# Medienanalyse-Webanwendung

Datenbericht

Letzte Änderung: 24.08.2021

## Rohdaten

### Übersichtstabelle der Rohdatensätze

| Datensatz Name | Quelle | Speicherort |
|---|---|---|
| New_data_vdss.tsv.xz | Swissdox@LiRI | Im Ordner namens data. |

### Details Rohdaten

- **Informationen über die Daten**: Swissdox@LiRI ist eine umfangreiche Datenbank, die etwa 23 Millionen veröffentlichte Medienartikel aus einer Vielzahl von Schweizer Medienquellen enthält. Diese Quellen umfassen sowohl Print- als auch Digitalmedien und decken viele Jahrzehnte ab. Täglich werden etwa 5000 bis 6000 neue Artikel hinzugefügt.
- **Details zur Datenquelle**: Die Daten stammen von unserem Partner CH Media, NZZ Mediengruppe, Ringier, Ringier Axel Springer Schweiz und TX Group (Tamedia), SRF/SRG und Wochenzeitung, insgesamt 250 Quellen mit geplanter weiterer Expansion.
- **Datenbeschaffung**: Die Datenbeschaffung erfolgt durch eine Kooperation zwischen LiRI und SMD (Schweizer Mediendatenbank AG). Die Initiative wurde von Prof. Dr. Noah Bubenhofer, Prof. Dr. Fabrizio Gilardi (UZH) und Roberto Nespeca (SMD) ins Leben gerufen und wird von der Universität Zürich UZH (Technologieplattform-Kommission) und den folgenden Unterstützern finanziert: Zürcher Hochschule für Angewandte Wissenschaften (Abteilung für Angewandte Linguistik), Universität Basel/Universitätsbibliothek Basel, ETHZ Bibliothek, Universitätsbibliothek Bern.
- **Wichtige Punkte aus den Nutzungsbedingungen**:
    - Die Daten dürfen ausschließlich für Forschungs- und akademische Zwecke verwendet werden. Eine kommerzielle Nutzung der Daten sowie jegliche Derivate sind nicht erlaubt.
    - Die Daten dürfen nur für das angegebene Forschungsprojekt verwendet werden. Die Wiederverwendung von Daten oder Teilen davon für andere Zwecke als das angegebene Forschungsprojekt ist nicht gestattet.
    - Es ist nicht erlaubt, die gesamten Daten herunterzuladen. Die für ein individuelles Forschungsprojekt erhaltenen Daten dürfen also nicht alle Texte innerhalb des Korpus enthalten.
    - Die Daten dürfen nur lokal auf Geräten der Forscher und Studierenden oder auf der Infrastruktur des Vertragspartners (akademische Institution) gespeichert werden. Insbesondere ist die Speicherung auf Cloud-Plattformen Dritter nicht gestattet.
    - Die Daten dürfen nicht mit Dritten geteilt werden.
    - Die aus dem Korpus erhaltenen Daten müssen spätestens sechs Monate nach Beendigung des Forschungsprojekts gelöscht werden. Eine Archivierung der erhaltenen Daten ist nicht erlaubt. Dies gilt nicht für aggregierte Daten oder andere Derivate, sofern diese nicht die gesamten Rohdaten oder bezogenen Texte enthalten.
    - Bestimmte Details des Projekts werden Swissdox mitgeteilt und können auf den SMD/Swissdox und LiRI-Websites erwähnt werden.
    - Bei wissenschaftlichen Publikationen, die Ergebnisse auf Basis von Swissdox-Daten präsentieren, muss eine bestimmte Anerkennungsnote enthalten sein.

- Geschäftsrelevant

## Datenkatalog Rohdaten

Beispiel eines Data Dictionary

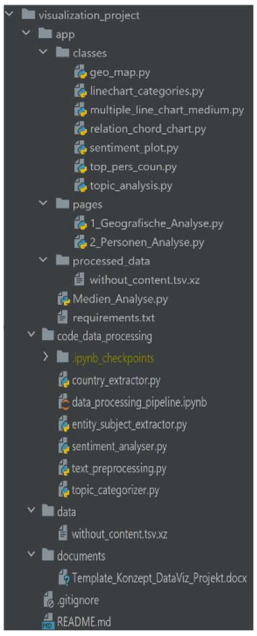
| Spaltenindex | Spaltenname | Datentyp | Werte (Wertebereich, ev. Validierungsregeln) | Kurze Beschreibung | Primary / Foreign Key |
|---|---|---|---|---|---|
| 1 | id | Int64 | | Eine eindeutige Identifikationsnummer für jeden Eintrag in der Datenbank. | Primary Key |
| 2 | Pubtime | object | | Das Datum und die Uhrzeit der Veröffentlichung des jeweiligen Medienartikels. | |
| 3 | Medium_code | object | | Ein Code, der das Medium (die Quelle) des Artikels repräsentiert. Jede Quelle hat einen eindeutigen Code. | |
| 4 | Medium_name | object | | Der Name des Mediums (der Quelle), aus dem der Artikel stammt. | |
| 5 | Rubric | object | | Rubrik des Artikels. | |
| 6 | Regional | object | | Geografische Region, auf die sich der Artikel bezieht. | |
| 7 | Doctype | object | | Ein Code, der den Dokumenttyp des Artikels repräsentiert. Jeder Dokumenttyp hat einen eindeutigen Code. | |
| 8 | Doctype_description | object | | Textuelle Beschreibung des Dokumenttyps. | |
| 9 | Language | object | | Die Sprache, in der der Artikel geschrieben ist. | |
| 10 | Char_count | object | 52 - 588653 | Die Anzahl der Zeichen im Inhalt des Artikels. | |
| 11 | Dateline | object | | Schlagzeilen oder kurze Beschreibungen der Artikel. | |
| 12 | Head | object | | Die Überschrift oder der Titel des Artikels. | |

