Quantitative Approaches to Discourse on Social Media

Workshop, Computational Humanities Summer School

Heidelberg

Tatjana Scheffler, Universität Potsdam

tatjana.scheffler@uni-potsdam.de

@tschfflr

July 16, 2019

Plan

- Collecting and storing corpora
- Conversation structure on social media
- Tools, methods, and tutorials
- Non-standard language

Work book (ipynb) for part 2

https://github.com/TScheffler/
2019HCH-conv

Introduction

Computational Linguistics and Social Media

Why Social Media?

for (computational) linguists:

- very large (and growing) amount of data
- machine-readable, online, easy access
- current topics
- a lot of metadata
- spontaneous 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
- social media as a source of empirical data
- …

Getting Social Media Data

Social Media with Text

- Twitter: relatively easy API access (more soon)
- Facebook: only public groups, some datasets available
- Wikipedia comments: from Wikipedia dump, e.g. https://figshare.com/articles/Wikipedia_Talk_Corpus/4264973
- Amazon reviews: http://jmcauley.ucsd.edu/data/amazon/
- Reddit: 2015 corpus or through the API https://archive.org/details/2015_reddit_comments_corpus
- http://www.clips.ua.ac.be/pages/pattern-web APIs

- Blogs: RSS and BeautifulSoup (get last few posts)
- …

Twitter

- http://www.twitter.com
- microblog
- 140 characters (now 280)
