2. Twitterdaten – Korpuserstellung

Seminar

Computerlinguistische Analyse von Twitterdaten

Tatjana Scheffler, Universität Potsdam

tatjana.scheffler@uni-potsdam.de

17.4.2013

Heute

- Deutsche Twitterdaten / Korpuserstellung
- Weitere Seminarplanung:

  Themenbesprechung

  Interessensbekundungen

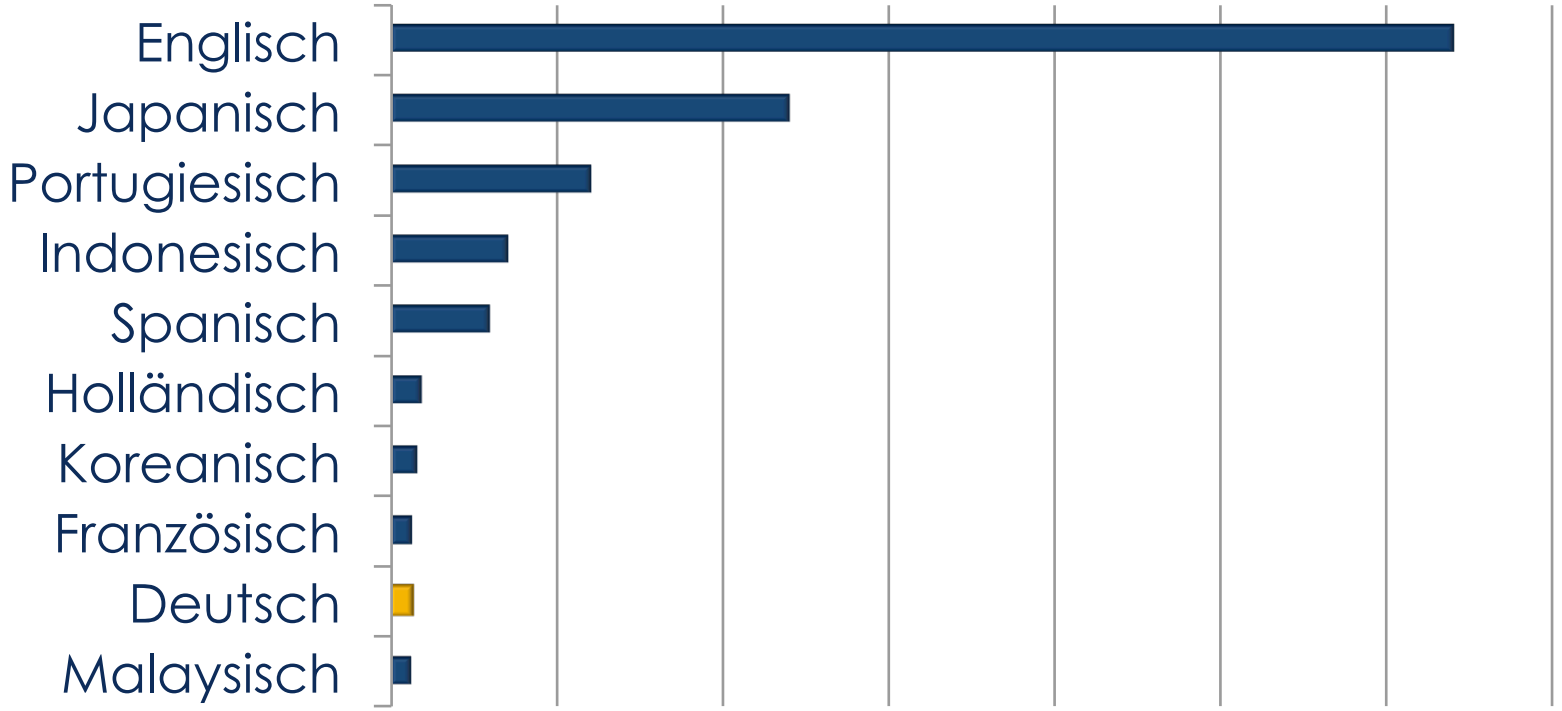
Hausaufgabe

- Erstelle einen Twitteraccount
- Folge mind. 5 Leuten, schreibe mind. 20 Tweets
- Zugriff auf den Twitter-Stream
  - Folge der Anleitung auf:
    http://www.ling.uni-potsdam.de/~scheffler/twitter/index.html
  - Wähle ein oder mehrere Stichwörter (keywords.txt)
  - Speichere die Tweets zu diesem Thema für eine Zeit
  - Schreibe während der Zeit selbst einen Tweet mit dem Suchwort