| | | | | | |
|---|---|---|---|---|---|
| 13 | Subhead | object | | Die Unterüberschrift oder der Untertitel des Artikels. | |
| 14 | Content_id | object | | Eine eindeutige Identifikationsnummer für den Inhalt des Artikels. | |
| 15 | Content | object | | Der tatsächliche Inhalt oder Text des Artikels. | |

## Datenaufbereitung Rohdaten

### Vorgehensweise

Der erste Schritt bestand darin, die Daten zu laden und einige Vorverarbeitungsschritte durchzuführen, um unerwünschte Spalten zu entfernen. Danach wurde eine Textvorverarbeitung auf den Inhalt der Artikel angewendet, und anschließend wurde eine Länderextraktion durchgeführt, um zu erkennen, welche Länder in den Artikeln erwähnt werden. Es folgte eine Sentiment-Analyse, um die Polarität und Subjektivität der Artikel zu bestimmen. Als nächstes wurde eine Kategorisierung der Themen der Artikel durchgeführt und Entitäten wurden aus den Artikeltiteln extrahiert. Zuletzt wurden die Ländernamen übersetzt und Personen aus dem verarbeiteten Inhalt extrahiert. Die Daten wurden während der Analyse mehrmals gespeichert, um die Ergebnisse zu sichern und die Reproduzierbarkeit zu gewährleisten.




In der Abbildung wird die Struktur unseres Projektes veranschaulicht:

- Der Ordner "classes" enthält alle Python-Module, die wir für die Erstellung von Illustrationen verwendet haben.
- Der Ordner "pages" beherbergt die zweite und dritte Seite unseres Dashboards.
- Im Ordner "processed_data" befinden sich diverse Elemente. Dazu gehört die Hauptseite des Dashboards, der verarbeitete Datensatz sowie eine Datei, die die technischen Voraussetzungen (z.B. benötigte Python-Pakete) auflistet, die zur Ausführung des Dashboards erforderlich sind.
- Der Ordner "code_data_processing" enthält alle Dateien, die zur Verarbeitung der Rohdaten erforderlich sind.

# Prozessierte Daten

## Übersichtstabelle der Prozessierten Daten

| Name | Input-Datensätze | Speicherort |
|---|---|---|
| Without_content.tsv.xz | New_data_vdss.tsv.xz | Im Ordner namens data |

## Details prozessierte Daten

- Der verarbeitete Datensatz enthält Informationen wie das Veröffentlichungsdatum, den vorverarbeiteten Inhalt des Artikels, die im Artikel genannten Länder, die Sentiments, die Subjektivität und die Kategorie des Artikels sowie die aus der Überschrift extrahierten Entitäten.

- Die Datenprozessierungsschritte wurden mit mehreren selbst erstellten Python-Modulen durchgeführt. Diese umfassen:
    - **text_preprocessing**: Dieses Modul enthält die TextPreprocessing Funktion, die dazu dient, den Textinhalt der Artikel zu normalisieren und irrelevante Informationen zu entfernen. Dabei werden Schritte wie das Entfernen von Stoppwörtern, die Durchführung einer Lemmatisierung und das Entfernen von Sonderzeichen und Zahlen durchgeführt.
    - **country_extractor**: Das Modul country_extractor beinhaltet die Klasse CountryExtractor, die dazu verwendet wird, Ländernamen aus dem Textinhalt der Artikel zu extrahieren. Diese Information kann nützlich sein, um den geographischen Fokus eines Artikels zu bestimmen.
    - **sentiment_analyser**: Dieses Modul implementiert die SentimentAnalyser Klasse, die Sentiment-Polaritäten und Subjektivität aus dem Textinhalt extrahiert. Dies kann hilfreich sein, um die Tonalität und die Objektivität/ Subjektivität eines Artikels zu bestimmen.
    - **topic_categorizer**: Das topic_categorizer Modul enthält die TopicCategorizer Klasse, die dazu dient, die in einem Artikel behandelten Themen zu kategorisieren. Dies kann dabei helfen, die Hauptthemen eines Artikels zu identifizieren und zu verstehen.
    - **entity_subject_extractor**: Dieses Modul umfasst die EntityAndSubjectExtractor Klasse, die Entitäten und Subjekte in den Überschriften der Artikel identifiziert. Das Extrahieren dieser Informationen kann dabei helfen, den Fokus und die Hauptpunkte eines Artikels zu verstehen.
- Die final verarbeiteten Daten können aus der Datei `processed_data_final.tsv.xz` ausgelesen werden, die im '../data/' Verzeichnis gespeichert ist. Dies geschieht mit Hilfe des `pandas` Moduls in Python, das für Datenmanipulation und -analyse verwendet wird. Der Code zum Laden des Datensatzes lautet:
  `pd.read_csv('../data/processed_data_final.tsv.xz', sep='\t', compression='xz').`