NLP of Microblogs

ESSLLI 2016

Bolzano/Bozen, Italy

Tatjana Scheffler & Manfred Stede

tatjana.scheffler@uni-potsdam.de / stede@uni-potsdam.de

August 22, 2016

Lecturers

Tatjana Scheffler
tatjana.scheffler@uni-potsdam.de

Manfred Stede
stede@uni-potsdam.de

Course Structure

Monday: Introduction, Data Collection, Preprocessing

Tuesday: Working with Metadata

Wednesday: Sentiment

Thursday: Conversations and Discourse

Friday: Linguistic Aspects

Introduction

Natural Language Processing of Microblogs

Twitter

… as a data source for linguists and computational linguists

Why Twitter?

for (computational) linguists:

- very large (and growing) amount of data
- machine-readable, online, easy access
- current topics
- a lot of metadata
- sponaneous language from different genres
- particular style (phenomena of both spoken and written language)

Application: Social Media Monitoring

- *presence analysis*: statistical analysis that indicates the presence of a concept on the web/in social media
- *trend analysis*: what is developing right now?
- *sentiment analysis*: opinions of a target group
- *buzz analysis*: involvement of a target group in a particular topic
- *profiling*: detect opinion leaders and multiplicators
- *source analysis*: significant locations on the web

In addition…

- sociolinguistics
- corpus linguistics
- discourse analysis
- Twitter as a source of empirical data

Twitter

- http://www.twitter.com
- microblog
- 140 characters
- based on follower-friend relations between users
- user timeline aggregates all posts by friends in real time
- @-replies, retweets, #tag topics
- access via the Twitter API (JSON format)

Problems with the analysis of Twitter data

- majority of previous work only on English data
- Twitter's terms of service prevent research-relevant uses of the data
- Twitter search yields incomplete results
- rate limiting on the Twitter stream access
  - but less of a problem for non-English languages!
- http://www.buzzfeed.com/nostrich/how-twitter-gets-in-the-way-of-research

Twitter data – an example

- simplified JSON representation of one tweet
- attribute value matrix
- (4 slides)

```
$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
)
```

