***Transcrire Radio Acidule***

**Rapport de projet**

[Introduction 2](#_Toc136217080)

[Contexte 2](#_Toc136217081)

[But du projet 2](#_Toc136217082)

[Organisation 2](#_Toc136217083)

[Prototype 2](#_Toc136217084)

[Abord de la problématique 3](#_Toc136217085)

[Diagramme de Gantt 3](#_Toc136217086)

[Méthodologie 4](#_Toc136217087)

[Transcription 4](#_Toc136217088)

[Base de données 5](#_Toc136217089)

[Nettoyage des données 5](#_Toc136217090)

[Mots fréquents 5](#_Toc136217091)

[Topic Modeling 5](#_Toc136217092)

[Web App 6](#_Toc136217093)

[Conclusion 6](#_Toc136217094)

[Résultats 6](#_Toc136217095)

[Améliorations à envisager 6](#_Toc136217096)

# Introduction

## Contexte

Radio-Acidule était une station locale lausannoise ayant diffusé des émissions variées entre 1984 et 1999. Une partie des émissions diffusées sur Radio-Acidule est actuellement disponible en ligne sur la Web TV des Archives de la Ville de Lausanne, hébergée par Dartfish TV.

Les archives de Radio-Acidule se composent de :

* 879 enregistrements d’émissions, dont certains sont des enregistrements d’émissions musicales. Le plus long enregistrement dure 5 heures 5, le plus courte 1 minute 43.

Les enregistrements sont annotés avec certains thèmes présents dans l’émission, la date de diffusion, les intervenant.es, le format et le type de document (cf. figure 1). Un texte de description accompagne parfois l’enregistrement. Il est parfois très bref, notamment pour les enregistrements de séances du conseil communal, et ne permet pas de comprendre les thèmes abordés au sein de l’émission.

## But du projet

Les annotations actuelles ne permettent pas de se rendre bien compte du contenu de l’émission. Afin de compléter les mots-clés associés aux émissions, il serait pertinent **d’analyser la transcription textuelle des émissions.**

L’objectif principal du projet est ainsi de transcrire les émission « parlées » (càd qui ne sont pas uniquement composées de musique) de Radio Acidule afin d’en faire une analyse textuelle, qui aurait pour but la déduction des thèmes les plus abordés au sein de l’émission.

# Organisation

## Prototype

J’ai commencé par établir un prototype de l’interface souhaitée. Il me semblait important de mettre en place deux moyens d’accéder aux transcriptions :

* Recherche par mots présents fréquemment dans les transcriptions.
* Recherche par thèmes, déduits par topic modeling.

Une image contenant texte

Description générée automatiquement

## Abord de la problématique

Ayant constaté les difficultés de navigation de la Web TV des AVL, je souhaitais mettre en place une manière d’accéder aux émissions par thèmes et par mots fréquents, qui donnerait une idée plus générale du contenu disponible et faciliterait la navigation au sein du corpus.

## Diagramme de Gantt

Le calendrier mis en place lors de la réalisation du diagramme de Gantt n’a pas été respecté pour chaque échéance, mais il a servi de guide général pour suivre l’avancement du projet. Avec du recul, je constate qu’il aurait été très utile de mieux définir la manière dont l’app allait être organisée avant de me lancer dans la création de la base de données. En effet, certaines informations pourraient être plus accessibles et facilement intégrables à l’app si leur enregistrement avait été prévu à l’avance (par exemple pour les langues, les probabilités des topics, …)

Une image contenant texte, capture d’écran, logiciel, Page web

Description générée automatiquement

# Méthodologie

## Transcription

Le modèle de reconnaissance vocale Whisper a été utilisé pour la transcription.

<https://github.com/openai/whisper>

## Base de données

La bibliothèque SQLite était une bonne option pour la taille du projet. En effet, une base de données MySQL aurait été trop lourde pour une utilisation avec aussi peu de documents.

Une image contenant texte, diagramme, reçu, ligne

Description générée automatiquement

## Nettoyage des données

Avant d’utiliser les transcriptions pour déterminer les thèmes et mots fréquents, un léger nettoyage des données a été nécessaire. Il reste quelques erreurs dans le texte, mais les erreurs de transcription les plus fréquentes ont été corrigées dans le fichier corrections.py.

Les stopwords par défaut de du corpus de spacy « fr\_core\_news\_sm » ont été retirés des calculs, ainsi qu’une liste de stopwords déterminée en observant le corpus. La liste est la suivante :

Stopwords = ["beaucoup", "bien", "oui", "non", "bon", "chose", "être", "avoir", "pouvoir", "faire", "envoyer", "aller", "prendre", "devoir", "voir", "permettre", "falloir", "croire", "vouloir" ]

Elle est composée de mots très fréquents et ayant peu d’importance sémantique, tels les verbes les plus fréquents en français.

## Mots fréquents

La bibliothèque de traitement automatique des langues a été utilisée pour déterminer, pour chaque texte, les 10 mots les plus fréquents.

Ceux-ci ne peuvent contenir que des mots qui ont le Part-of-Speech-tag ‘NOUN', 'ADJ', 'ADV', 'VERB', ou 'PROPN'.

## Topic Modeling

La bibliothèque Gensim (<https://radimrehurek.com/gensim/>)a été utilisée pour conduire une LDA (Latent Dirichlet Allocation).

## Web App

Afin de mettre en page les résultats, et de proposer une manière de naviguer au sein du corpus Acidule, le Framework pour applications web Streamlit a été utilisé.

# Conclusion

## Résultats

Le but de ce projet de transcription était en partie **d’évaluer l’intérêt et la pertinence d’un projet de transcription plus large des archives radio.**

Les retranscriptions pourront être **mises à disposition** des archives, ainsi qu’une App permettant d’explorer le corpus Acidule.

## Améliorations à envisager

* La qualité des transcriptions pourrait être améliorée grâce à un programme de correction automatique, ou par la relecture. Le son de certaines émissions est très peu clair, il y a souvent de la musique en arrière-plan. La transcription ne serait donc peut-être pertinente que pour certaines parties du corpus sélectionnées manuellement, et qui pourraient être relues.
* Les noms des émissions pourraient être mieux formulés en utilisant le fichier XML fourni par les AVL, afin de rendre l’utilisation de la Web App plus agréable à la lecture.
* L’adresse web des émissions en ligne sur la Web TV des AVL ne correspond à aucune donnée disponible dans le fichier XML ou dans le nom du fichier. Il n’est donc pas possible de lier directement la page d’une émission à l’enregistrement de celle-ci.
* La langue des émissions a été détectée, elle pourrait être ajoutée comme critère de recherche.