Exposé
RadioMining: Eine Analyse von Radiosendern und den deutschen Charts.

Projektgruppe 13 (Moodle Gruppe 13)

15. Juni 2025

Kurze Projektbeschreibung

Das Projekt **„RadioMining“** untersucht einige Radiosender aus dem DACH-Raum, sowie deren Zusammenhang mit den offiziellen Deutschen Charts. Im Detail werden die Wiedergabe bzw. Playlisten, sowie die Startseiten (engl. landing page) der Radiosender untersucht. Hierzu wurde ein GitHub-Projekt erstellt[1].
Zustätzlich werden auch gesprochene Inhalte aus Radiosendungen aufgezeichnet, transkripiert und anaylsiert. Hierzu wurde das Tool `audio_miner`[2] erstellt, welches mittels `ffmpeg` die Livestreams der Sender lokal speichert. Im Anschluss werden diese mittels `PyTorch` und `pyannote`[3] einer simplen Sprechererkennung durchgeführt und dann der augefzeichnete Ton mithilfe von `OpenAI Whisper`[4] transkribiert.
Ein `Streamlit`-gestütztes Tagging-Tool erlaubt das komfortable Labeln jeder Transkriptzeile mit Kategorien wie `music`, `traffic` oder `advertisement`. Anschließend trainiert ein `spaCy`-Klassifikator die automatische Einordnung neuer Segmente.
Zuletzt möchte das Projekt prüfen, inwieweit es möglich ist, die mittels AudioMining erfassten Informationen mit den per Webscraping erfassten Informationen zu kombinieren.

Projektziele und Scope

- **Datenerhebung:** Kontinuierliches Scraping der Playlists und Landingpages, sowie der Livestreams von einigen der Radiosender.
- **Transkription & Tagging:** Batch-Transkription mit `audio_miner` und manuelle Annotation mittels des Tagging-Tools.
- **Modelltraining:** Training mittels eines `spaCy`-Textklassifikators zur automatischen Segmentklassifikation.
- **Vergleich mit Charts:** Abgleich der ermittelten Musikrotationen mit den Top 100 der offiziellen Charts, zur Analyse der Sendernähe zu Mainstream-Trends.

[1] `https://github.com/fanonwue/fhswf-radio-scraper`
[2] `https://github.com/smilchsack/audio_miner`
[3] `https://huggingface.co/pyannote`
[4] `https://github.com/openai/whisper`

- **Vergleich von Nachrichten Überschriften und Transkripten:** Untersuchung der gesprochenen Inhalte der Radiosender im Vergleich zu den Nachrichtenüberschriften der Radiosender.
- **Visualisierung & Bericht:** Auswertung der Ergebnisse, Visualisierung der Daten und Erstellung eines Abschlussberichts.

Genutzte Datenquellen

Die nachfolgende Tabelle listet die genutzten Datenquellen auf, die für das Projekt verwendet werden:

Tabelle 1: Erfasste Datenquellen

Sender	**Audio (Livestream)**	**Playlists**	**Landingpage**
SWR1 RLP	✓	✓	✓
SWR3	✓	✓	✓
SRF3	–	✓	✓
WDR2	✓	✓	–
DLF Nova	–	✓	–
Radio MK	–	✓	–
Offizielle Charts	–	✓	–

– = keine Erfassung, ✓ = Erfassung erfolgt.

Vorgehensweise

Nachfolgend wird die Vorgehensweise des Projekts beschrieben. Das Projekt ist in vier Phasen unterteilt, die jeweils unterschiedliche Aufgaben und Ziele umfassen: 1 – Entwicklung, 2 – Datenerhebung, 3 – Analyse und 4 – Auswertung.

Tabelle 2: Aufgabenübersicht für das Projekt *RadioMining*

Nr.	**Aufgabe**	**Bearbeiter**	**Phase**
1	Exposé erstellen	Sebastian Milchsack	1
2	Web-Crawler-Basis & `audio_miner` entwickeln	Fabian Wünderich, Sebastian Milchsack	1
3	Web-Crawler erstellen	Sebastian Dornack, Sebastian Milchsack, Nils Robinet	1
4	Crawler deployen u. Daten sammeln	Sebastian Dornack	2
5	Audiostreams deployen, Daten sammeln u. transkribieren	Sebastian Milchsack	2
6	Audiostreams für n-Tage taggen	TBD	3
7	Datenvorbereitung (Landingpages, Playlists)	TBD	3
8	spaCy-Klassifikator trainieren	TBD	3
9	Musikrotationen ↔ Top 100 abgleichen	TBD	4
10	News-Titel ↔ Transkripte prüfen	TBD	4
11	Zwischenpräsentation	gesamtes Team	4
12	Ergebnisse evaluieren	TBD	4
13	Resultate visualisieren, Endpräsentation u. Bericht erstellen	gesamtes Team	4