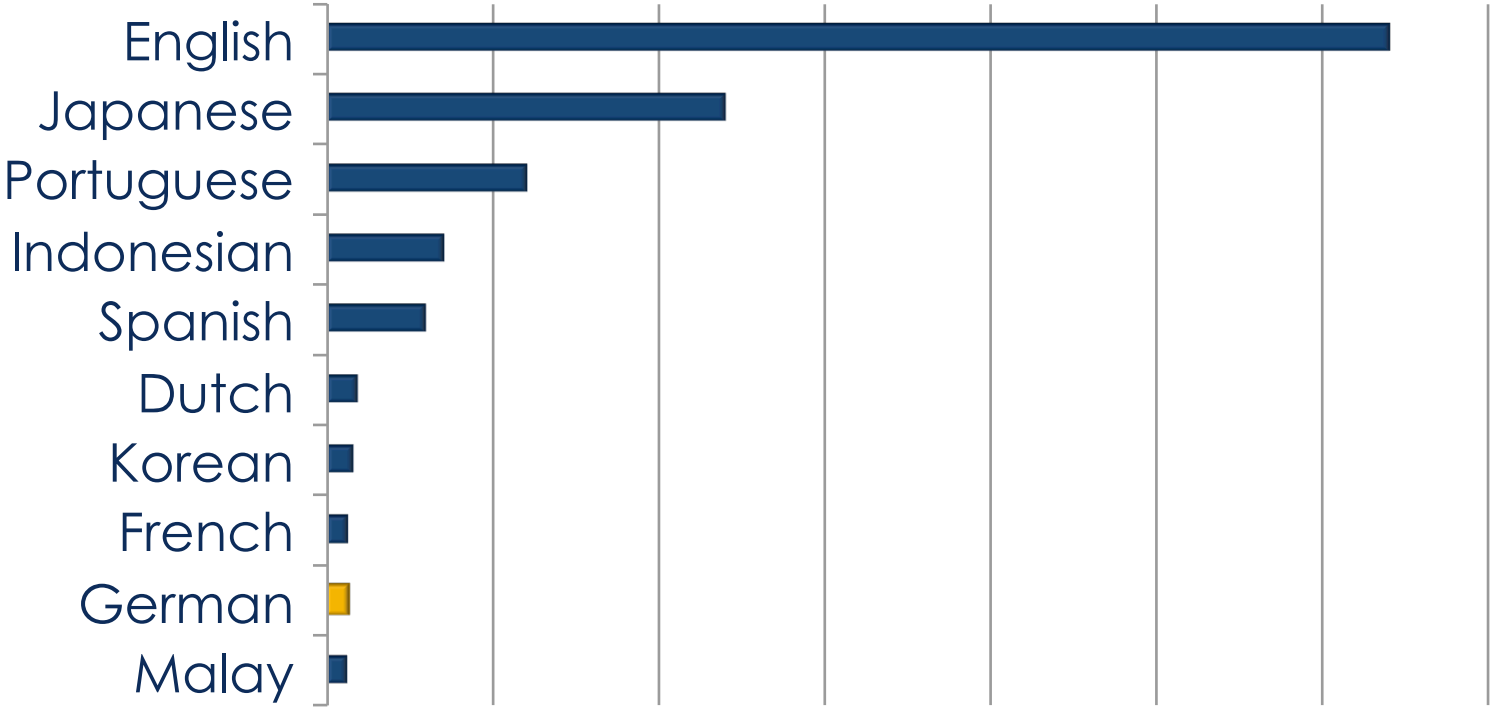
Creating a Twitter corpus

approach, problems

Twitter-APIs for creating corpora

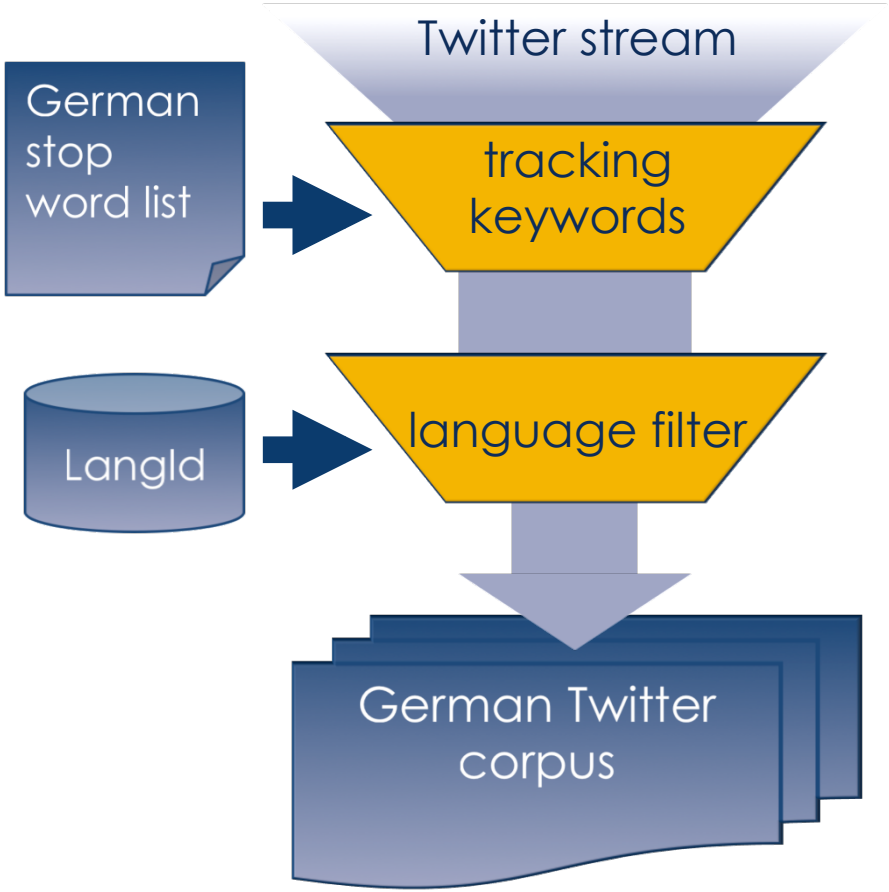
- Search API or Streaming API
- Search API: key words, up to 7 days into the past
- Streaming API:
  - real time stream of posted tweets
  - rate limitation
  - many non-German tweets
  - filter by:
    - geo-location (location)
    - up to 5000 user ids (follow)
    - up to 400 keywords (track)

Languages on Twitter



Source: Hong, Lichan, Convertino, Gregorio, and Chi, Ed. "Language Matters In Twitter: A Large Scale Study" International AAAI Conference on Weblogs and Social Media (2011)

Corpus creation



~ 500.000.000 tweets / day

~ xx.000.000 tweets / day

~ 1.000.000 tweets / day

Tools: access Twitter's streaming API

1. Python package: tweepy `https://github.com/tweepy/tweepy`
2. register own application, get access keys
3. create key word list
   - e.g.: filter stream for 397 most common German stop words
   - exclude foreign homographs: "war", "die", "des", …
   - loss of only ~5% of German tweets
4. Tweepy + langId for language identification
5. for example, use twython script: `http://www.ling.uni-potsdam.de/~scheffler/twitter/`

Language identification

- Twitter's own language identification is not accurate (seems to be based on user profile)
- Google Compact Language Detector:

  `pypi.python.org/pypi/chromium_compact_language_detector/`
- Langid: `https://github.com/saffsd/langid.py`
  by Lui/Baldwin "langid.py: An Off-the-shelf Language Identification Tool" (ACL 2012)

