Corpora from the blogosphere: why and how?

CLARIN-PLUS Workshop "Creation and Use of Social Media Resources"

Adrien Barbaresi

Austrian and Berlin-Brandenburg Academies of Sciences

18 May 2017

Outline

Among the most frequent n-grams on Reddit

Phatic expressions

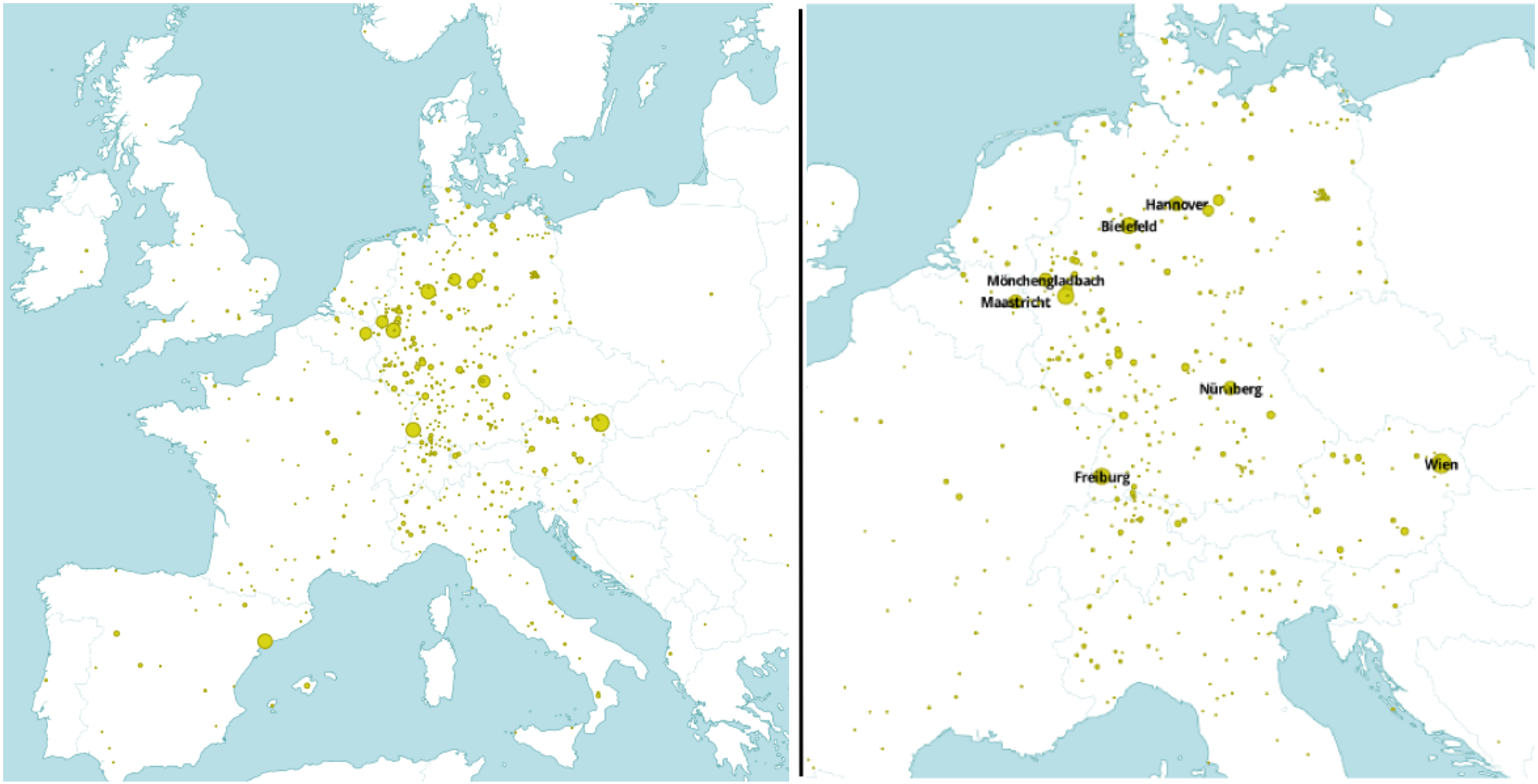
- Danke schön
- Ausgezeichnet!
- Viel Glück!