- 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 =  149963070435893248
|    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 =  "149963070435893248"
|    in_reply_to_screen_name =  NULL
)
```

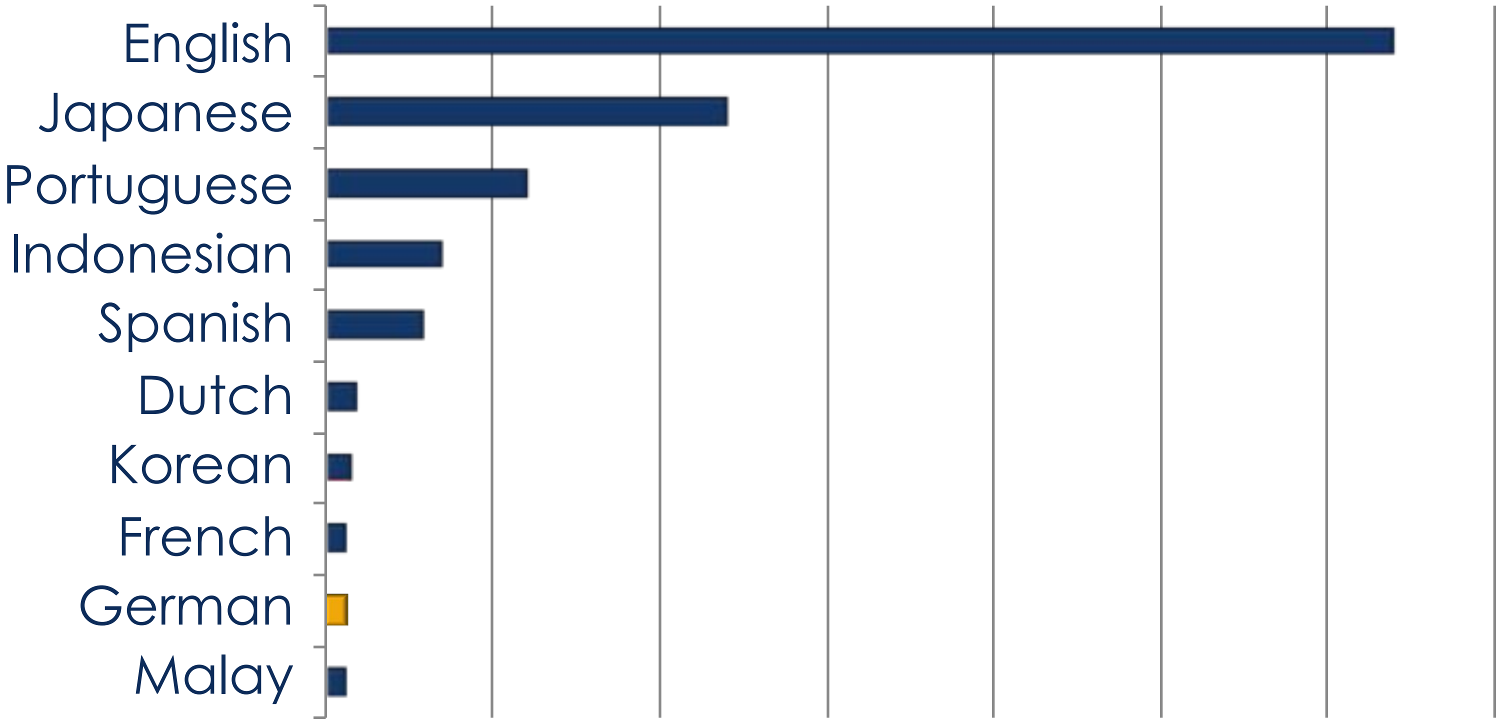
Creating a Twitter corpus

approach, problems

Twitter-APIs for creating corpora

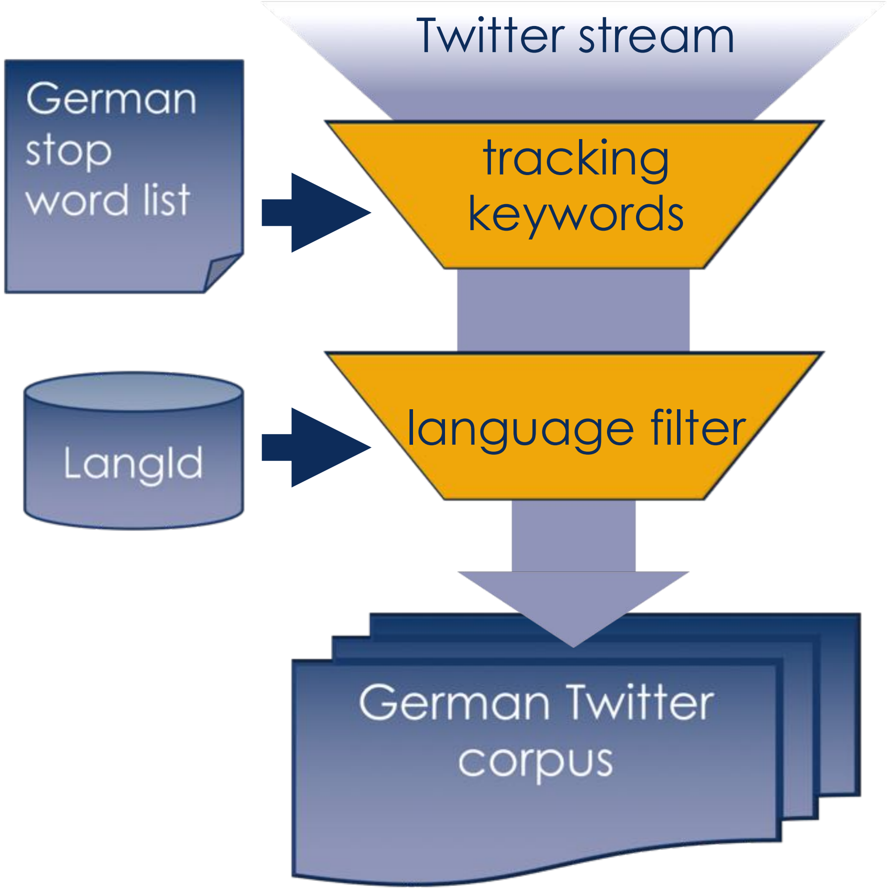
- Search API or Streaming API
- Search API: key words, up to 7 days into the past
- Streaming API:
  - real time stream of posted tweets
  - rate limitation
  - many non-German tweets
  - filter by:
    - geo-location (location)
    - up to 5000 user ids (follow)
    - up to 400 keywords (track)

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. register own application, get access keys
2. Python package: tweepy
   `https://github.com/tweepy/tweepy`
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 twarc (hydrate): `https://github.com/DocNow/twarc`
- deletion of tweets and/or accounts: 21,2% of the Tweets2011 corpus were unretrievable after 9 months

Ethics

- How to anonymize tweets in scientific papers?
    - removal of @handles -> still googleable

- recommendation:
    - use celebrities
    - get consent if possible

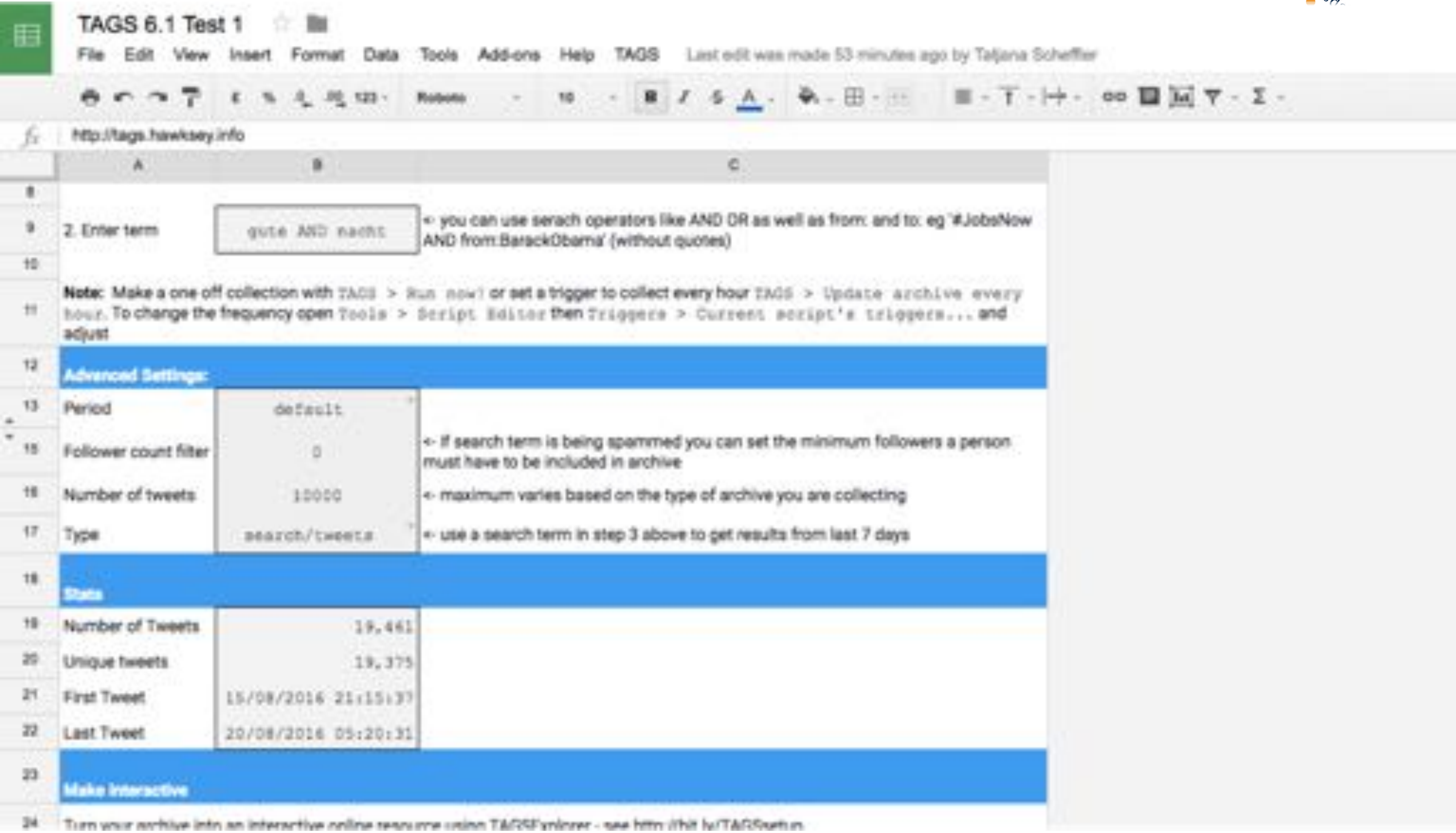
- Williams/Burnap/Sloan, 2017: Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation http://journals.sagepub.com/doi/full/10.1177/0038038517708140

Twarc