German tweets	Langid	Google CLD	Twitter
precision	97%	96%	~ 40%

Dealing with Twitter corpora

- **Twitter ToS prohibits sharing of aggregated tweets (=corpora)!**
- corpus sharing only via tweet IDs; time-consuming recrawling of individual tweets, e.g. via

  `https://github.com/lintool/twitter-tools`
- deletion of tweets and/or accounts: 21,2% of the Tweets2011 corpus were unretrievable after 9 months
- How to anonymize tweets in scientific papers?
  - removal of @handles
  - often still googleable

Other tools: TAGS

- Twitter Archiving Google Sheet: https://tags.hawksey.info/
- automatically run API queries in a Google Sheets doc
- save / export the archive

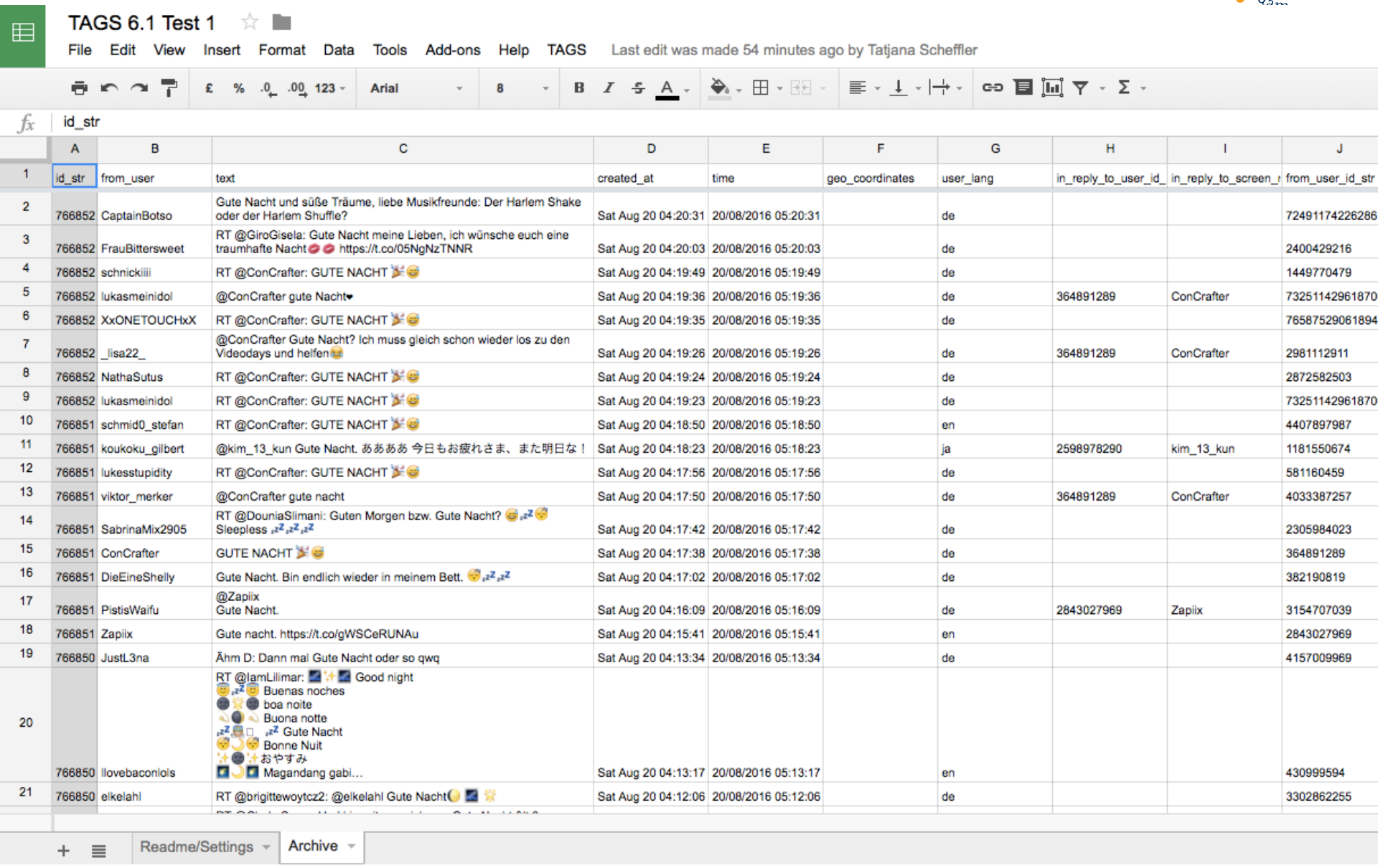
TAGS 6.1 Test 1

File Edit View Insert Format Data Tools Add-ons Help TAGS Last edit was made 53 minutes ago by Tatjana Scheffler