Germany as seen by Americans? (e.g. in films)

- Arbeit macht frei
- Mein Kampf
- Papiere bitte

Proper nouns

- Arjen Robben
- Kurt Vonnegut


Hannover
Bielefeld
Mönchengladbach
Maastricht
Nürnberg
Freiburg
Wien

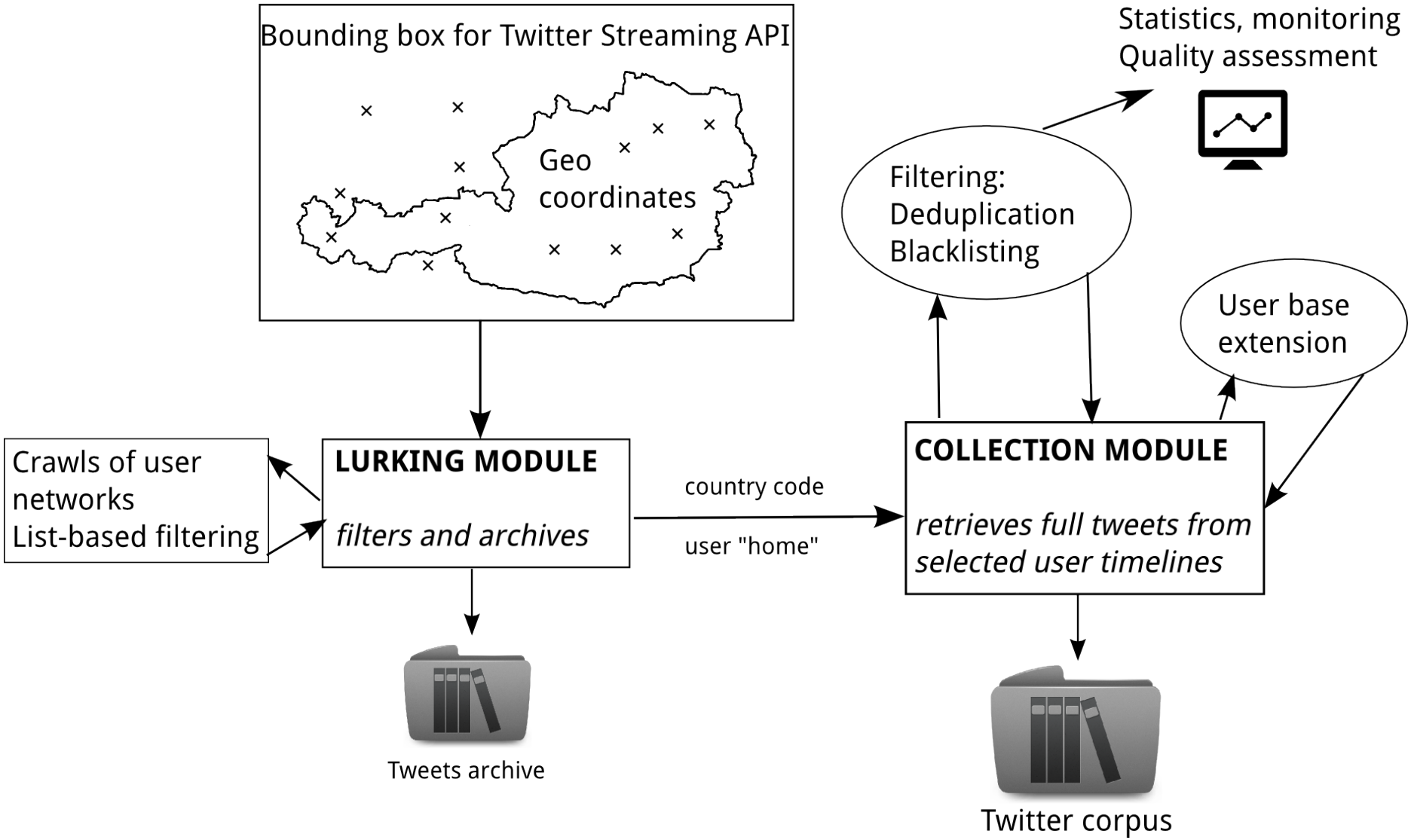
Conclusions on Reddit

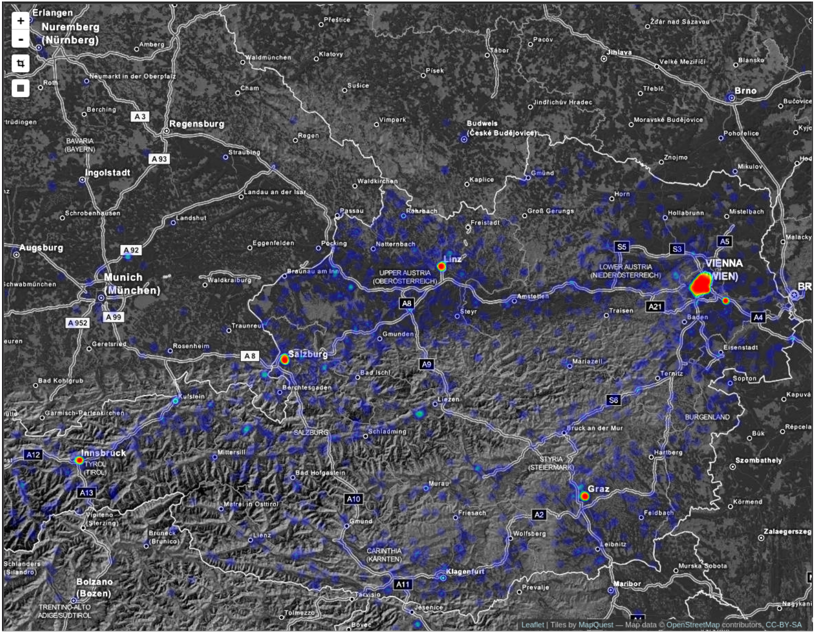
- A (small) corpus focusing on German can be built using the publicly available Reddit comment dataset
- The structural properties of the corpus are in line with the expectations concerning CMC
- Visualization of place names shows an interesting distribution

Reddit is a well-established platform, currently gaining traction in Germany
regular updates announced by Jason Baumgartner
$\rightarrow$ there could soon be more text

Barbaresi A. (2015): Collection, Description, and Visualization of the German Reddit Corpus. In: *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication, GSCL Tagung*, Essen, pp. 7-11.

Twitter: focusing on users related to Austria




Erlangen
Nuremberg (Nürnberg)
Amberg
Přeštice
Waldmünchen
Klatovy
Tábor
Pacov
Jihlava
Žďár nad Sázavou
Velké Meziříčí
Blansko
Brno
Roth
Neumarkt in der Oberpfalz
Cham
Sušice
Písek
Třebíč
Jindřichův Hradec
Berching
A 3