- `https://github.com/DocNow/twarc`
- Python package and command line interface
- retrieve conversations based on a tweet
- dehydrate/hydrate tweet ids

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
A
B
C

http://tags.hawksey.info

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)

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

Advanced Settings:

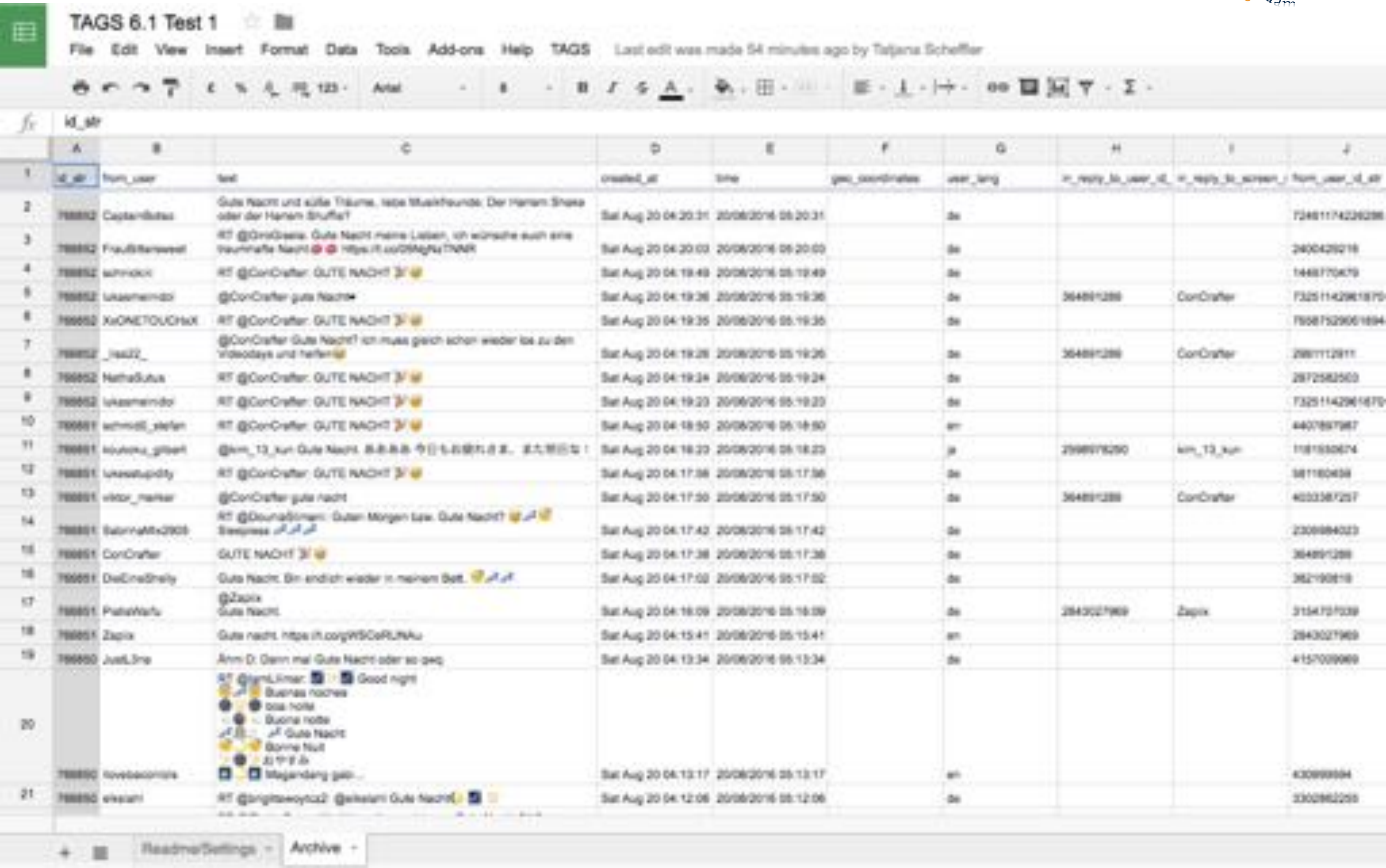
Period	default	
Follower count filter	0	<- If search term is being spammed you can set the minimum followers a person must have to be included in archive
Number of tweets	10000	<- maximum varies based on the type of archive you are collecting
Type	search/tweets	<- use a search term in step 3 above to get results from last 7 days

Stats

Number of Tweets	19,461
Unique tweets	19,375
First Tweet	15/08/2016 21:15:37
Last Tweet	20/08/2016 05:20:31

Make Interactive

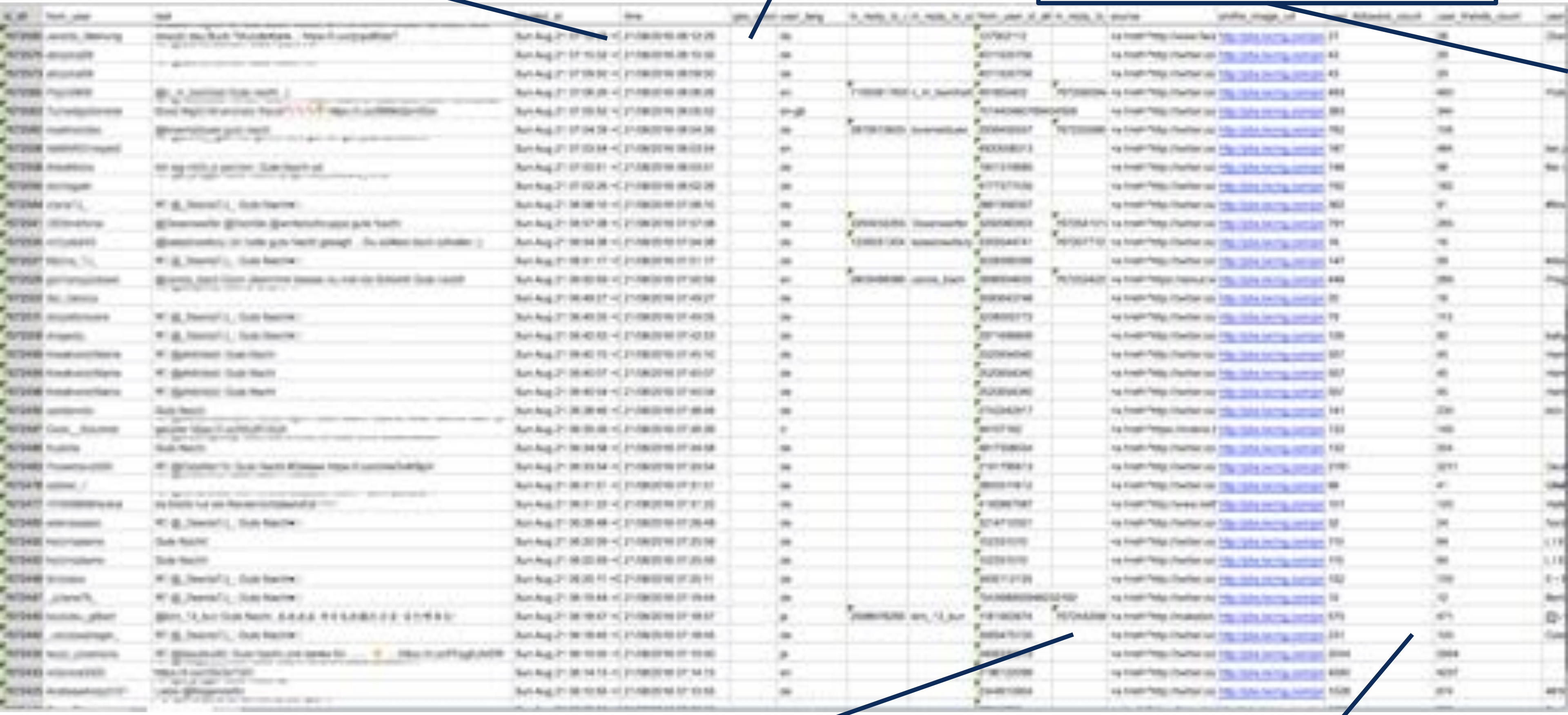
Turn your archive into an interactive online resource using TAGSExplorer - see http://bit.ly/TAGSsetup



TAGS 6.1 Test 1

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	766652	CaptainBotsu	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			72481174226296
3	766652	FrauBittersweet	RT @GinaGisela: Gute Nacht meine Lieben, ich wünsche euch eine traumhafte Nacht🌸🌸 https://t.co/D9MgNuTNNR	Sat Aug 20 04:20:03	20/08/2016 05:20:03		de			2400429219