Roboto 10

http://tags.hawksey.info

	A	B	C
8			
9	2. Enter term	gute AND nacht	<- you can use serach operators like AND OR as well as from: and to: eg '#JobsNow AND from:BarackObama' (without quotes)
10			
11	**Note:** Make a one off collection with `TAGS > Run now!` or set a trigger to collect every hour `TAGS > Update archive every hour`. To change the frequency open `Tools > Script Editor` then `Triggers > Current script's triggers...` and adjust		
12	**Advanced Settings:**		
13	Period	default	
15	Follower count filter	0	<- if search term is being spammed you can set the minimum followers a person must have to be included in archive
16	Number of tweets	10000	<- maximum varies based on the type of archive you are collecting
17	Type	search/tweets	<- use a search term in step 3 above to get results from last 7 days
18	**Stats**		
19	Number of Tweets	19,461	
20	Unique tweets	19,375	
21	First Tweet	15/08/2016 21:15:37	
22	Last Tweet	20/08/2016 05:20:31	
23	**Make interactive**		
24	Turn your archive into an interactive online resource using TAGSExplorer - see http://bit.ly/TAGSsetup		



TAGS 6.1 Test 1

File Edit View Insert Format Data Tools Add-ons Help TAGS Last edit was made 54 minutes ago by Tatjana Scheffler

id_str

	A	B	C	D	E	F	G	H	I	J
1	id_str	from_user	text	created_at	time	geo_coordinates	user_lang	in_reply_to_user_id_	in_reply_to_screen_	from_user_id_str
2	766852	CaptainBotso	Gute Nacht und süße Träume, liebe Musikfreunde: Der Harlem Shake oder der Harlem Shuffle?	Sat Aug 20 04:20:31	20/08/2016 05:20:31		de			72491174226286
3	766852	FrauBittersweet	RT @GiroGisela: Gute Nacht meine Lieben, ich wünsche euch eine traumhafte Nacht💋💋 https://t.co/05NgNzTNNR	Sat Aug 20 04:20:03	20/08/2016 05:20:03		de			2400429216
4	766852	schnickiiii	RT @ConCrafter: GUTE NACHT 🎉😇	Sat Aug 20 04:19:49	20/08/2016 05:19:49		de			1449770479
5	766852	lukasmeinidol	@ConCrafter gute Nacht♥	Sat Aug 20 04:19:36	20/08/2016 05:19:36		de	364891289	ConCrafter	73251142961870
6	766852	XxONETOUCHxX	RT @ConCrafter: GUTE NACHT 🎉😇	Sat Aug 20 04:19:35	20/08/2016 05:19:35		de			76587529061894
7	766852	_lisa22_	@ConCrafter Gute Nacht? Ich muss gleich schon wieder los zu den Videodays und helfen😂	Sat Aug 20 04:19:26	20/08/2016 05:19:26		de	364891289	ConCrafter	2981112911
8	766852	NathaSutus	RT @ConCrafter: GUTE NACHT 🎉😇	Sat Aug 20 04:19:24	20/08/2016 05:19:24		de			2872582503
9	766852	lukasmeinidol	RT @ConCrafter: GUTE NACHT 🎉😇	Sat Aug 20 04:19:23	20/08/2016 05:19:23		de			73251142961870
10	766851	schmid0_stefan	RT @ConCrafter: GUTE NACHT 🎉😇	Sat Aug 20 04:18:50	20/08/2016 05:18:50		en			4407897987
11	766851	koukoku_gilbert	@kim_13_kun Gute Nacht. あああああ 今日もお疲れさま、また明日な！	Sat Aug 20 04:18:23	20/08/2016 05:18:23		ja	2598978290	kim_13_kun	1181550674
12	766851	lukesstupidity	RT @ConCrafter: GUTE NACHT 🎉😇	Sat Aug 20 04:17:56	20/08/2016 05:17:56		de			581160459
13	766851	viktor_merker	@ConCrafter gute nacht	Sat Aug 20 04:17:50	20/08/2016 05:17:50		de	364891289	ConCrafter	4033387257
14	766851	SabrinaMix2905	RT @DouniaSlimani: Guten Morgen bzw. Gute Nacht? 😅💤😴 Sleepless 💤💤💤	Sat Aug 20 04:17:42	20/08/2016 05:17:42		de			2305984023
15	766851	ConCrafter	GUTE NACHT 🎉😇	Sat Aug 20 04:17:38	20/08/2016 05:17:38		de			364891289
16	766851	DieEineShelly	Gute Nacht. Bin endlich wieder in meinem Bett. 😴💤💤	Sat Aug 20 04:17:02	20/08/2016 05:17:02		de			382190819
17	766851	PistisWaifu	@Zapiix Gute Nacht.	Sat Aug 20 04:16:09	20/08/2016 05:16:09		de	2843027969	Zapiix	3154707039
18	766851	Zapiix	Gute nacht. https://t.co/gWSCeRUNAu	Sat Aug 20 04:15:41	20/08/2016 05:15:41		en			2843027969
19	766850	JustL3na	Ähm D: Dann mal Gute Nacht oder so qwq	Sat Aug 20 04:13:34	20/08/2016 05:13:34		de			4157009969
20	766850	Ilovebaconlols	RT @IamLilimar: 🌌✨🌌 Good night 😇💤😇 Buenas noches ⚫🌟⚫ boa noite 💫⚫💫 Buona notte 💤💤 Gute Nacht 😴🌙😴 Bonne Nuit ✨⚫✨ おやすみ 🌌🌙🌌 Magandang gabi…	Sat Aug 20 04:13:17	20/08/2016 05:13:17		en			430999594
21	766850	elkelahl	RT @brigittewoytcz2: @elkelahl Gute Nacht🌕🌌🌟	Sat Aug 20 04:12:06	20/08/2016 05:12:06		de			3302862255

