Concert Program OCR

Jonah Broyer and Samson DeVol

Oregon State University Cascades

CS.67

Background

When attending professional symphony orchestra concerts, audiences are often given booklets called programs. They typically contain pages of information pertaining to the composer, performers, and commentary about the pieces audiences are hearing.

The Concert Program Optical Character Recognition (OCR) project collects historical concert data from the web and organizes it in data repository so users can download the information, filtered by desired attributes. This system allows this text data to be seamlessly fed to various Deep Learning algorithms.



Philharmonic Society of New York

FOUNDED 1842

FIFTY-SECOND SEASON, 1893-1894.

ANTON SEIDL, - - - - CONDUCTOR.

Second Public Rehearsal, Friday, Dec. 15th, 1893,
AT 2 P. M.

Second Concert, Saturday, December 16th, 1893,
AT 8:15 P. M.

MUSIC HALL, 7th Avenue and 57th Street.

Soloist, - - - - Mr. HENRI MARTEAU.

— PROGRAMME: —

PART I.

Figure 1: Example of a Program

Contributors



Jonah Broyer
Team Member
jonahbroyer.com



Samson DeVol
Team Member
samsondevol.com



Patrick Donnelly
Project Partner
soundbendor.com

Implementation

The music concert program data is gathered from existing digital archives using web scraping scripts. The first 14,000 programs were gathered from the New York Philharmonic from concerts dating back to the 1840's. The system connects data collection scripts to the Microsoft SQL database which facilitates a direct pipeline to the front-end web application that allows users to view, filter, and download datasets as `csv` files. This system enables training large-language models with historical concert programs.

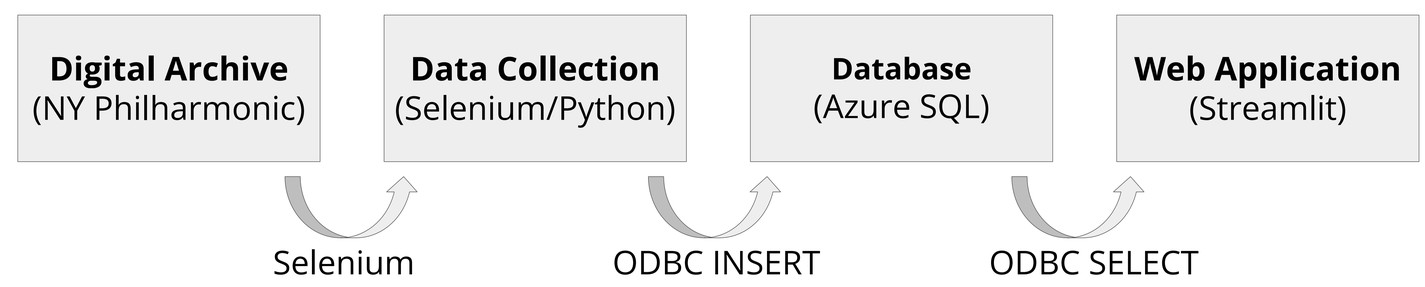
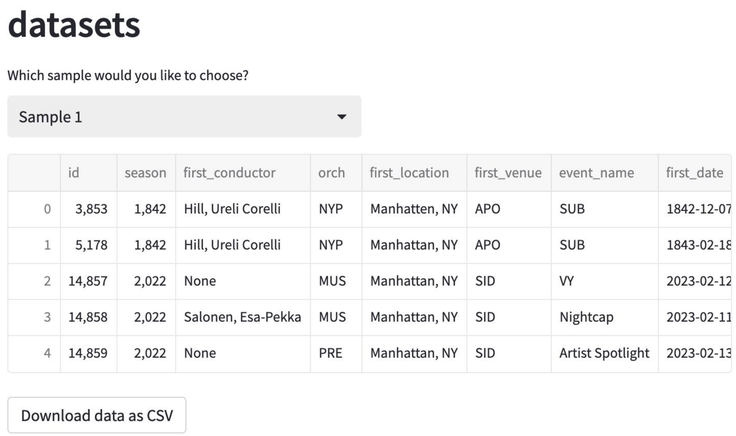


Figure 2: Pipeline of Concert Program OCR

Objective

Collecting, cleaning, and filtering this data is a cumbersome yet essential part of training deep machine learning models. Although information on historical music has been digitized, such data is often inaccessible and inconvenient for immediate use in deep learning experiments. This project aims to provide machine learning researchers a robust source of metadata, OCR text, and supplementary information for classical music concerts. Using the concert program data in deep learning algorithms can advance neural networks associating descriptive terms with program attributes that deepen our written and oral understanding of how music affects humans emotionally.



datasets

Which sample would you like to choose?

Sample 1

	id	season	first_conductor	orch	first_location	first_venue	event_name	first_date
0	3,853	1,842	Hill, Ureli Corelli	NYP	Manhatten, NY	APO	SUB	1842-12-07
1	5,178	1,842	Hill, Ureli Corelli	NYP	Manhattan, NY	APO	SUB	1843-02-18
2	14,857	2,022	None	MUS	Manhattan, NY	SID	VY	2023-02-12
3	14,858	2,022	Salonen, Esa-Pekka	MUS	Manhattan, NY	SID	Nightcap	2023-02-11
4	14,859	2,022	None	PRE	Manhattan, NY	SID	Artist Spotlight	2023-02-13

Download data as CSV

Figure 3: Streamlit Page with Concert Program Data

soundbendor lab