Regensburg
Vimperk
Budweis (České Budějovice)
Moravské Budějovice
Pohořelice
BAVARIA (BAYERN)
Regen
Straubing
A 93
Znojmo
Mikulov
Ingolstadt
Waldkirchen
Kaplice
Gmünd
Landau an der Isar
Horn
Schrobenhausen
Landshut
Passau
Rohrbach
Freistadt
Groß Gerungs
Hollabrunn
Mistelbach
Augsburg
A 92
Eggenfelden
Pocking
Natternbach
Linz
S5
S3
A5
VIENNA (WIEN)
Munich (München)
Waldkraiburg
Braunau am Inn
UPPER AUSTRIA (OBERÖSTERREICH)
LOWER AUSTRIA (NIEDERÖSTERREICH)
A8
Amstetten
A21
A 952
A 99
Steyr
Traisen
Baden
A4
Traunreut
Gmunden
Eisenstadt
Geretsried
Rosenheim
A 8
Salzburg
A9
Mariazell
Bad Ischl
Ternitz
Sopron
Bad Kohlgrub
Berchtesgaden
Kufstein
Liezen
S6
Garmisch-Partenkirchen
BURGENLAND
SALZBURG
Schladming
Bruck an der Mur
A12
Innsbruck
TYROL (TIROL)
Mittersill
STYRIA (STEIERMARK)
Hartberg
Szombathely
Bad Hofgastein
Murau
A13
A10
Graz
Körmend
Matrei in Osttirol
Friesach
A2
Feldbach
Vipiteno (Sterzing)
Gmünd
Wolfsberg
Bruneck (Brunico)
Lienz
Leibnitz
Zalaegerszeg
CARINTHIA (KÄRNTEN)
Schlanders (Silandro)
Klagenfurt
Murska Sobota
Bolzano (Bozen)
A11
Maribor
Tarvisio
Prevalje
TRENTINO-ALTO ADIGE/SÜDTIROL
Jesenice
Tolmezzo
Bovec
Leaflet | Tiles by MapQuest — Map data © OpenStreetMap contributors, CC-BY-SA