4	766652	schnickX	RT @ConCrafter: GUTE NACHT 🌛😴	Sat Aug 20 04:19:49	20/08/2016 05:19:49		de			1449770479
5	766652	lukasmeindoi	@ConCrafter gute Nacht	Sat Aug 20 04:19:36	20/08/2016 05:19:36		de	364891288	ConCrafter	73251142961870
6	766652	XxONETOUCHxX	RT @ConCrafter: GUTE NACHT 🌛😴	Sat Aug 20 04:19:35	20/08/2016 05:19:35		de			76587529061894
7	766652	_Isa22_	@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	364891288	ConCrafter	2981112911
8	766652	NathaSutus	RT @ConCrafter: GUTE NACHT 🌛😴	Sat Aug 20 04:19:24	20/08/2016 05:19:24		de			2872582503
9	766652	lukasmeindoi	RT @ConCrafter: GUTE NACHT 🌛😴	Sat Aug 20 04:19:23	20/08/2016 05:19:23		de			73251142961870
10	766651	schmidt_stefan	RT @ConCrafter: GUTE NACHT 🌛😴	Sat Aug 20 04:18:50	20/08/2016 05:18:50		en			4407897987
11	766651	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	766651	lukasstupidity	RT @ConCrafter: GUTE NACHT 🌛😴	Sat Aug 20 04:17:58	20/08/2016 05:17:58		de			581160438
13	766651	viktor_marker	@ConCrafter gute nacht	Sat Aug 20 04:17:50	20/08/2016 05:17:50		de	364891288	ConCrafter	4033387257
14	766651	SabrinaMix2905	RT @DounaSimani: Guten Morgen bzw. Gute Nacht! 😴🌛😴 Sleepness 🌛🌛🌛	Sat Aug 20 04:17:42	20/08/2016 05:17:42		de			2309984023
15	766651	ConCrafter	GUTE NACHT 🌛😴	Sat Aug 20 04:17:38	20/08/2016 05:17:38		de			364891288
16	766651	DieEineShelly	Gute Nacht. Bin endlich wieder in meinem Bett. 😴🌛🌛	Sat Aug 20 04:17:02	20/08/2016 05:17:02		de			362190818