Readme/Settings | Archive

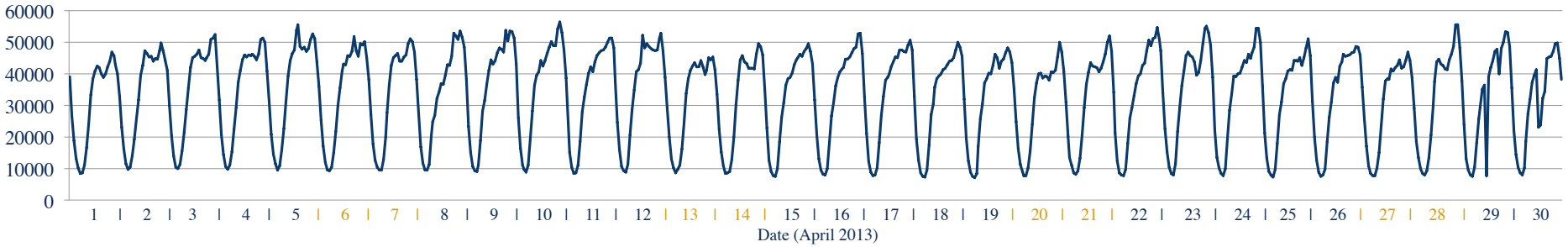
TAGS – create one tonight!

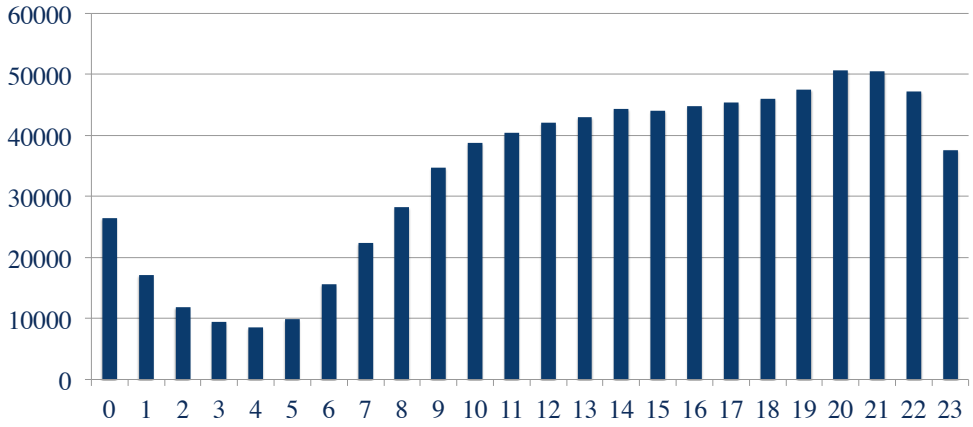
1. get TAGS, a Twitter and a Google account, log in
2. click Make a Copy
3. TAGS -> Setup Twitter Access, authorize
4. insert search terms and settings
5. TAGS -> Start updating archive every hour

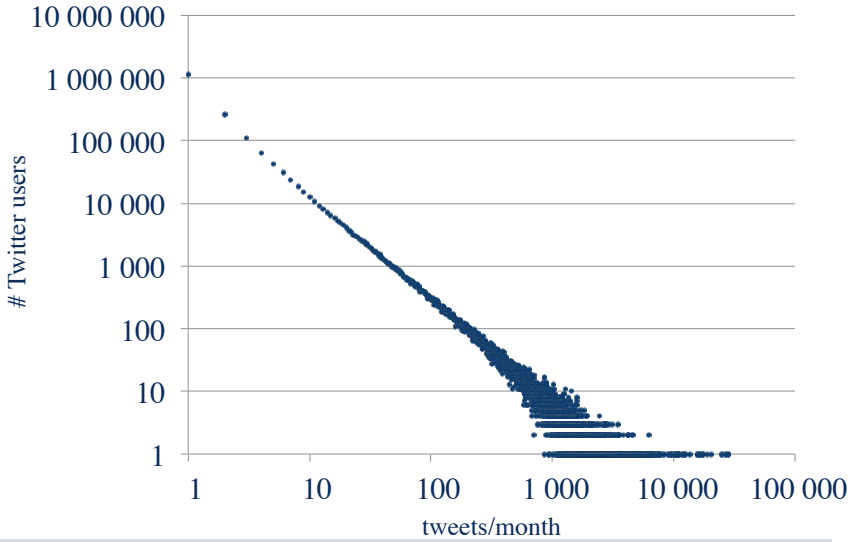
Finished! It will run in the background even if you're not online.