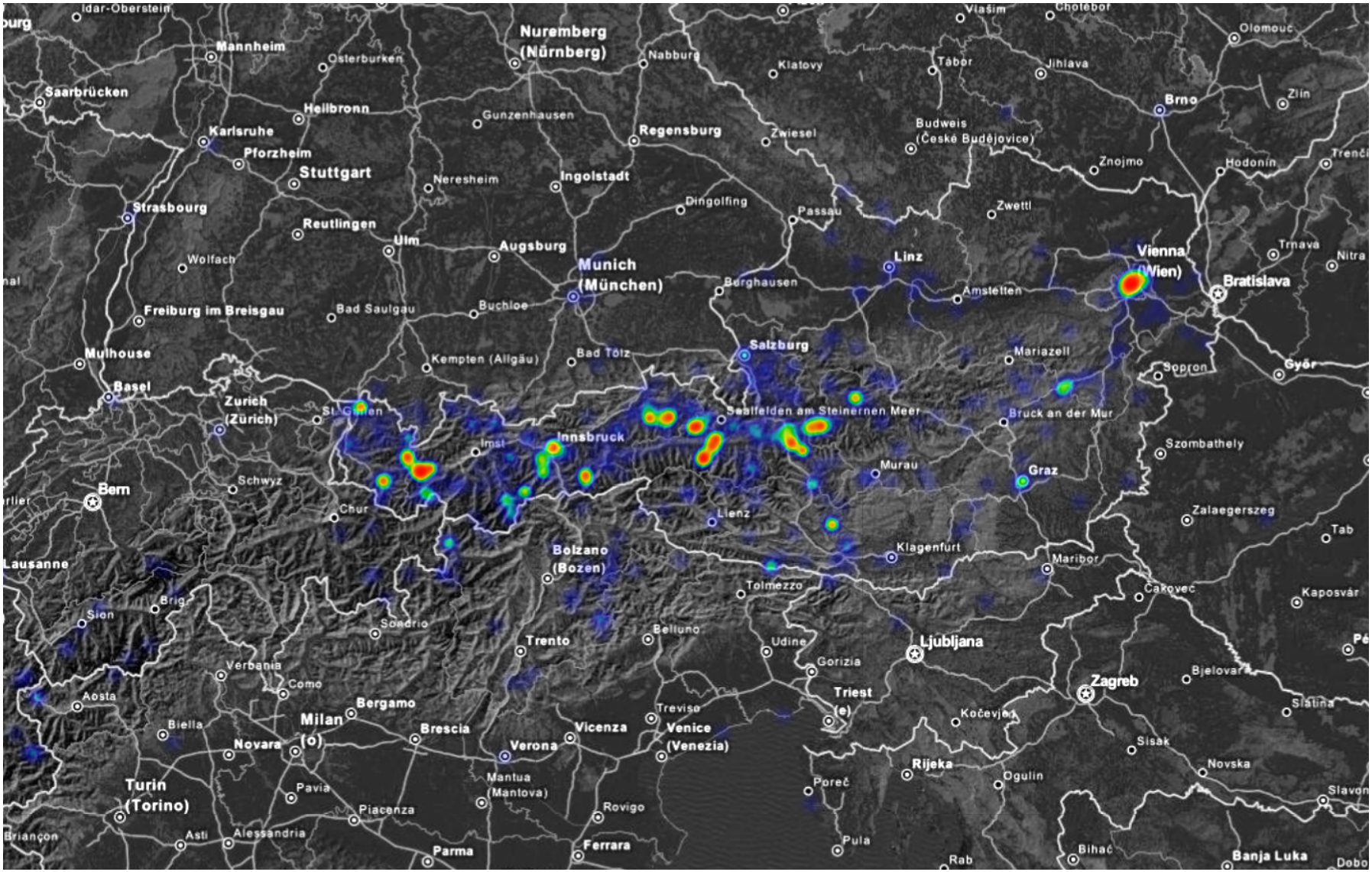


Figure: Query *text:ski**

Indexing and querying of the corpus

Design decisions to focus on users related to Austria

Implementation: Elasticsearch + Kibana

Barbaresi, A. (2016): Collection and Indexing of Tweets with a Geographical Focus. *LREC 2016 – Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC)*, pp. 24-27.

