

# Twitter-Daten sammeln und mit Elasticsearch analysieren

Nobutake Kamiya (Universität Zürich, Universitätsbibliothek)

# Was ich hier anbiete:

- Wie Twitter-Daten gesammelt werden (Python)
- Einführung der Tools zur Datenanalyse (Elasticsearch)
- Denkanstoß zur Handhabung der Daten / Datenmanagement

# Hier geht's zum Ziel

Der Link führt zu meinem lokalen Server. Deshalb funktioniert er nur von meinem PC...

# Was ich hier **NICHT** anbiete:

- Eine wissenschaftliche Erkenntnis
- Eine fachspezifische Forschungsmethode

# Tweets sammeln

# Tweets sammeln: Methode

- Twitter API v2 - Academic Research
- Python

# Übersicht Twitter API

Essential	Elevated	Academic Research
500'000 tweets / Monat	2 Mio. Tweets / Monat	10 Mio. Tweets / Monat
Full-Archive-Suche nicht möglich	Full-Archive-Suche nicht möglich	Full-Archive-Suche und Full-Archive-Count
XX	XX	Advanced search operator

# Python

Die gesamten Codes sind [hier](#) zu finden.



# Python: Query 1

Datum und Uhrzeit Japanischer Zeit auf UTC (koordinierte Weltzeit)

```
jst_st = datetime.datetime(2021, 3, 1, 0, 0, 0, 0, datetime.timezone(timedelta(hours=+9)))
jst_et = datetime.datetime(2021, 4, 1, 0, 0, 0, 0, datetime.timezone(timedelta(hours=+9)))

utc_st = jst_st.astimezone(timezone.utc)
utc_st = utc_st.isoformat()
utc_et = jst_et.astimezone(timezone.utc)
utc_et = utc_et.isoformat()
```

# Python: Query 2

## Suchbedingungen

```
query_params = {'query': '(オリンピック lang:ja -is:retweet) OR (五輪 lang:ja -is:retweet)',  
                'tweet.fields': 'created_at,id,text,public_metrics',  
                'start_time': utc_st,  
                'end_time': utc_et,  
                'max_results': 500}
```

Weitere Regeln siehe [hier](#)

# Python: Data cleaning

Beispiel: [Ein Tweet des IOCs](#)  
([japanisch](#)).

The image shows a tweet from the official Olympic account (@gorin). The tweet is in Japanese and celebrates a gold medal win in the women's skateboarding park event at the Tokyo 2020 Olympics. The text of the tweet is: "1年前の今日、#東京2020 スケートボード 女子パークで、四十住さくら選手が金メダルを獲得🏆" and "決勝で唯一60点台を叩き出しました👏👏". The tweet includes a video player showing a skateboarder performing a trick. The video player has a title "決勝ラン1 安定した滑りを披露" and "6,570 views". The video is 0:09 / 0:45 long. The tweet was posted at 1:00 AM on August 4, 2022.

オリンピック  
@gorin · Follow

1年前の今日、**#東京2020** スケートボード 女子パークで、四十住さくら選手が金メダルを獲得🏆

決勝で唯一60点台を叩き出しました👏👏

**#オリンピック #スケートボード**

Continue watching on Twitter

決勝ラン1  
安定した滑りを披露  
6,570 views

0:09 / 0:45

1:00 AM · Aug 4, 2022

# Python: Data cleaning

```
# User-Name (Erwähnungen), URL, Hash-tags und Zeilenumbrüche aus dem Tweets entfernen
tw_text = re.sub(r'@\w+', '', ['text'])
tw_text = re.sub(r'(http|https)://[0-9a-zA-Z\./]+', '', tw_text)
tw_text = re.sub(r'#+?(\s|$)', '', tw_text)
tw_text = re.sub(r'\n', '', tw_text)
# Kana soll immer in Fullwidth-Zeichen dargestellt werden
tw_text = mojimoji.han_to_zen(tw_text, kana=True, digit=False, ascii=False)
# Digit und Ascii sollen immer in Halfwidth-Zeichen dargestellt werden
tw_text = mojimoji.zen_to_han(tw_text, kana=False, digit=True, ascii=True)
```

# Besonderheiten von der Twitter-Daten

- "Academic researchers are permitted to distribute an unlimited number of Tweet IDs and/or User IDs [...]", aber mehr nicht [Developer terms \(unter "content redistribution"\)](#).
- Tweets können gelöscht/geändert werden!

