

Erstellung und Nachnutzbarkeit eines (deutschsprachigen) Reddit-Korpus

Reddit als Datenbasis

Plattformstruktur

- Über 7 Milliarden Aufrufe pro Monat - eines der größten Online-Foren weltweit
- Subreddits als thematisch spezialisierte Communitys

Archivierte Reddit-Daten: Pushshift

- Reddit-Kommentare und -Einreichungen
- Bereitstellung als vollständige Dumps oder Subreddit-spezifische Dateien

Linguistische Relevanz

- Hoher Grad an konzeptioneller Mündlichkeit: informell, dialogisch, spontan
- Netzjargon, Neologismen und (sprachliche) Memes als linguistische Phänomene

Von Pushshift zu TEI-XML

Input

- Pushshift-Archivdateien (.zst)
- mit Reddit-Kommentaren

Schritte der Datenbereinigung

- Entfernen gelöschter Kommentare und Bots
- Bereinigung von URLs, Zitaten, Escape-Zeichen u. Ä.

Konvertierung in TEI-XML

- JSON als Zwischenformat (automatisch generiert)
- Metadaten: Autor, Zeitstempel, Thread-ID, URL etc.
- Gruppierung nach Threads oder Einzelkommentaren

Output

- Bereinigte Reddit-Kommentare inkl. Metadaten
- Reproduzierbare Ergebnisse für eigene Forschungsfragen

```
<TEI>
<text>
  <body>
    <div type="comments">
      <list>
        <item source="https://www.reddit.com/r/Garten/comments/15bgtgi/comment/jtraadp/">
          Wenn es eine Formschnithecke sein soll, empfehle ich Liguster, Hainbuche, Rotbuche oder Eibe. Allesamt schnittverträglich und heimisch, du tust damit also auch der Vogelwelt was Gutes und die ein oder andere Schmetterlingsraupe findet Futter darin.
          <date>2023-07-28</date>
          <name>WWConny</name>
        </item>
        <item source="https://www.reddit.com/r/Garten/comments/15bgtgi/comment/jtrt07z/">
          Für so eine niedrige Hecke, die pflegeleicht sein soll, würde ich eher Kleinwüchsige Pflanzen nehmen.
          <date>2023-07-28</date>
          <name>blackdevilsisland</name>
        </item>
        <item source="https://www.reddit.com/r/Garten/comments/15bgtgi/comment/jtsw34x/">
          Ach ja, Abstand zum Zaun: mindestens(!) 50 cm.
          <date>2023-07-28</date>
          <name>kannsnedsein</name>
        </item>
      </list>
    </div>
  </body>
</text>
</TEI>
```

reddit-d

Datenumfang

- 34 Mio. Kommentare
- 90 Mio. Sätze
- 1,34 Mrd. Tokens
- Zeitraum: 2006-2023
- Daten aus 40 Subreddits mit breitem thematischen Spektrum
- POS-getaggt und lemmatisiert