17	766651	PiataWarfu	@Zapix Gute Nacht	Sat Aug 20 04:16:09	20/08/2016 05:16:09		de	2843027969	Zapix	3154707039
18	766651	Zapix	Gute nacht. https://t.co/gWSCeRUNAu	Sat Aug 20 04:15:41	20/08/2016 05:15:41		en			2843027969
19	766650	JustLJna	Ahm D: Dann mal Gute Nacht oder so gwq	Sat Aug 20 04:13:34	20/08/2016 05:13:34		de			4157009969
20	766650	lovebaconlols	RT @ZeynLilmar: 🌃 🌃 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			430999994
21	766650	elesiani	RT @brigitteweyrtca2: @elesiani Gute Nacht🌛 🌃	Sat Aug 20 04:12:06	20/08/2016 05:12:06		de			3302862255



time

geo_coordinates

user profile info

in_reply_to

user network

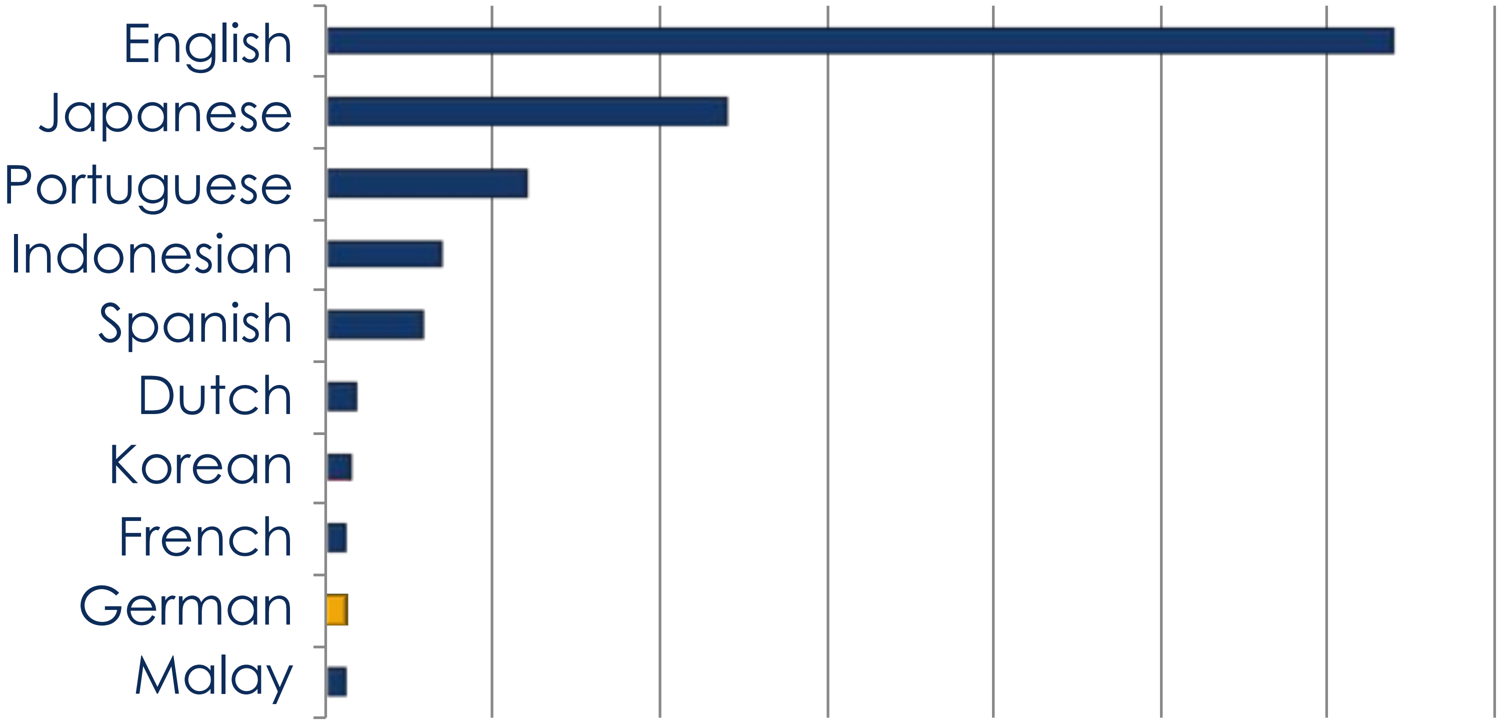
TAGS – create one tonight!

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.

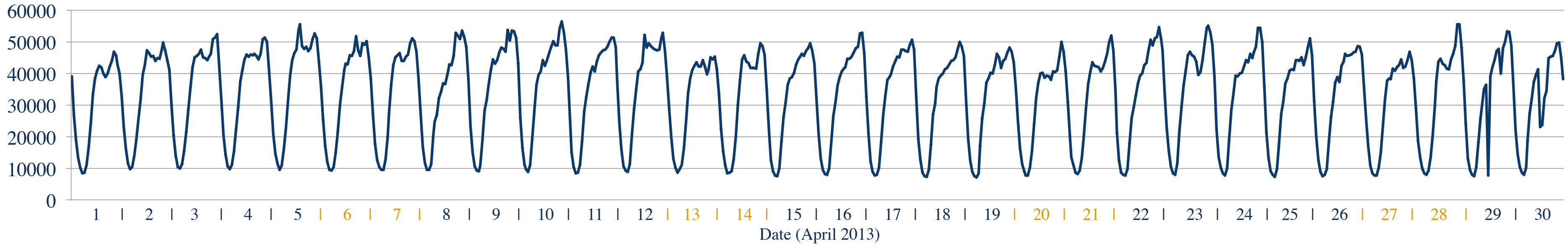
What is Twitter data like?

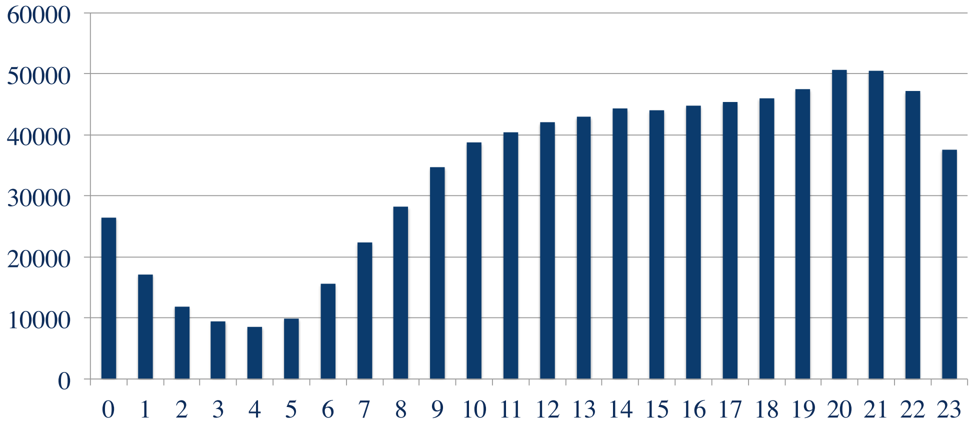
Languages on Twitter

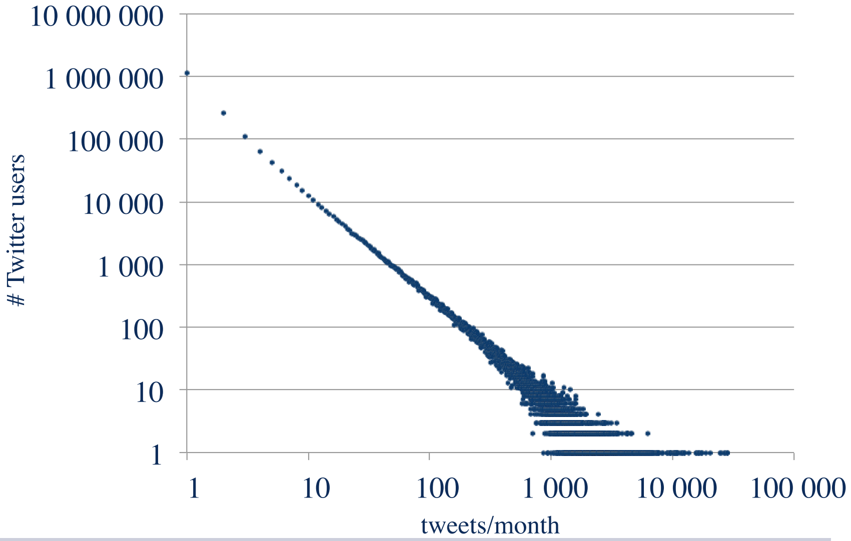


Source: Hong, Lichan, Convertino, Gregorio, and Chi, Ed. "Language Matters In Twitter: A Large Scale Study" International AAAI Conference on Weblogs and Social Media (2011)

German Twitter data







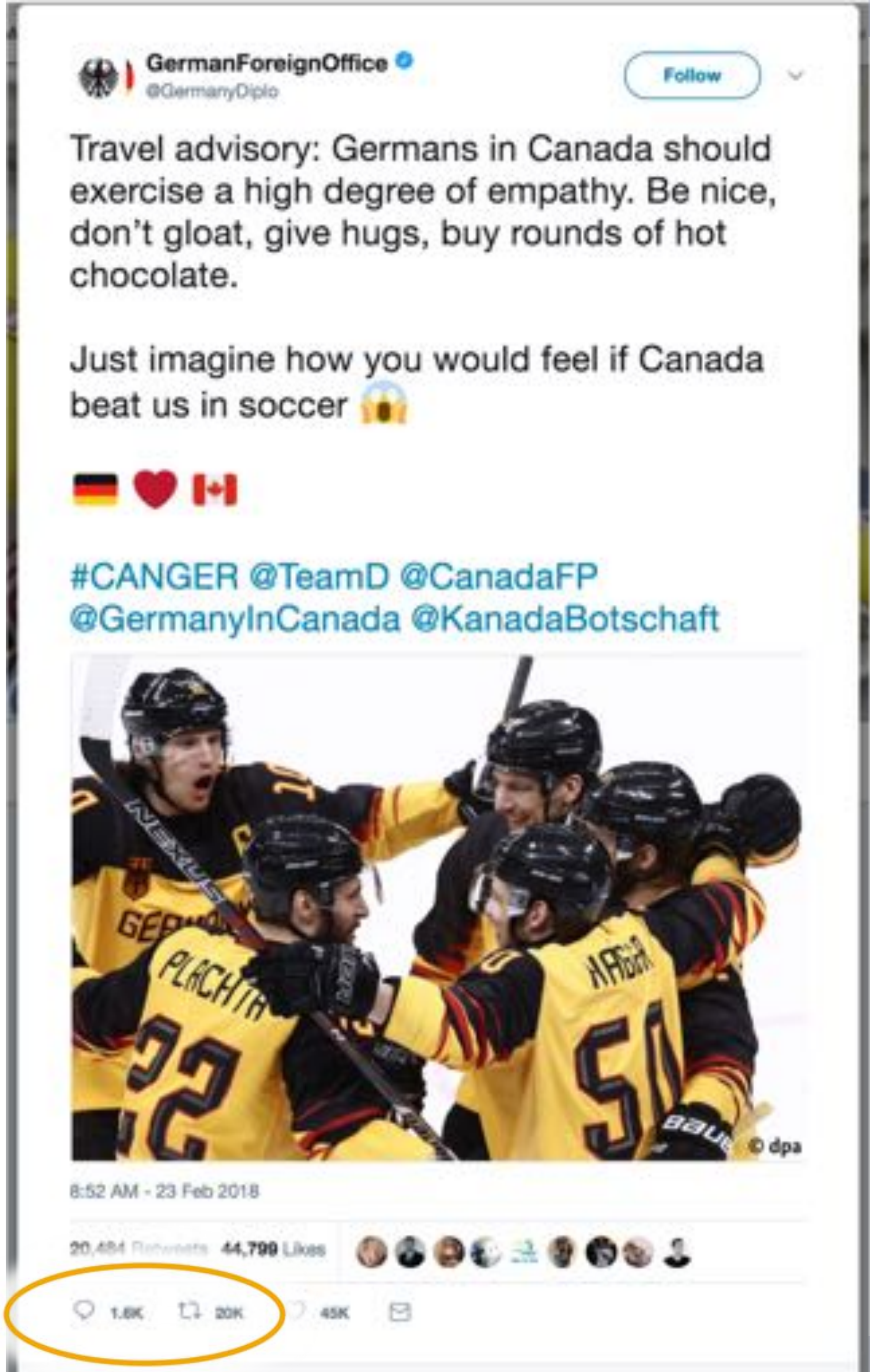
(Scheffler 2014)

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



GermanForeignOffice
@GermanyDiplo

Travel advisory: Germans in Canada should exercise a high degree of empathy. Be nice, don't gloat, give hugs, buy rounds of hot chocolate.