What is Twitter data like?

German Twitter data

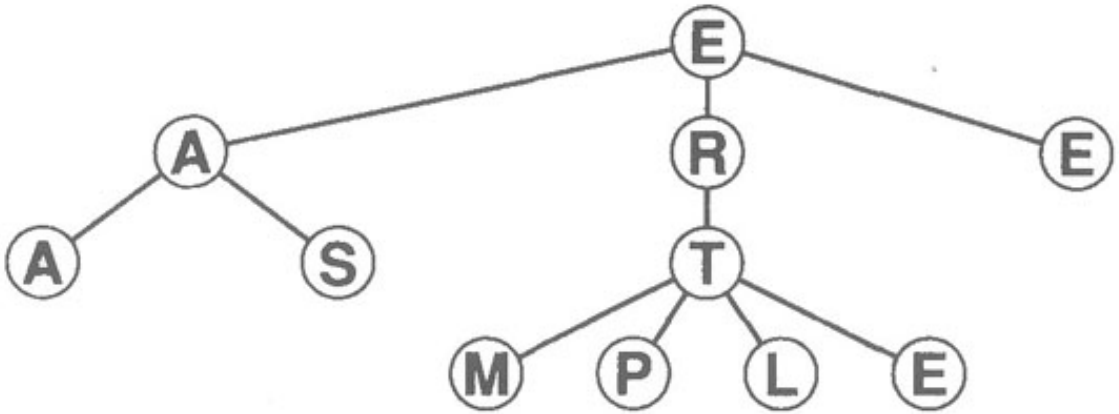






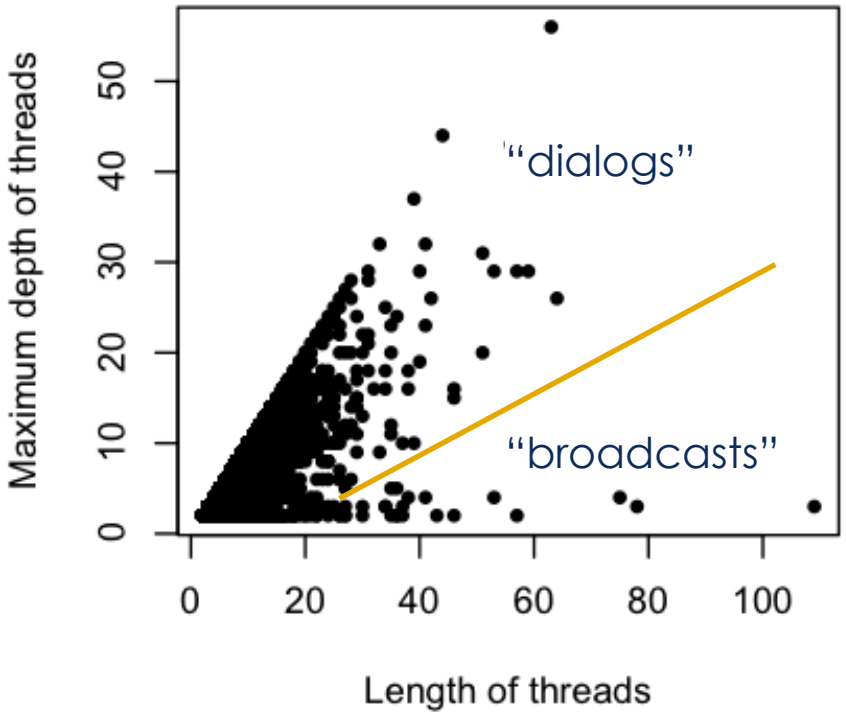
Microblogs = conversations

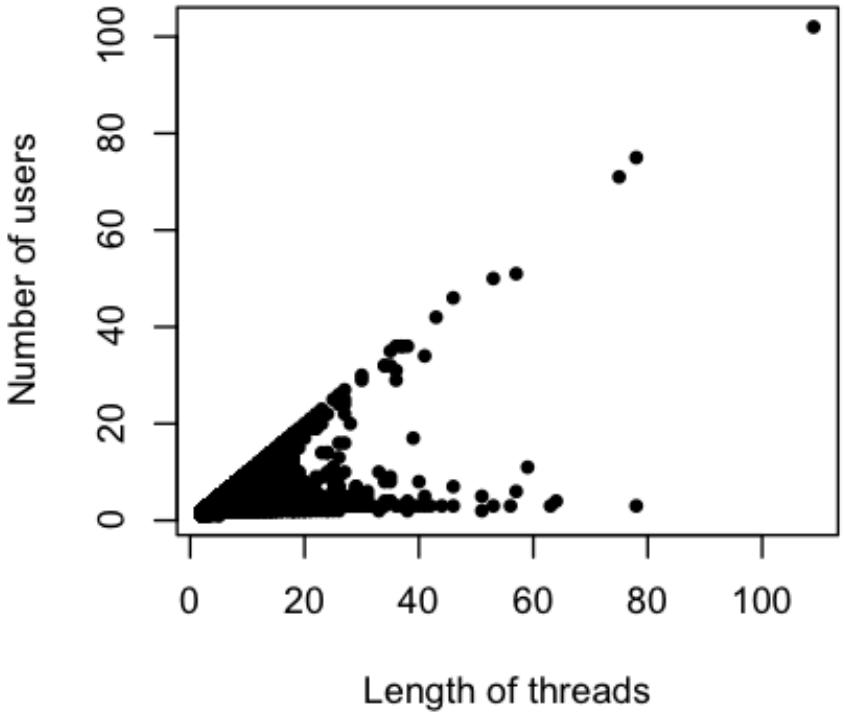
- reply-to-function creates conversations on Twitter
- ~20-25% of tweets are replies
- ~40% of tweets = part of conversations
- tree structure:



Types of "conversations"

- tree topology:






SF QuakeBot @earthquakesSF · 17. März
1.8 magnitude earthquake occurred 6.84mi E of Quarry near Portola Valley, CA. Details: eqbot.com/jLE Map: eqbot.com/jLW
1
2
SF QuakeBot @earthquakesSF · 17. März
A 1.3 magnitude earthquake occurred 5.59mi ESE of Livermore, California. Details: eqbot.com/jBF Map: eqbot.com/jBr
1
3
BlonderEngelTV @DirtyTinaTV · 4 Min.
Neue Features im geilen #Erotik #Browsergame. Jetzt #Gratis anmelden unter tinyurl.com/mazg8dj #gamescom2014
Girls-Check @GirlsCheck1 · 6 Min.