Erstellung eines deutschen Twitterkorpus

Probleme, Vorgehensweise

Sprache auf Twitter

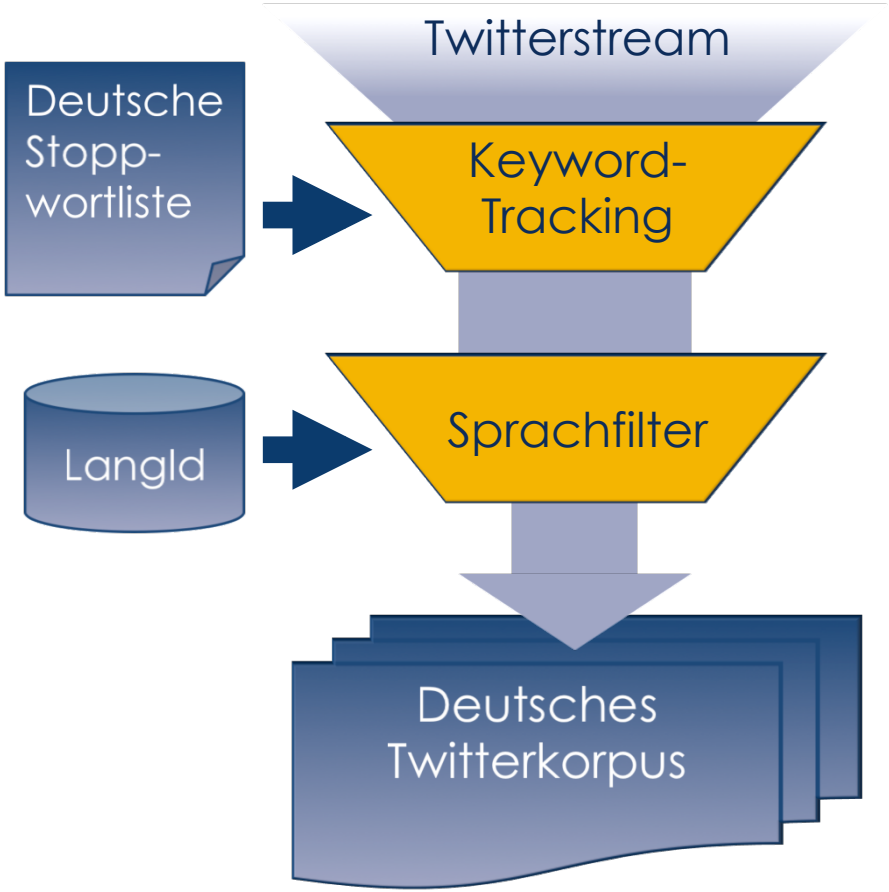


Quelle: Hong, Lichan, Convertino, Gregorio, and Chi, Ed. "Language Matters In Twitter: A Large Scale Study" International AAAI Conference on Weblogs and Social Media (2011)

Twitter-API zur Korpuserstellung

- Search API oder Streaming API
- Search API: Suchworte, ca. 7 Tage in die Vergangenheit
- Streaming API:
  - Echtzeitstream der entstehenden Tweets
  - Quotenlimitierung
  - Viele nicht-deutsche Tweets
  - Filter
    - Geolokation (location) – nur ca. 2% der dt. Tweets
    - bis zu 5000 User-Ids (follow)
    - bis zu 400 Stichwörter (track)

Korpuserstellung



~ 500.000.000 Tweets / Tag

~ xx.000.000 Tweets / Tag

~ 1.500.000 Tweets / Tag

Bisher verfügbare Daten:

2.-12. Dez. 2011	ca. 4,5 Mio. Tweets
20.-27. Dez. 2011	ca. 4 Mio. Tweets
Dez./Jan. 2013	> 8 Mio. Tweets
Mär./Apr. 2013	21 Mio. Tweets ↗

Tools: Twitterstream mitschneiden

1. Python-Paket: tweepy `https://github.com/tweepy/tweepy`
2. Eigene Anwendung bei Twitter registrieren und Access/ Consumer Keys erhalten
3. Wortliste der mitzuschneidenden Stichwörter erstellen
   - Z.B.: Filtere Stream nach 397 häufigen deutschen Wörtern
   - Ausschluss von fremdsprachigen Homographen: “war”, “die”, “des”, …
   - Verlust nur ca. 2-5% der deutschen Tweets
4. Twitter für Linguisten-Paket Twython starten
   `http://www.ling.uni-potsdam.de/~scheffler/twitter/`

Sprachidentifikation

- Twitter-eigene Sprachklassifikation ist zu inakkurat; scheint auf Eigenschaften im User-Profil zu basieren
- Google Compact Language Detector:

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

Deutsche Tweets	Langid	Google CLD	Twitter
Präzision	97%	96%	~ 40%

Twitterdaten als Korpus

- Enthält spezielle Tokens (Emoticons, URLs, # Hashtags)
- Umgangssprache, Slang und Dialekte
- **Vorverarbeitung ist wichtig:**
  - Normalisierung (Umlaute, Prolongationen, Tippfehler?)
  - Behandlung von Spezialtokens (@Handles, #Tags)
  - Tokenisierung
  - Satzgrenzenbestimmung

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

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

Twitter Terms of Service – Probleme

- **Keine Weitergabe von aggregierten Tweets (=Korpus) erlaubt**
- Korpusweitergabe nur über Tweet-IDs möglich; einzelne Tweets müssen dann zeitaufwändig wieder gecrawlt werden, z.B. mit `https://github.com/lintool/twitter-tools`
- Löschung von Tweets und/oder Accounts: 21,2% des Tweets2011-Korpus verschwanden in den ersten 9 Monaten
- Anonymisierung von Tweets in Papieren
  - @Handles entfernen
  - Trotzdem auffindbar

Weitere Seminarplanung

Sie sind gefragt!

Planung

Heute: Vorverarbeitung (TS)

24.4. Technisches (TS)

1.5. Tag der Arbeit (kein Seminar)

8.5.	29.6.	Vorträge
15.5.	5.6.	
22.5.	12.6.	
19.6.	26.6.	Projektarbeit

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