Blogs

"A reverse chronological sequence of dated entries" (Kumar et al. 2003)

"The cross-linking that takes place between blogs, through blogrolls, explicit linking, trackbacks, and referrals has helped create a strong sense of community in the weblogging world." (Glance et al. 2004)

Problems to solve

1. Which blogs and where?

2. Which information should we extract?

The chosen solution and its advantages

A software (WordPress) & its platform (wordpress.com)

wordpress.com: potentially more than 1,350,000 blogs in German

+ all the self-hosted websites using WordPress (approx. 1/4 worldwide)
  - Host diversity = various user profiles?
  - Same software

⇒ Comparable if not same content **structure**
⇒ Potentially identical extraction procedures

For a fistful of blogs...

Focused crawling of wordpress.com, German part

158,719 blogs found

the unseen rest: no links/tags, closed, crawl not long enough

Result of blog scans (homepage $>$ 5 posts)

- 12.7 % with comments (20,181)
- 0.8 % under CC license at best (1,201)
- 0.2 % with comments and under CC license (324)

Barbaresi A. & Würzner K.M. (2014) For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In: *Proceedings of NLP 4 CMC Workshop, KONVENS*, Hildesheim, 2014. pp. 2-10.

Last but not least: License issues

Statistics on German blogs on wordpress.com

158,719 blogs found, results of scans on homepage ($>$ 5 posts):

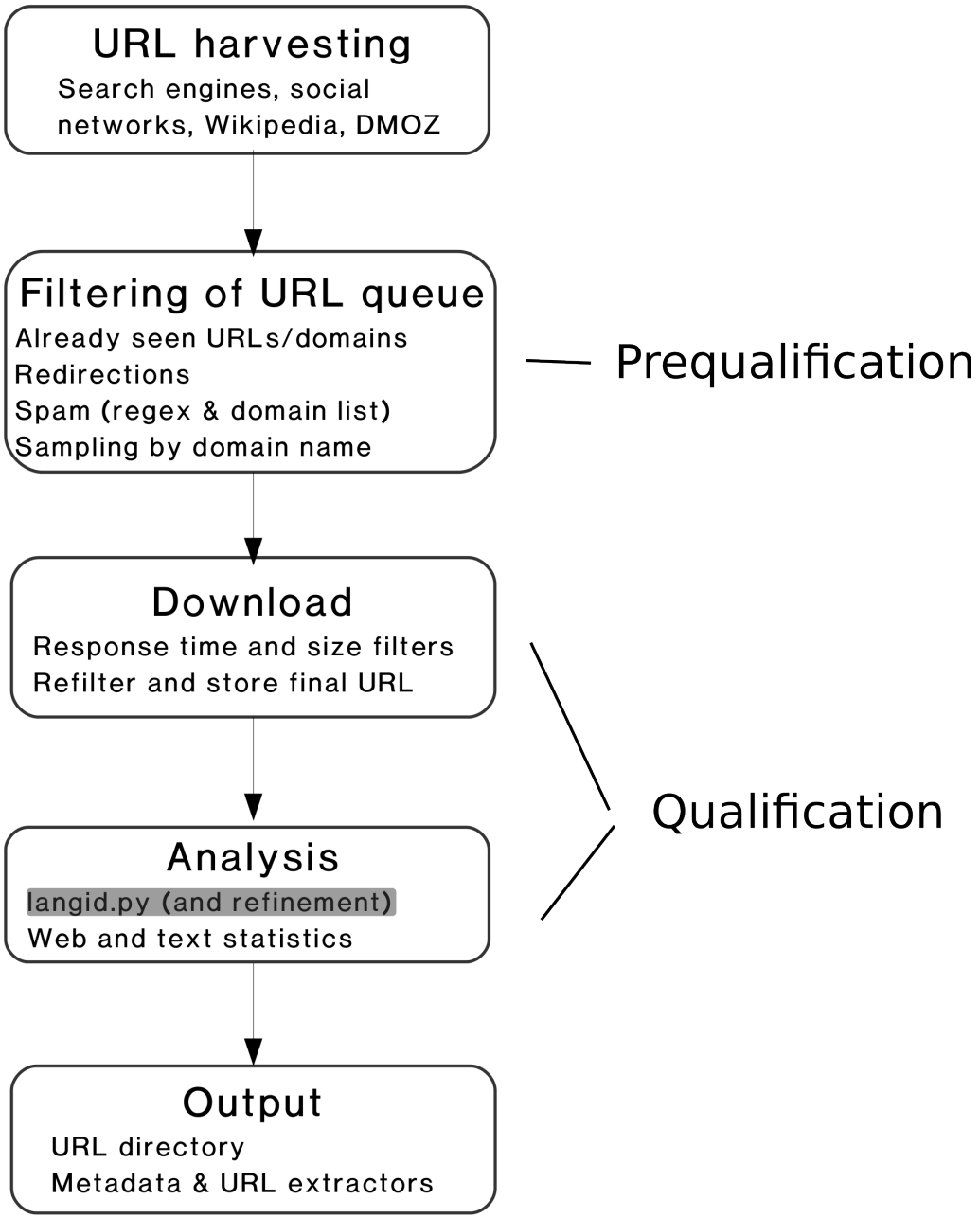
- 0.8 % under CC license at best (1,201)

