Computerlinguistische Analyse von Twitterdaten

Seminar

1. Einführung und Motivation

Tatjana Scheffler, Universität Potsdam

tatjana.scheffler@uni-potsdam.de

10.4.2013

Heute

- Einführung
- Kleines Beispiel
- Seminarplanung

Einführung

Motivation, Allgemeines

Motivation

Für Linguisten/Computerlinguisten:

- sehr große Datenmenge (noch wachsend)
- in maschinenlesbarer Form im Netz
- aktuelle Themen
- viele Metadaten
- Spontansprache aus verschiedenen Genres
- spezieller Stil (zwischen geschriebener und gesprochener Sprache)

Praxis: Social Media Monitoring

- *Präsenzanalyse*: Statistische Analyse, die die Präsenz eines Zielkonzeptes im Web/Social Media angibt
- *Trendanalyse*: Was entsteht gerade?
- *Tonalitätsanalyse*: Meinungsbild der Zielgruppe
- *Buzz-Analyse*: Involvement einer Zielgruppe zu einem bestimmten Thema
- *Profiling*: Erkenne Meinungsführer und Multiplikatoren
- *Quellenanalyse*: Bedeutende Orte im Netz

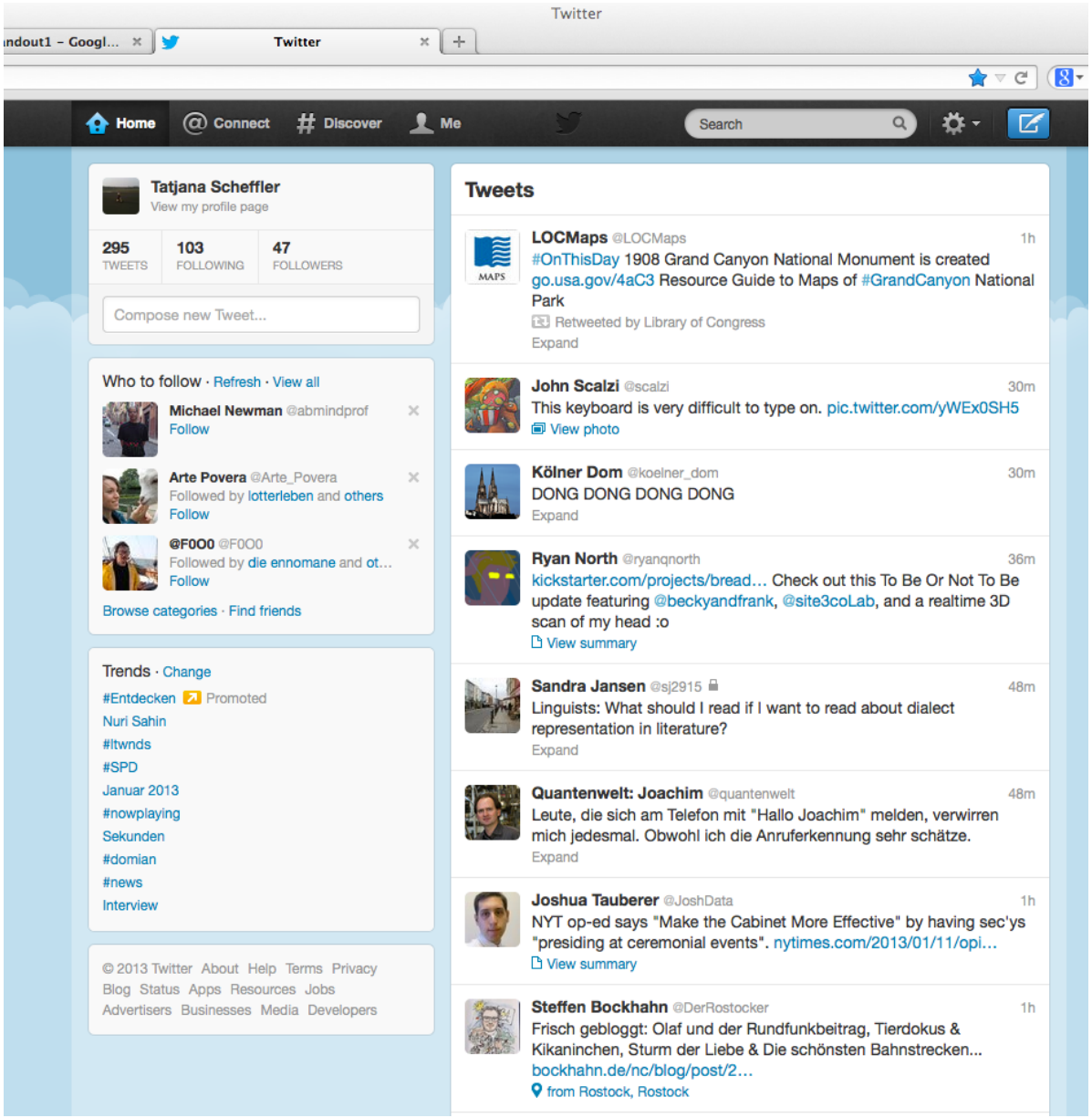
Dank an Tina Klüwer für die Zusammenstellung =)