<https://www.dwds.de/d/korpora/reddit>

Korpusbelege Reddit-d

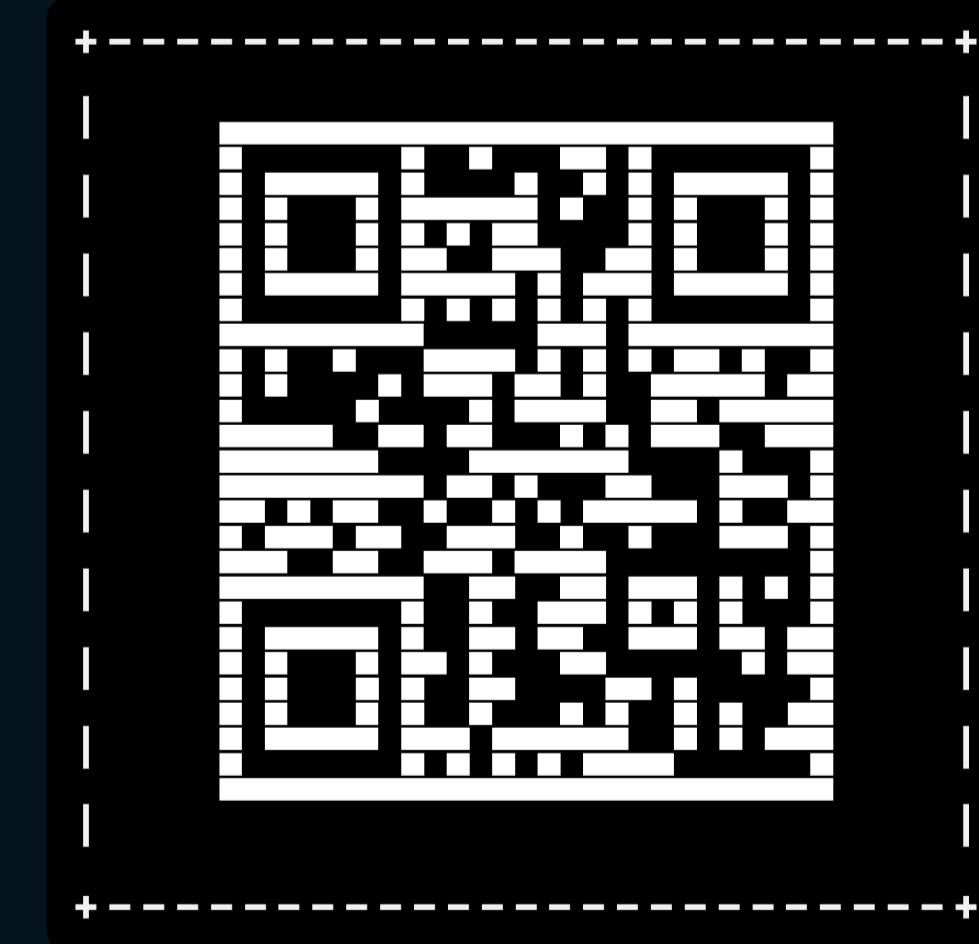
"@{Das das} @eskalierte #1 \$p=ADJ"

Korpus:	Start:	Ende:
Reddit-d	2006	2023
Anzeige:	Sortierung:	Treffer pro Seite:
<input type="radio"/> KWIC <input checked="" type="radio"/> voll <input type="radio"/> maximal	Datum absteigend	50

- 1-50 von 286 Treffern
- | | | | | | | | | | | | |
|-----|----|---|---|---|---|---|---|---|----|-----|---|
| -10 | -5 | ← | 1 | 2 | 3 | 4 | 5 | → | +5 | +10 | → |
|-----|----|---|---|---|---|---|---|---|----|-----|---|
- Reddit/de/jahrelang_wartete_ich_darauf_heute_ist_es_endlich, 2023-12-15
Das eskalierte schnell in eine dunkle Richtung.
 - Reddit/FragReddit/frage_an_die_frauen_was_war_die_schlimmste_spruch, 2023-12-06
Das eskalierte deutlich mehr als ich zunächst erwartet hatte.
 - Reddit/ich_ieh_iichel, 2023-11-28
Das eskalierte schneller als gedacht
 - Reddit/Ratschlag/leute_die_nicht_zurück_grüßen, 2023-10-27
Also, **das eskalierte schnell**.
 - Reddit/de/unangenehmer_zwischenfall_airbus_a350_von_delta, 2023-09-06
Das eskalierte ziemlich schnell
 - Reddit/Ratschlag/mein_ausbilder_ist_ein_rassist_ich_bin_damit, 2023-08-20
Das eskalierte schnell.
 - Reddit/VeganDE/über_die_ethik_des_eieressens, 2023-08-15
Das eskalierte schnell 😅 Bei uns zuhause gibt es auch keine Eier/Eiprodukte
 - Reddit/FragReddit/wofür_verurteilt_du_heimlich_menschen, 2023-06-09
Das eskalierte mittelmäßig zügig.

[►►►] Create your own corpus! [◀◀◀]

Datenquelle:
Pushshift-Archives via
Academic Torrents



- (ND) JSON im zst-Format
- Top 40.000 Subreddits als separate Dateien

Open-Source-
Pipeline:
RedTEI



GitHub-Repositorium mit
vollständiger Dokumentation