Most frequent licence types

652	BY-NC-SA
532	BY-NC-ND
351	BY-SA
282	BY
129	BY-NC
58	BY-ND

NC	non-commercial
ND	no derivative works
SA	share alike

$\rightarrow$ The usage of *-ND is a problem


URL harvesting
Search engines, social networks, Wikipedia, DMOZ
Filtering of URL queue
Already seen URLs/domains
Redirections
Spam (regex & domain list)
Sampling by domain name
Prequalification
Download
Response time and size filters
Refilter and store final URL
Analysis
langid.py (and refinement)
Web and text statistics
Qualification
Output
URL directory
Metadata & URL extractors

Extraction

1. HTML parse
2. subtree selection with XPATH-expressions
3. tag conversion and pruning
4. output in XML TEI format

Targets

- Title of post, title of blog, date of publication, canonical URL, author, categories, and tags
- Posts // Comments
  (text structuration: titles, paragraphs, bold and italic, no links)

⇒ Extraction as proxy for quality assessment:
Full duplicates, short documents, and documents without date removed

DMHDF – DAS BLOG DER MITTELMÄSSIGKEIT

ÜBER…

15 FEB

Tweets der KW 7 (2014)

Hallo alle Miteinander,

hier meine Tweets aus dieser Woche @dmhadf:

#Begriff: #Volksmehr bei WP http://de.wikipedia.org/wiki/Volksmehr #Mehrheit der gültig stimmenden Personen in einem Schweizer fakultativen Referendum

#Person: Toni Brunner bei WP http://de.wikipedia.org/wiki/Toni_Brunner einflussreicher Schweizer Politiker der #SVP, #gegenmasseneinwanderung

#Kommunalpolitik: Kritik an Krampnitz-Plänen in #Potsdam http://www.pnn.de/potsdam/827212/ Was wird aus der ehemaligen Krampnitz-Kaserne?

#Meinung: Eine Zäsur für die #Schweiz http://www.nzz.ch/meinung/kommentare/eine-zaesur-fuer-die-schweiz-1.18239317 Was die #Volksmehr bei der letzten Volksinitiative bewirkt. #Einwanderung

#Begriff: #Kurantmünze bei WP http://de.wikipedia.org/wiki/Kurantm%C3%BCnze Münze, die durch ihren Metallwert gedeckt ist. Nicht mehr verwendet #Geld

#Organisation: #Reichsbank (im dt. Reich) bei WP http://de.wikipedia.org/wiki/Reichsbank Zentralnotenbank des dt. Reiches, wurde von privaten finanziert.

#Kommunalpolitik: Neue (alte) Schule in #Hennigsdorf geplant http://www.moz.de/heimat/lokalredaktionen/hennigsdorf/hennigsdorf-artikel/dg/0/1/1245986/ Vorbereitungen laufen, einige Fragen ungeklärt.