Probleme bei der Analyse von Twitterdaten

- Bisherige Studien fast ausschließlich auf englischen Daten
- Twitter-Terms of Service verbieten viele forschungsrelevante Verwendungen der Daten
- Suchfunktion Twitter Search liefert unvollständige Ergebnisse
- Twitter-Stream-Zugang ist ratenlimitiert
  - Aber für Deutsch meist kein Problem
- http://www.buzzfeed.com/nostrich/how-twitter-gets-in-the-way-of-research

Twitter

- http://www.twitter.com
- Kurznachrichtendienst
- 140 Zeichen
- Follower-Friend-Beziehungen zwischen Nutzern
- Timeline aggregiert alle Nachrichten der Friends in Echtzeit
- @-Replies, Retweet-Relation, #Tag Themen
- Abrufen über Twitter API (JSON-Format)


Twitter
Twitter
Home
Connect
Discover
Me
Search
Tatjana Scheffler
View my profile page
295 TWEETS
103 FOLLOWING
47 FOLLOWERS
Compose new Tweet...
Who to follow · Refresh · View all
Michael Newman @abmindprof
Follow
Arte Povera @Arte_Povera
Followed by lotterleben and others
Follow
@F0O0 @F0O0
Followed by die ennomane and ot…
Follow
Browse categories · Find friends
Trends · Change
#Entdecken Promoted
Nuri Sahin
#ltwnds
#SPD
Januar 2013
#nowplaying
Sekunden
#domian
#news
Interview
© 2013 Twitter About Help Terms Privacy
Blog Status Apps Resources Jobs
Advertisers Businesses Media Developers
Tweets
LOCMaps @LOCMaps 1h
#OnThisDay 1908 Grand Canyon National Monument is created go.usa.gov/4aC3 Resource Guide to Maps of #GrandCanyon National Park
Retweeted by Library of Congress
Expand
John Scalzi @scalzi 30m
This keyboard is very difficult to type on. pic.twitter.com/yWEx0SH5
View photo
Kölner Dom @koelner_dom 30m
DONG DONG DONG DONG
Expand
Ryan North @ryanqnorth 36m
kickstarter.com/projects/bread… Check out this To Be Or Not To Be update featuring @beckyandfrank, @site3coLab, and a realtime 3D scan of my head :o
View summary
Sandra Jansen @sj2915 48m
Linguists: What should I read if I want to read about dialect representation in literature?
Expand
Quantenwelt: Joachim @quantenwelt 48m
Leute, die sich am Telefon mit "Hallo Joachim" melden, verwirren mich jedesmal. Obwohl ich die Anruferkennung sehr schätze.
Expand
Joshua Tauberer @JoshData 1h
NYT op-ed says "Make the Cabinet More Effective" by having sec'ys "presiding at ceremonial events". nytimes.com/2013/01/11/opi…
View summary
Steffen Bockhahn @DerRostocker 1h
Frisch gebloggt: Olaf und der Rundfunkbeitrag, Tierdokus & Kikaninchen, Sturm der Liebe & Die schönsten Bahnstrecken... bockhahn.de/nc/blog/post/2…
from Rostock, Rostock

