Karlsruhe Institute of Technology

Dezentrale Systeme und Netzdienste
Institut für Telematik

Lehrstuhl
Prof. Dr. Hannes Hartenstein

Fakultät für Informatik

Diplomarbeit
2014

Mein Titel

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

Todo list

1 Einleitung

1.1 Motivation und Hintergründe

Motivation aus Proposal (abändern).

1.2 Problembeschreibung

Lokalisierung von Tweets ohne konkrete geografische Angaben. Problem das sehr wenige Tweets Geotags haben.

2 Grundlagen und Stand der Technik

2.1 Geoinformationen in Social Media Daten

selbst definiert okay?

1. gesicherte Geoinformationen vs. ungesicherte Geoinformationen
2. konkrete Geolocations zu allgemeineren (Städte <—> Länder) Referenz: From Justin Biebers Heart
3. Lokalisierung von Social-Media Elementen (Videos, User, Nachrichten, Bilder) kleine Übersicht
4. Hinleitung zu Twitter

2.2 Twitter

Allgemeine Informationen zu Twitter.

1. Was ist Twitter -> Tweets/Mechanismen/"Wie wird Twitter genutzt"

Paper raussuchen

2. Einfluss von Twitter auf Weltbild/Meinung/ usw.
3. Twitter als Nachrichtenmedium (Can Twitter Replace Newswire (Petrovic et. al))
4. Anatomie eines Tweets
   a) Welche Informationen sind in einem Tweet enthalten?
   b) Konzentration auf Daten die Hinweise zur räumlichen Lage geben könnten aber auch allgemein auf die Daten eingehen.

2.3 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

2.4 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

Allgemeine geografische Grundbegriffe.

Evetnuell hier fehl am Platz

2.5 N-Gramme

1. NGramme allgemein, Verwendung, Beispiele.
2. Zusammenhang zwischen Länge/Grad eines N-Grammes und Wahrscheinlichkeiten. -> mathematische Herleitung?!

Nochmal genau prüfen, Zusammenhang zu Markov Modell und NGram Statistik herausstellen

2.6 Stand der Technik

1. Naiver Ansatz -> Geotagging mit Google Maps API V3
   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) allgemeiner Ansatz : Geotagged Tweets analysieren (Inhalt/andere Indikatoren usw.), zuordnen zu geografischen B̈ereichenünd daraus lernen.
   b) Inhaltsanalyse
   c) Indikatoransatz
   d) Multiindikatoransatz
   e)

in allen anderen Arbeiten gleiches Prinzip?

2.6.1 Probleme früherer Ansätze

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

2.6.2 Vorteile neuer Ansatz bei Mapping auf Geografische Daten

Notwendigkeit/Vorteile von Hierarchiebeziehungen im Mapping auf Geograohie Daten

3 Entwurf

3.1 Indikatoren zur Bestimmung der geografischen Lokation

Wie detailliert hier auf Framework eingehen? Preprocessor Konzept zur universellen Vorverarbeitung

3.1.1 Vorverarbeitung der Indikatoren

3.2

4 Implementierung

5 Leistungsbewertung

6 Schlussfolgerungen, Ausblick und Fragen

7 Zusammenfassung

Literaturverzeichnis

[FVMF13] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: Social butterflies or frequent fliers? *CoRR*, abs/1310.2671, 2013.

[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.

[PCV13] Reid Priedhorsky, 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.

[ti13] twitter inc. Final initial public offering(ipo) prospectus, 11 2013.