#Amateure die #reale #Dates suchen - 20 Coins #Gratis #geschenkt auf tinyurl.com/o5arsfp #date #vorsaetze


Netflix Bot @netflix_bot · 6. März
Jail Caesar (2012) NR [Movie] is now available on Netflix Instant - ift.tt/1QWVKIs
1
Netflix Bot @netflix_bot · 6. März
Policeman (2011) NR [Movie] is now available on Netflix Instant - ift.tt/1psF7Pd
1
Big Ben @big_ben_clock · 3 Std.
BONG BONG BONG BONG BONG BONG
57
47
Big Ben @big_ben_clock · 4 Std.
BONG BONG BONG BONG BONG
47
46

bots

- useful information: SF QuakeBot, weather info
- fun bots
- affiliate spam
- app-related bots

recognition of automatic content

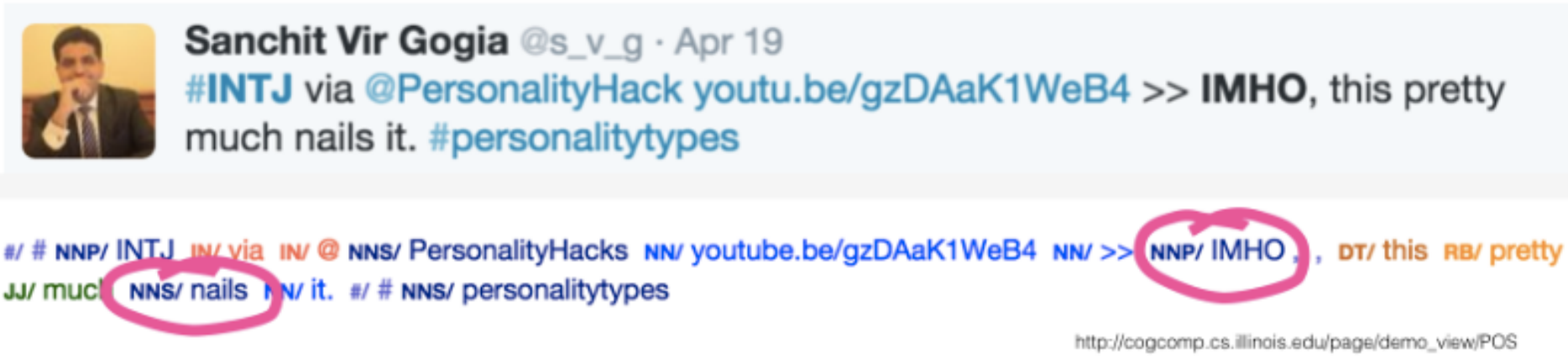
- clients: 10 most frequent clients = 80% of the data
- content: many hashtags, URLs
- time: frequent posts
- network structure: too few or too many followers
- interaction: not part of conversations

- Tool: BotOrNot? http://truthy.indiana.edu/botornot/

What is Twitter data like?