Twitterdaten – Beispiel

- Leicht Vereinfachte JSON-Darstellung eines Tweets
- Attribut-Value Matrix
- (4 Folien)

```
$json (
|    text =  "Cro: sehr, sehr dope! #XmasJam"
|    source =  "Twitter for iPhone"
|    retweeted =  FALSE
|    favorited =  FALSE
|    retweet_count =  0
|    entities (
|    |    user_mentions => Array (0)
|    |    ( )
|    |    hashtags => Array (1)
|    |    (
|    |    |    ['0'] (
|    |    |    |    text =  "XmasJam"
|    |    |    |    indices => Array (2)
|    |    |    |    (
|    |    |    |    |    ['0'] =  22
|    |    |    |    |    ['1'] =  30
|    |    |    |    )
|    |    |    )
|    |    )
|    |    urls => Array (0)
|    |    ( )
|    )
```

```
|   place (
|   |   country =  "Germany"
|   |   place_type =  "city"
|   |   country_code =  "DE"
|   |   name =  "Stuttgart"
|   |   full_name =  "Stuttgart, Stuttgart"
|   |   url =  "http://api.twitter.com/1/geo/id/e385d4d639c6a423.json"
|   |   id =  "e385d4d639c6a423"
|   |   bounding_box (
|   |   |   coordinates => Array (1) (
|   |   |   |   ['0'] => Array (4) (
|   |   |   |   |   ['0'] => Array (2) (
|   |   |   |   |   |   ['0'] =  9.038755
|   |   |   |   |   |   ['1'] =  48.692343 )
|   |   |   |   |   ['1'] => Array (2) (
|   |   |   |   |   |   ['0'] =  9.315466
|   |   |   |   |   |   ['1'] =  48.692343 )
|   |   |   |   |   ['2'] => Array (2) (
|   |   |   |   |   |   ['0'] =  9.315466
|   |   |   |   |   |   ['1'] =  48.866225 )
|   |   |   |   |   ['3'] => Array (2) (
|   |   |   |   |   |   ['0'] =  9.038755
|   |   |   |   |   |   ['1'] =  48.866225 ) ) )
|   |   |   type =  "Polygon" )
|   |   attributes ( )
|   )
```

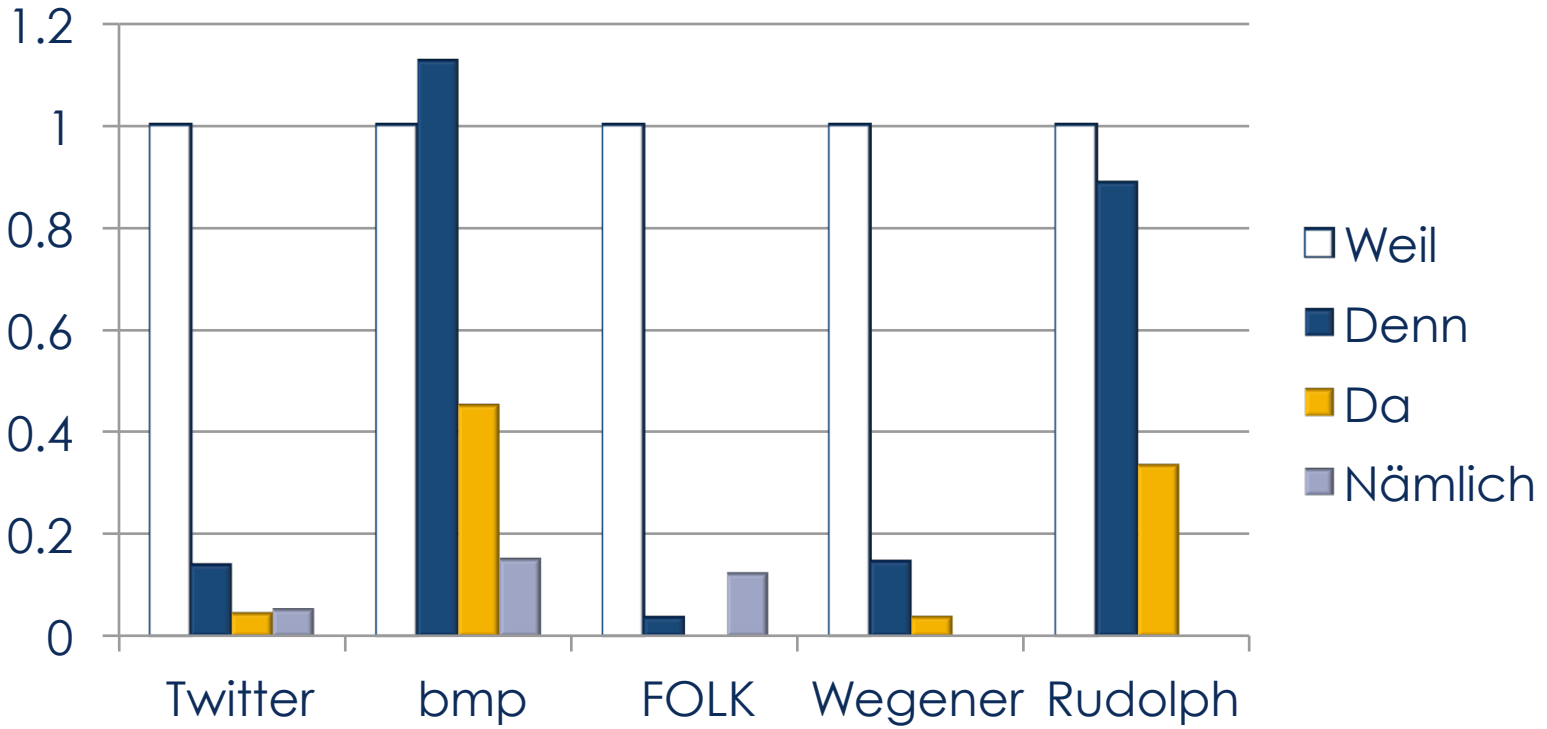
```
|   user (
|     |   friends_count =  1983
|     |   follow_request_sent =  NULL
|     |   profile_sidebar_fill_color =  "dbeefd"
|     |   profile_background_image_url_https = "https://si0.twimg.com/...0210.jpg"
|     |   profile_image_url =  "http://a3.twimg.com/…/twitter_normal.gif"
|     |   profile_background_color =  "f1f9ff"
|     |   url =  "http://christianfleschhut.de/"
|     |   id =  1182351
|     |   is_translator =  TRUE
|     |   screen_name =  "cfleschhut"
|     |   lang =  "en"
|     |   location =  "Karlsruhe, Germany"
|     |   followers_count =  1628
|     |   statuses_count =  3882
|     |   name =  "Christian Fleschhut"
|     |   description =  "93 â       til"
|     |   favourites_count =  166
|     |   profile_background_tile =  FALSE
|     |   listed_count =  54
|     |   created_at =  "Wed Mar 14 21:15:22 +0000 2007"
|     |   utc_offset =  3600
|     |   verified =  FALSE
|     |   show_all_inline_media =  TRUE
|     |   time_zone =  "Berlin"
|     |   geo_enabled =  TRUE
|   )
```

```
|    truncated =  FALSE
|    in_reply_to_status_id_str =  NULL
|    created_at =  "Thu Dec 22 21:22:36 +0000 2011"
|    in_reply_to_user_id =  NULL
|    id =  14996307043589324 8
|    in_reply_to_status_id =  NULL
|    geo (
|    |    coordinates => Array (2) (
|    |    |    ['0'] =  48.78509331
|    |    |    ['1'] =  9.18866308
|    |    )
|    |    type =  "Point"
|    )
|    in_reply_to_user_id_str =  NULL
|    id_str =  "14996307043589248"
|    in_reply_to_screen_name =  NULL
)
```

