... *cooperative work & social ...*, pages 1511–1522, 2014.
- [Gol08] Daniel W Goldberg. *A Geocoding Best Practices Guide*. North American Association of Central Cancer Registries (NAACCR), 2008.
- [HGG12] S Hale, D Gaffney, and M Graham. Where in the world are you? geolocation and language identification in twitter. *Proceedings of ICWSM'12*, (2013), 2012.

- [HHSC11] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber’s heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’11, pages 237–246, New York, NY, USA, 2011. ACM.
- [IAB] Interactive Advertising Bureau IAB. Social media ad metrics definitions. Internet.
- [JJ21] G. Jellinek and W. Jellinek. *Allgemeine Staatslehre*. J. Springer, 1921.
- [KA08] Balachander Krishnamurthy and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks (WOSP ’08)*, pages 19–24, 2008.
- [KCLC13] Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW ’13, pages 667–678, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter , a social network or a news media? In *The International World Wide Web Conference Committee (IW3C2)*, pages 1–10, 2010.
- [MND12] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. *ICWSM*, pages 511–514, 2012.
- [MPLC13] Fred Morstatter, J Pfeffer, H Liu, and KM Carley. Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose. *Proceedings of ICWSM*, pages 400–408, 2013.
- [NP03] M E J Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 68:036122, 2003.
- [PCV13] Reid Friedhorsky, Aron Culotta, and Sara Y. Del Valle. Inferring the origin locations of tweets with quantitative confidence. *CoRR*, abs/1305.3932, 2013.

- [POM⁺13] S. Petrovic, M. Osborne, R. Mccreadie, C. Macdonald, and I. Ounis. Can twitter replace newswire for breaking news? In *ICWSM - 13*, 2013.
- [SHP⁺13] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. *Seventh International AAAI Conference on Weblogs and Social Media*, pages 573–582, 2013.
- [SKD11] Martin Szomszor, Patty Kostkova, and Ed De Quincey. #swineflu: Twitter predicts swine flu outbreak in 2009. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, volume 69 LNICST, pages 18–26, 2011.
- [SMvZ09] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 484, 2009.
- [SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.
- [ti13] twitter inc. Final initial public offering(ipo) prospectus, 11 2013.
- [TSSW11] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welp. Election forecasts with twitter: How 140 characters reflect the political landscape, 2011.