Preprocessing

Why preprocessing?



Tagging Twitter #hard

ESSLLI 2016 Fortuitous Data: https://fortuitousdata.github.io/slides/fortuitous-data-ESSLLI-day1.pdf

Twitter as corpus data

- special tokens (emoticons/emoji, #tags, URLs)
- slang, dialect, colloquial language, typos
- **preprocessing**
  - normalization(diacritics, elongations, typos?)
  - treatment of special tokens (@handles, #tags)
  - tokenization
  - sentence segmentation

```
uuund der akku hält und hält....super :) #iphone4s

Der Tagesspiegel: Busemann: Keine Weisung an
Staatsanwaelte in Wulff-Affaere - http://t.co/
Xef3vrUj #Pressemitteilung
```

normalization

- Alegria (2008)
- Aw (2006)
- Beaufort (2010)
- Brody Diakopoulos (2011)
- Choudhury (2007)
- Clark and Araki (2011)
- Cook and Stevenson (2009)
- Han and Baldwin (2011)
- Kaufmann and Kalita (2010)
- Kobus (2008)
- Krawczyk (2009)
- Kukich (1992)
- Liu (2011)
- Melero (2012)
- Oliva (2012)
- Sproat (2001)
- Toutanova and Moore (2002)
- Wei (2011)
- Yvon (2010)
- …

Approaches for normalization

level:

- graphemes / phonemes
- words
- phrases

methodology:

- rule based
- statistical
- hybrid

steps for text normalization

1. clean noise
2. restore diacritics
3. normalize slang
4. squeeze multiple characters
5. split sentences
6. tokenize

(Wladimir Sidorenko)

1. clean noise

- removal or replacement of Twitter-specific entities: emoticons, @handles, #tags

```
@_0816_ That's nice :-)
I'll keep looking...*hihi* ;-)

That's nice %PosSmiley
I'll keep looking...%PosSmiley %PosSmiley
REPLACED      0       0                              @_0816_
REPLACED      21      10        %PosSmiley           :-)
REPLACED      62      10        %PosSmiley           *hihi*
REPLACED      73      10        %PosSmiley           ;-)
```

2. restore diacritics

Der israelisch-palaestinensische Konflikt ist ein Konflikt um Land, die Sicherheit von Grenzen und um die Staatlichkeit zweier Nationen.

Der israelisch-palästinensische Konflikt ist ein Konflikt um Land, die Sicherheit von Grenzen und um die Staatlichkeit zweier Nationen.

3. normalize slang

- more important in English?
  do tngrs luv 2 txt msg?

```
@rmmarchy Und da isser wieder, der VIP-Vertrag
für's Bobby Car. #cdu #wulff
```

```
@rmmarchy Und da ist er wieder, der VIP-
Vertrag für das Bobby Car. #cdu #wulff
```

4. squeeze multiple charaters

```
Ichhh haaasssseee diiieeseeen Tiiisch

Ich hasse diesen Tisch
```

- “Hase” or “hasse”?
- dictionary lookup for all strings with three or more identical characters in a row: Ichhh, Ichh, Ich
- all found variants are kept

5. split sentences

```
Dieter Golombek, Jurysprecher Dt. Lokalj.-
Preis: "Wir brauchen erklärenden Journalismus
mehr denn je." drehscheibe.org/weblog/?p=3926
#video #bpbwahl
```

```
Dieter Golombek, Jurysprecher Dt. Lokalj.-
Preis: "Wir brauchen erklärenden Journalismus
mehr denn je."<sentence/>
drehscheibe.org/weblog/?p=3926<sentence/>
#video #bpbwahl<sentence/>
```

6. tokenization

```
@_0816_ Das ist lieb von Dir :-) Ich suche
weiter...in der Kueche*hihi* ;-)
```

TreeTagger Tokenizer:

```
@_0816_ Das ist lieb von Dir :- ) Ich suche
weiter ... in der Kueche*hihi* ; - )
```

Custom Tokenizer:

```
@_0816_ Das ist lieb von Dir :-) Ich suche
weiter ... in der Kueche *hihi* ;-) <sentence/>
```

Summary Day 1

- Twitter data: easy to collect
- API or tool access: stream / search
- metadata
- non-human generated content, noisy text
- often: preprocessing; automatically try to make social media text closer to “standard” forms of text

Coming up…

Tuesday: Working with Metadata

Wednesday: Sentiment (classification and clustering)

Thursday: Conversations and Discourse

Friday: Linguistic Aspects

…. stay tuned!

Thank you.

tatjana.scheffler@uni-potsdam.de

Thanks / image references

- Normalization work/slides by Wladimir Sidorenko.
- graphics from:

  Tatjana Scheffler. A German Twitter Snapshot. In: *Proceedings of LREC*, Reykjavik, Iceland. 2014.

  und von den Postern unter:

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