Just imagine how you would feel if Canada beat us in soccer 😱

🇩🇪❤️🇨🇦

#CANGER @TeamD @CanadaFP @GermanyInCanada @KanadaBotschaft

© dpa

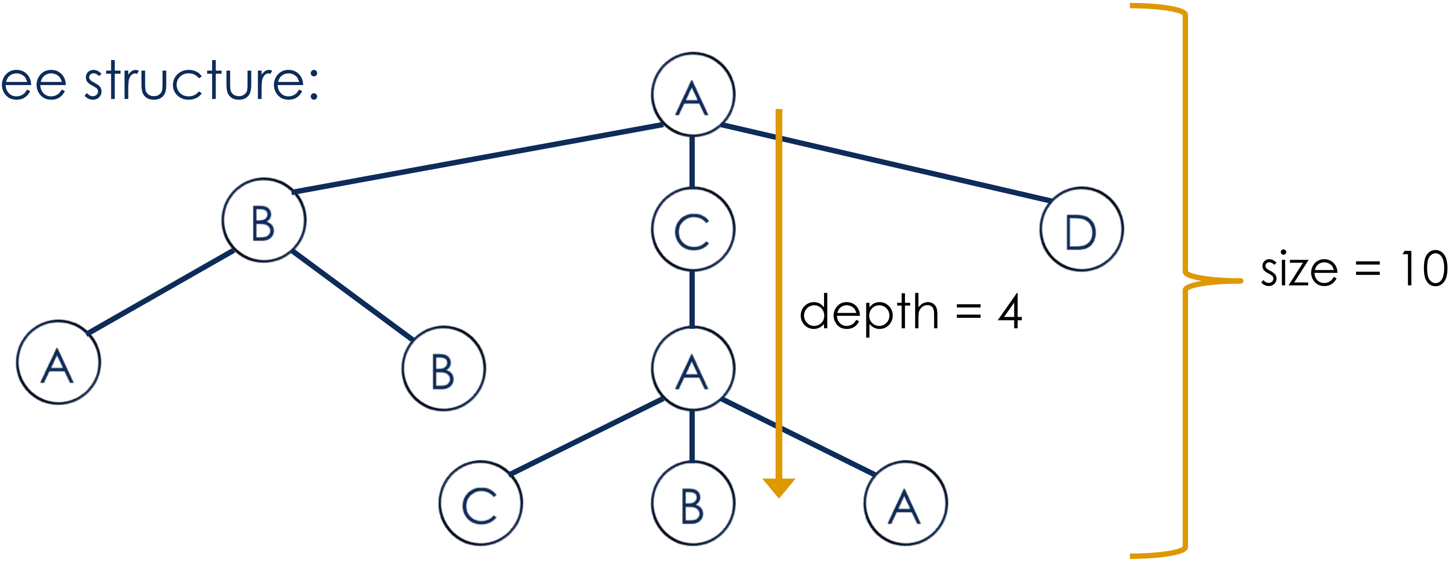
8:52 AM - 23 Feb 2018

20,484 Retweets 44,799 Likes

- What are the answers like?
- Is the conversation:
  - emotional?
  - deliberative?
  - information-seeking?
  - fair?
  - biased?
  - diverse?
- Is the dialog structure parallel to standard spoken schemas?
- What linguistic means are used to indicate it?

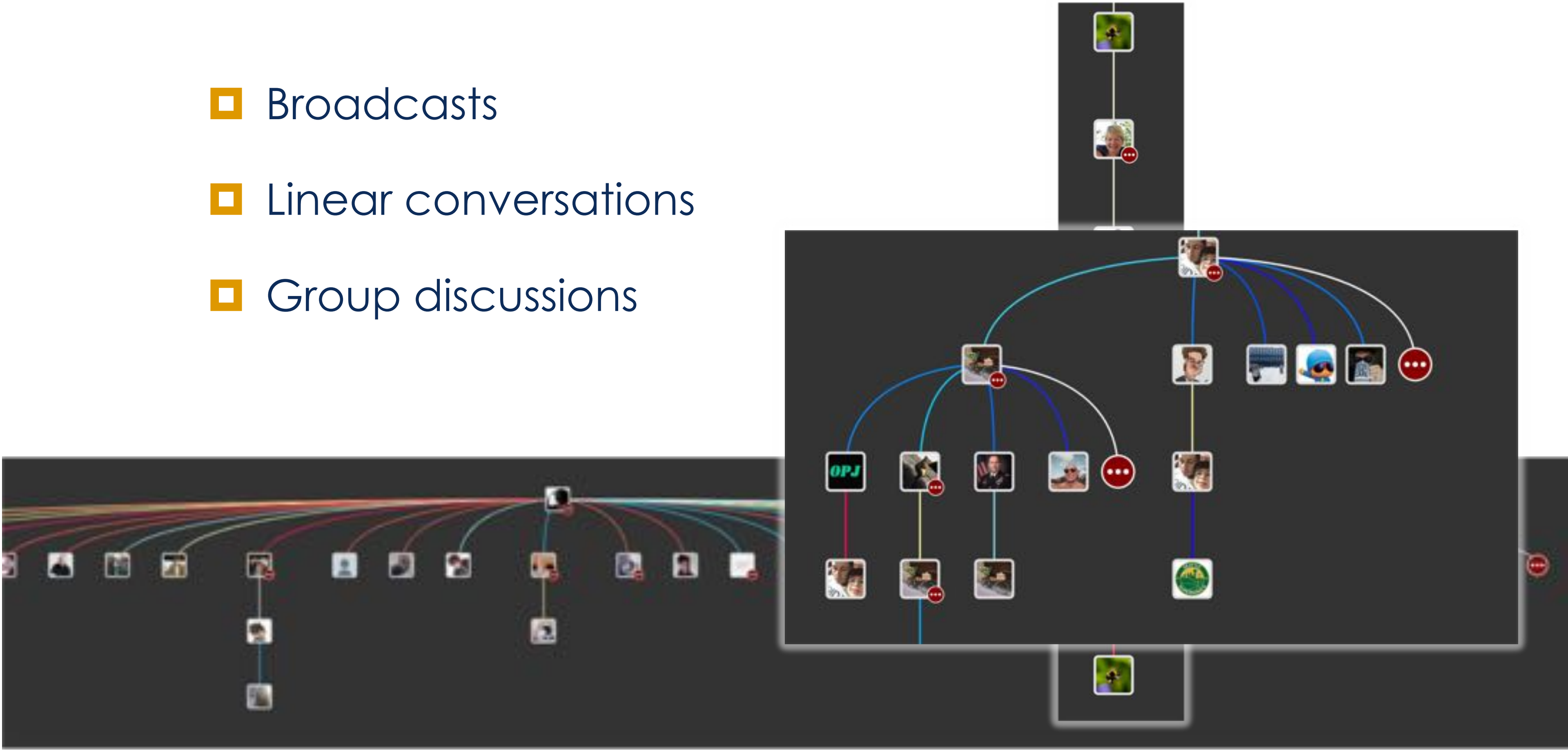
Microblogs = conversations

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

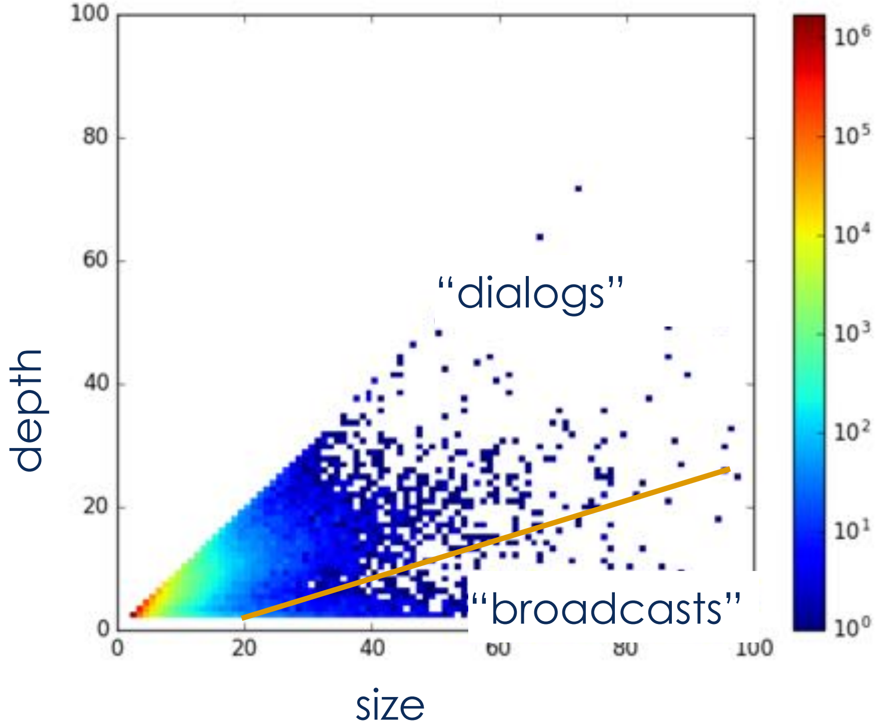


Types of Twitter conversations

- Broadcasts
- Linear conversations
- Group discussions



Types of conversations

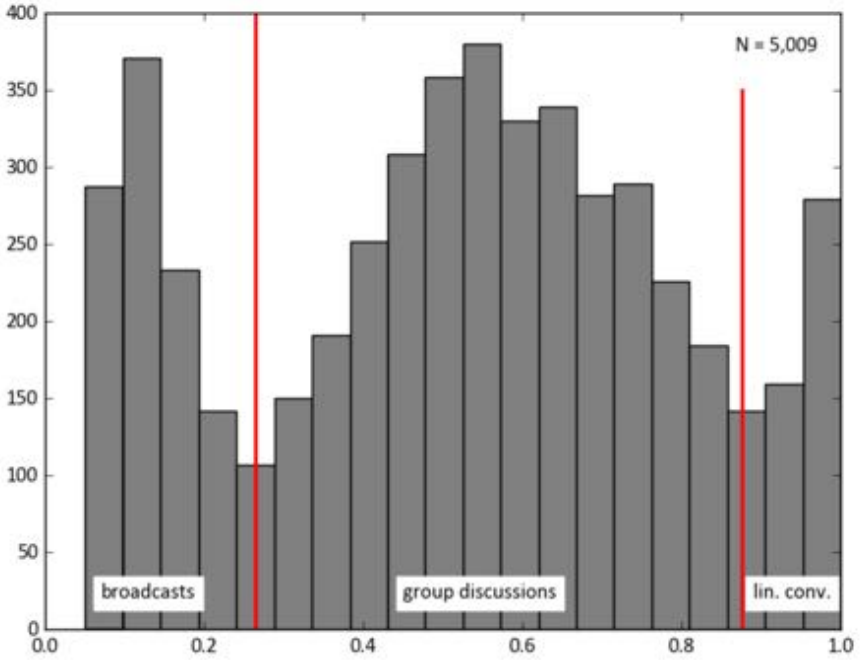


(Scheffler 2017)

Conversation type analysis

- Angle z in the size/depth-plot:

$$z(x) = \frac{4}{\pi}\arctan\left(\frac{\mathrm{depth}(x)}{\mathrm{size}(x)}\right)$$



(Scheffler 2017)

Sample Datasets

- TAGS output:

`http://bit.ly/2FSFvTX`

- Hockey thread, json format:

https://bit.ly/2YERjhD

- Hockey thread, tagged:

https://bit.ly/2XWMGyA

(pw: hch2019)

Part 2 – Tools and Case Studies

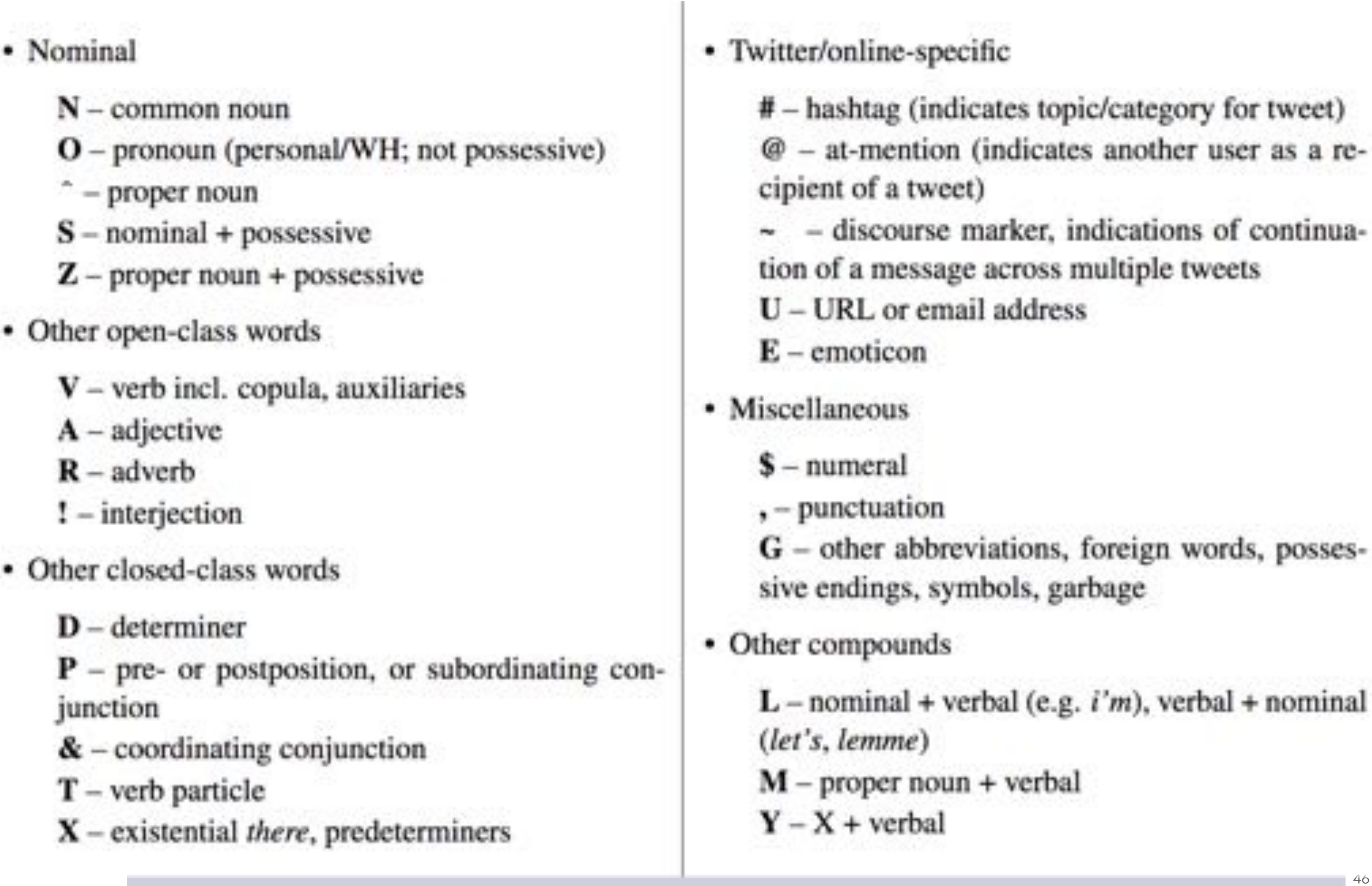
Pre-Processing

Tokenization & Tagging

- Tokenization: finding word boundaries