Beispiel

Analyse von Twitterdaten

Twitter-Stil: Kausalkonnektoren



Twitter = Wulff-Korpus; 253172 Deutsche Tweets über den Wulff-Skandal; bmp = Berliner Morgenpost-Teil von COSMAS II; FOLK = Forschungs- und Lehrkorpus Gesprochenes Deutsch, Dialoge; Wegener = Gesprochene Korpora 1980-1999 aus (Wegener 1999, Tab. 1); Rudolph = Geschriebene Texte (Rudolph 1982) zitiert in (Wegener 1999)

Seminarplanung

Sie sind gefragt!

Mögliche Themen

- Vorverarbeitung, Säuberung
- Topikerkennung
- Trenderkennung und -verfolgung
- Tonalitätsanalyse
- Soziolinguistik, Stil, Variabilität
- Profiling
- Information Retrieval/Document Retrieval
- Semantic Role Labelling
- Conversation Modelling

Seminarschein

- Vortrag zu einem der Themen, Diskussion
- Lesen, Beteiligung an der Diskussion
- Projekt und Ausarbeitung

Seminarwebseite mit Informationen:

http://www.ling.uni-potsdam.de/~scheffler/teaching/2013twitter.html

Planung

17.4.	Vorverarbeitung (TS)		
24.4.	xxx (TS)		
1.5.	Tag der Arbeit (kein Seminar)		
8.5.		5.6.	Vorträge
15.5.		12.6.	
22.5.		19.6.	
29.5.		26.6.	
3.7.	Kurzvorstellung der Projekte (alle)		
10.7.	Abschlussdiskussion, Weiteres (TS)		

Fragen

- tatjana.scheffler@uni-potsdam.de
- Sprechzeiten:
  Dienstags, 10-12 Uhr und nach Vereinbarung
  Haus 14, Raum 2.33
  Bitte per Email voranmelden!
- Aktuelle Informationen, Literatur, etc. auf der Webseite:

  http://www.ling.uni-potsdam.de/~scheffler/teaching/2013twitter.html