ARCHIV

- Februar 2014
- Januar 2014
- Dezember 2013
- November 2013
- Oktober 2013
- September 2013
- August 2013
- Juli 2013
- Juni 2013
- Mai 2013
- April 2013
- März 2013
- Februar 2013
- Januar 2013
- Dezember 2012
- November 2012
- Oktober 2012
- September 2012
- August 2012
- Juli 2012

Follow

essen feinheiten fukushima führerschein geek getränk golden week grammatik heimat herbst hitachi japan japanisch kultur neujahr off-topic picknick radtour regen reise sakura sprache start taifun technik tokio tokyo travel trinken tsunami tsuyu unfall uni video vlog werbung wetter winter

Kategorien

- Anfang
- Daheim
- Deutsch
- English
- Essen
- Gesundheit
- Japanische Getränke
- Kanji
- myNichi / マイ日
- Nur in Japan
- Probleme
- Reise
- Sport
- Sprache
- Sprachliche Feinheiten
- Start
- Technik
- Tokio
- Tokyo
- Tradition
- Travel
- Trinken
- Uncategorized
- Video
- Werbung

JAN 06 2014

日本にコンニ Vlog 11: Jigokudani

Über Weihnachten und Neujahr habe auch ich mir (wie alle Japaner) ein paar Tage freigenommen. Da ich ausserdem hohen Besuch aus Deutschland hatte, hatte ich mir ein paar besondere Sachen ausgedacht, die unter anderem eine kleine inner-japanische Reise beinhalteten. Einen Ausschnitt davon möchte ich euch in diesem Vlog gerne zeigen.

Hier noch ein paar nützliche Informationen und Links zu dem Video:
Homepage des Parks (auf Englisch und Japanisch): http://www.jigokudani-yaenkoen.co.jp/
Google Maps des Parks: http://bit.ly/jigokudani
Fahrpläne der Nagaden (長野電鉄):
- von 長野駅 (Nagano Hbf.) nach 湯田中温泉 (Yudanaka Onsen)
- und zurück von 湯田中温泉 (Yudanaka Onsen) nach 長野駅 (Nagano Hbf.)
Fahrpreis dieser Verbindung: ¥1.130 + ¥100 Expresszuschlag (one-way)
Fahrplan des Busses zwischen 湯田中温泉 (Yudanaka Onsen) und 上林温泉 (Kanbayashi Onsen)
Fahrpreis dieser Verbindung: ¥280 (one-way)
Der Fußweg ab der Bushaltestelle 上林温泉 (Kanbayashi Onsen) ist sehr gut ausgeschildert aber nicht geräumt oder gestreut. Gutes (winterliches) Schuhwerk und/oder Spikes sind sehr anzuraten!
Der Eintritt in den Park selbst kostet ¥500.
Und hier nochmal der Link zur Live-Kamera: http://www.jigokudani-yaenkoen.co.jp/livecam/monkey/index.htm

3 Kommentare

DEZ 07 2013

[redacted]

Data so far

Wordpress-AT corpus: 0.5 GTokens
Wordpress-DE corpus: 2.1 GTokens

General AT+CH+DE refined corpus (2016): ca. 3 GTokens from 6,978,183 pages and ca. 200,000 different domains

Barbaresi, A. (2016): Efficient construction of metadata-enhanced web corpora. *Proceedings of the 10th Web as Corpus Workshop*, Association for Computational Linguistics, pp. 7-16.

Conclusions on the experiments

All blogs in a formal sense, but strong differences

Typological gap between original and current studies as well as between users of a platform and users of a content management system

Relatively few interlinking and interaction

Sketches the typical profile of a **passive internet consumer, a "prosumer" at best**, which should be taken in consideration

Attempt at a typology

- Blogs **mimic existing text types**, audiences, and motivations, with a focus on information (general, specialized, or community-based) as well as on promotional goals
- Websites whose **finality is to sell** information, entertainment, or concrete products and services

Potential uses

- Comparison with reference corpora (K-M. Würzner and others, BBAW)
- Detection of phrasal compounds (Katrin Hein, IDS Mannheim)
- Mapping of linguistic variation in tweets (Antonio Ruiz Tinoco, Sophia University)
- Used in bachelor's and master's theses (students strongly relate to CMC data)
- other exciting projects??

Thank you for your attention!


barbaresi@bbaw.de


@adbarbaresi


http://adrien.barbaresi.eu/