# Tweets sammeln - Zusammenfassung

- Nutzungsbedingungen der Daten kennenlernen!
- Codes für Query und Data cleaning publizieren
- Datum der Ausführung dokumentieren

# Elasticsearch (ES)

# Was ist Elasticsearch?

- Eine Suchmaschine basierend auf Lucene
- Nutzung der Standarddistribution ist kostenlos
- Auch in [次世代デジタルライブラリー](#) von der NDL genutzt



# Plugins für Japanisch

- [Kuromoji](#)
- [ICU \(International Components for Unicode\)](#)
- [Sudachi \(die neueste Version für ES 5.6?\)](#)

In diesem Beispiel werden **Kuromoji** und **ICU** für Analyser verwendet

# Einstellung des Analyzers

- Konfiguration kann in JSON-Format geschrieben werden
- Char\_filter (Normalisierung der Schriftzeichen, fakultativ)
- Tokenizer (Worttrennung [z.B. N-gram], notwendig, nur ein Tokenizer anwendbar)
- Token\_filter (Wörter [z.B. Stopwörter] werden nach bestimmten Regeln gefiltert, fakultativ)

## Einstellung mit dem Beispielsatz

"そのオリンピック選手は身長196センチという長身だった。人々はおどろいた。二〇〇〇人がオリンピックを観戦しながらコンピューターをつかっていた"

# Char\_filter - Einstellung

```
{
  "settings": {
    "analysis": {
      "analyzer": {
        "my_kuromoji_analyzer": {
          "type": "custom",
          "char_filter": ["icu_normalizer", "kuromoji_iteration_mark"],
          "tokenizer": "keyword"
        }
      }
    }
  }
}
```

Der "Keyword"-Tokenizer gibt den Satz so zurück (s. [hier](#))

# Ergebnis

そのオリンピック選手は身長**196センチ**という長身だった。**人人**はおどろいた。二〇〇〇人がオリンピックを観戦しながらコンピューターをつかっていた

# Tokenizer-Einstellung

```
{
  "settings": {
    "analysis": {
      "analyzer": {
        "my_kuromoji_analyzer": {
          "type": "custom",
          "char_filter" : ["icu_normalizer", "kuromoji_iteration_mark"],
          "tokenizer": "kuromoji_tokenizer"
        }
      }
    }
  }
}
```

# Ergebnis

[Ergebnis in JSON-Format](#)

# Token\_filter-Einstellung 1

```
[...]  
  "filter": ["kuromoji_baseform", "kuromoji_part_of_speech", "kuromoji_stemmer", "ja_stop", "kuromoji_number", "synonym_filter"]  
    },  
    "filter": {  
      "synonym_filter": {  
        "type": "synonym",  
        "synonyms": ["オリンピック,五輪"]  
      }  
    }  
  }  
}
```



# Token\_filter-Einstellung 2

Einige Token\_filter von Kuromoji werden verwendet

- kuromoji\_baseform
- kuromoji\_part\_of\_speech

Zusätzlich wird der "synonym\_filter" eingesetzt.

- "オリンピック" und "五輪"

# Ergebnis

[Hier](#)

# Die endgültige Einstellung

...Falls man sich dafür interessiert... [Hier](#)

# Elasticsearch - Zusammenfassung

- Mit ES können die große Datenmengen behandeln und analysieren werden
  - Mit API können die Daten weiter verarbeitet werden, z.B. für Netzwerkanalyse
  - [Kibana](#) (ein Plugin von ES) erlaubt eine einfache GUI-Verarbeitung
- Die Konfiguration sollte bekannt gemacht werden

# Denkanstoß - 1

- Bearbeitung großer Mengen von japanischen Texten ist durchaus möglich
- Für die wissenschaftlichen Kommunikation ist die Nachvollziehbarkeit wichtig
- Open-Data-Gedanken werden sehr wichtig

# Denkanstoß - 2

- Falls die Rohdaten nicht publiziert werden können, sollten mindestens die Codes für Datensammlung und Verarbeitung publiziert werden
- Für Bibliothekare ist es interessant, Information über anwendbaren Datenquellen und deren Nutzungsbedingungen zu wissen/vermitteln

**Vielen Dank!**