- Part of speech tagging: tagging word classes
- TweetNLP: standalone project (Gimpel et al., 2011)

@GermanyDiplo @TeamD @CanadaFP @GermanyInCanada @KanadaBotschaft I'll take 2 cups and a hug please . :) Congrats on the win , you deserved it . 👍

@ @ @ @ @ L V $ N & D N V , E ! P D N , O V O , E



- Nominal

  N – common noun
  O – pronoun (personal/WH; not possessive)
  ^ – proper noun
  S – nominal + possessive
  Z – proper noun + possessive

- Other open-class words

  V – verb incl. copula, auxiliaries
  A – adjective
  R – adverb
  ! – interjection

- Other closed-class words

  D – determiner
  P – pre- or postposition, or subordinating conjunction
  & – coordinating conjunction
  T – verb particle
  X – existential *there*, predeterminers

- Twitter/online-specific

  # – hashtag (indicates topic/category for tweet)
  @ – at-mention (indicates another user as a recipient of a tweet)
  ~ – discourse marker, indications of continuation of a message across multiple tweets
  U – URL or email address
  E – emoticon

- Miscellaneous

  $ – numeral
  , – punctuation
  G – other abbreviations, foreign words, possessive endings, symbols, garbage

- Other compounds

  L – nominal + verbal (e.g. *i'm*), verbal + nominal (*let's*, *lemme*)
  M – proper noun + verbal
  Y – X + verbal

TweetNLP

- http://www.cs.cmu.edu/~ark/TweetNLP/
- Run on a text file (one tweet per line):

```
./runTagger.sh --no-confidence inputfile >
outputdir
```

- Import output into Excel (for example)

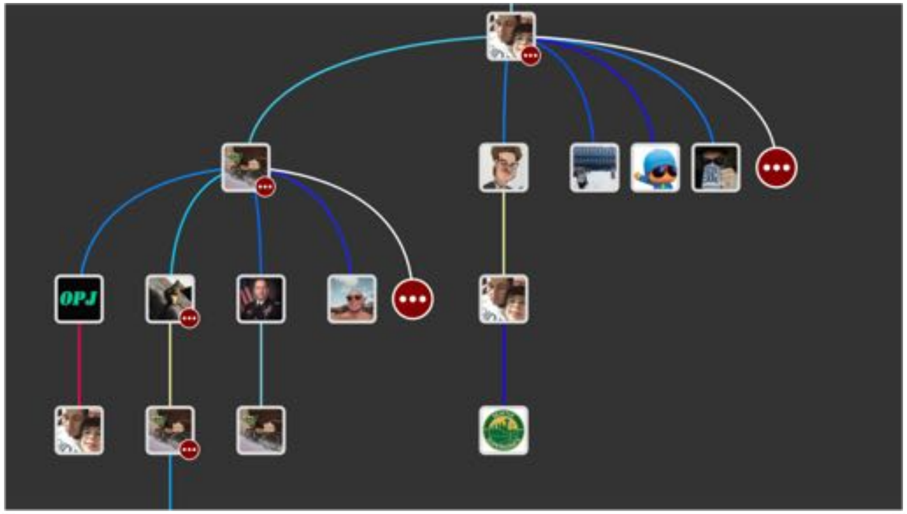
File > Import > Text file > delimited (UTF-8!) > Tab separated

Twitter & Social Media Tools

- http://www.tweepy.org/
- German:
  - tokenizer: https://pypi.python.org/pypi/SoMaJo
  - tagger (not sm specific): http://www.clips.ua.ac.be/pages/pattern-de

Visualization

- Twarc / TreeVerse
- https://github.com/paulgb/Treeverse
- Google Chrome extension
- Visualize conversations



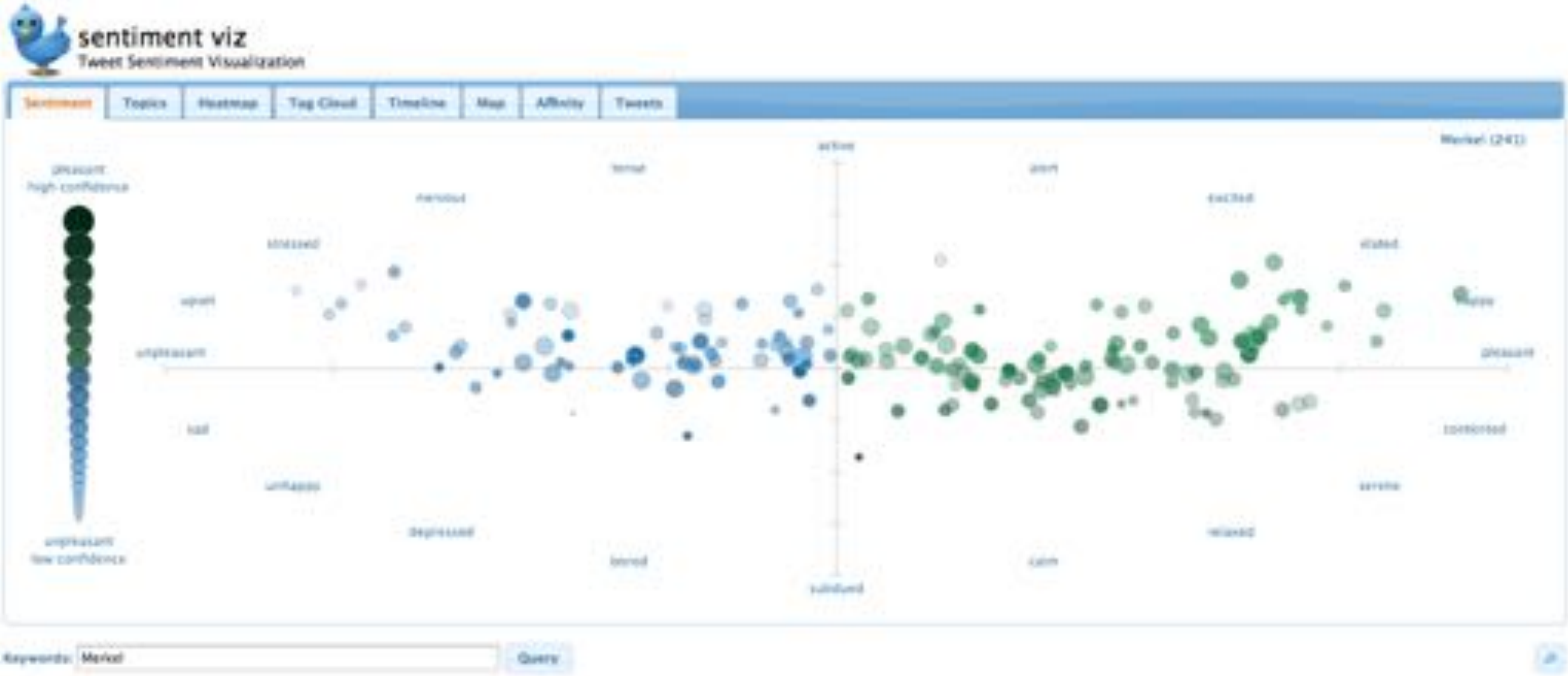
Sentiment Analysis

Sentiment Analysis

> WTF? **I** have green energy and have to co-finance coal and nuclear? What nonsense. WHAT NONSENSE!

- Finding subjective utterances
  - opinion
  - target of opinion
  - source of opinion (attitude holder)
- Corpus annotation of training data
- Machine learning (e.g., based on words used)

SentiViz



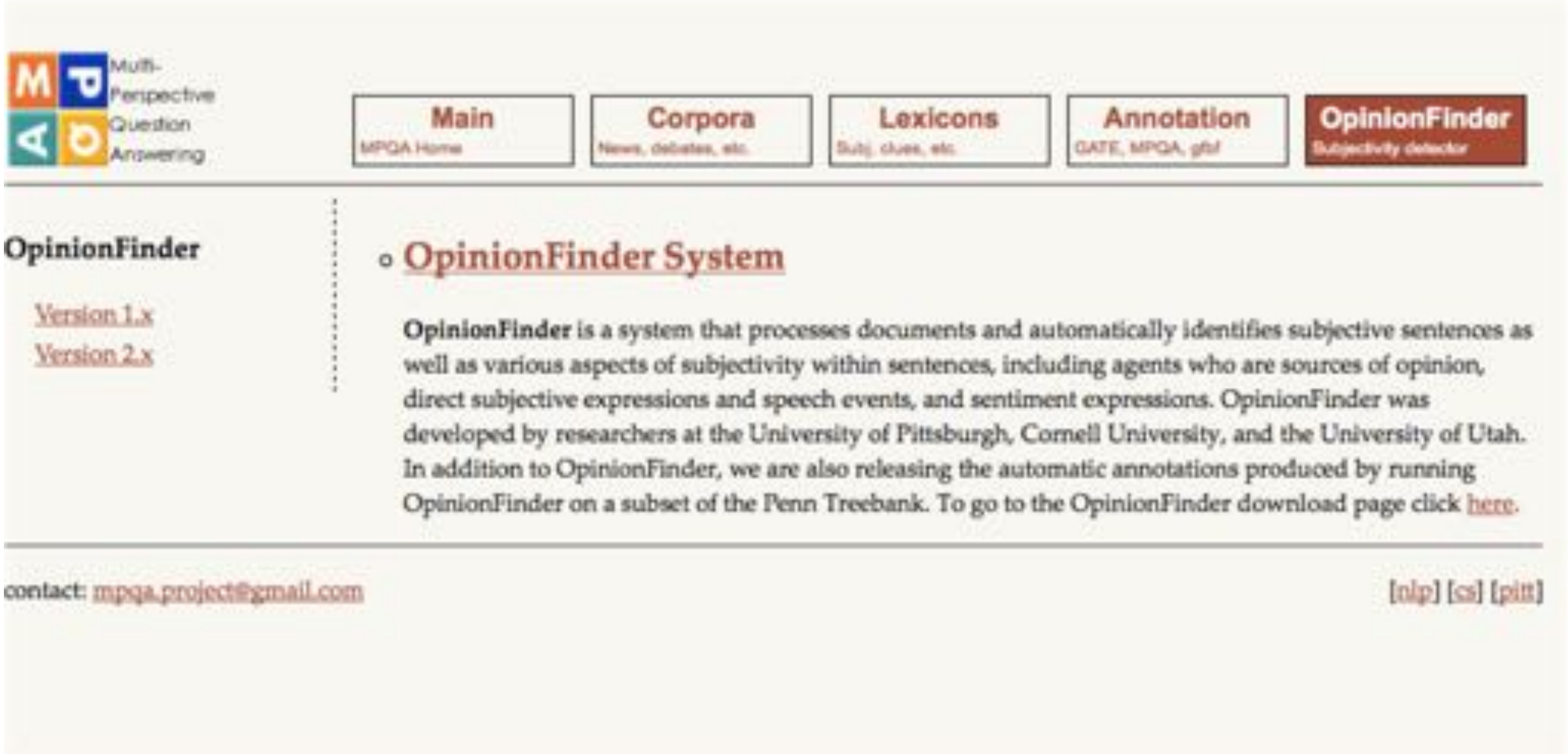
Keywords: Merkel

http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Sentiment Analysis Systems

- OpinionFinder (Wiebe et al., 2005)
  - Java program
- SentiStrength (Thelwall et al., 2010)
  - Windows program (Java version can run on any system)
  - `http://sentistrength.wlv.ac.uk/`
- SoCal (Taboada et al., 2011)
  - Python program (can be run from command line)
  - Needs Stanford CoreNLP
  - `https://github.com/sfu-discourse-lab/SO-CAL`

OpinionFinder



OpinionFinder

Version 1.x

Version 2.x

○ OpinionFinder System

OpinionFinder is a system that processes documents and automatically identifies subjective sentences as well as various aspects of subjectivity within sentences, including agents who are sources of opinion, direct subjective expressions and speech events, and sentiment expressions. OpinionFinder was developed by researchers at the University of Pittsburgh, Cornell University, and the University of Utah. In addition to OpinionFinder, we are also releasing the automatic annotations produced by running OpinionFinder on a subset of the Penn Treebank. To go to the OpinionFinder download page click here.

contact: mpqa.project@gmail.com

[nlp] [cs] [pitt]

http://mpqa.cs.pitt.edu/opinionfinder/

Emoji

Resources on Emoji

- Sentiment of Emoji: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144296
- MoJiSem: Varying linguistic purposes of emoji in (Twitter) context (ACL Student Research Workshop 2017) http://www.aclweb.org/anthology/P17-3022
- http://emojitracker.com/
- https://emojipedia.org/

Other Tools:

- Great Python introduction:
  - http://greenteapress.com/wp/think-python-2e/
- Unix for Poets (command line interface):
  - https://web.stanford.edu/class/cs124/kwc-unix-for-poets.pdf
- NLTK (natural language toolkit) package for Twitter:
  - http://www.nltk.org/howto/twitter.html

Questions?

tatjana.scheffler@uni-potsdam.de