OCR

OCR until yesterday

Input image

Layout Inference

Segmentation
- Layout Postprocessing
- U-Net Inference

Baseline
- U-Net Inference
- Baseline Postprocessing
- Baseline Ascender Descendender Endpoints Regions

Text-Line Inference

LSTM Inference

Optical Character Recognition

Köln-
Gießener
Eifenbahn.
Mit Bezug auf unfere Veröffentlichung
vom 23. v. Mts. machen wir hierdurch be-
kannt, daß vom Mittwoch den 11. d. Mts.
ab auf den Strecken Deutz-Wetzlar und Betz-
dorf-Siegen alle Perfonenzüge wieder fahr-
vlanmäßig abgelaffen und Güter ohne Be-
fchränkung befördert werden, wogegen auf der
Strecke Wetzlar-Gießen bis auf Weiteres noch
jeder Verkehr unterbrochen bleibt.
Köln, den 9. Juli 1866.
Die Direction der Köln-Mindener Eifenbahn-
Gefellfchaft.

EXAMPLE: Musical "The New Records"

The New Records
A BULLETIN FOR THOSE INTERESTED IN RECORDED MUSIC

Issued Monthly by

H. ROYER SMITH COMPANY
"The World's Record Shop"

10TH & WALNUT STREETS   -   -   -   PHILADELPHIA 7, PA., U. S. A.

VOL. 18, NO. 1   MARCH, 1950   By mail to any address 50c per year

THIS is the 205th issue of this little bulletin. With it we begin the eighteenth volume. Many of our readers have been with us since Vol. 1, No. 1 and quite a few of them have kept a complete file of all issues. We know this because occasionally a copy goes astray in the mail and they hasten to write for another copy stating that their files are complete except for that one number. We hope that the present volume will prove as interesting to these old-timers as the previous ones.

During the last few months we have been advertising THE NEW RECORDS in a number of leading newspapers throughout the country including *The New York Times, Chicago Tribune, Los Angeles Times, Boston Herald, Detroit Free Press, Cleveland Press, St. Louis Star Times, Cincinnati Enquirer, Minneapolis Star Tribune* and *Washington Times Herald.* Hundreds of new subscribers have been secured by this means and we take this opportunity of welcoming them with the hope that they will continue to find our little bulletin of value.

Last month we included a renewal blank with each copy mailed, and so many renewals were received that it has been impossible for us to check our mailing list. We are, therefore, carrying over all of our subscribers, whether they renewed or not. We cannot continue this practice beyond this month, so if you have not as yet sent in your renewal, you should do so at once in order that you will not miss any issues.

* * *

The demand for LP recordings, especially of classical music, is snowballing all over the country. Our readers are continually requesting us to list as many of them as we possibly can so that they may know what is available. This month we are making a very special effort to bring our readers up-to-date on the LP releases.

In doing so we have had to shorten many of our reviews in order to list more recordings in the space available. In so doing we hope we are pleasing the majority of our subscribers.

BACK COPIES

We are continually receiving requests for back copies of THE NEW RECORDS. Most of the issues published during the last seventeen years are available. The price is 5¢ each. A file of all of the available issues (at least 115 copies) is $5. These prices are postpaid within U. S. A.

NEW LP RELEASES

Schubert: *Symphony No. 1 in D.* Winterthur Symphony Orchestra — Swoboda. 1-12" disc (CH-CHC-23). $4.85.

Haydn: *Symphony No. 77 in B-flat.* One side, and Haydn: *Symphony No. 78 in C minor.* Concert Hall Symphony Orchestra—Swoboda. 1-12" disc (CH-CHC-30). $4.85.

Bach: *Concerto in E major.* Hans Andreae (harpsichord) with the Winterthur Symphony Orchestra —Kromer. One and one-half sides, and Bach: *Fugue in A minor.* Hans Andreae (harpsichord). 1-10" disc (CH-CHC-33). $3.85.

Brahms: *Variations and Fugue on a Theme of Handel, Op. 24.* Ella Goldstein (piano). 1-12" disc (CH-CHC-35). $4.85.

Vivaldi: *Concerto in B-flat.* One side, and Vivaldi: *Concerto in E minor.* Leo Rostal (violoncello) and Concert Hall String Ensemble. 1-12" disc (CH-CHC-37). $4.85.

Tchaikovsky: *Dumka, Op. 59.* One side, and Prokofiev: *Summer Day—Suite.* Ray Lev (piano). 1-12" disc (CH-CHC-26). $4.85.

Fauré: *Theme and Variations, Op. 73.* One side, and Fauré: *Nocturne No. 7* and Fauré: *Barcarolle in A minor, Op. 26.* Paul Loyonnet (piano). 1-12" disc (CH-CHC-16). $4.85.

Handel: *Sonata No. 5 in A.* One side, and Handel: *Sonata No. 6 in E.* Joseph Scripka (violin) and Elizabeth Rich (piano). 1-10" disc (CH-CHC-15). $3.85.

Schumann: *Carnaval, Op. 9.* Ella Goldstein (piano). 1-12" disc (CH-CHC-32). $4.85.

Villa-Lobos: *Quartet No. 6 in E* (Quartetto Brasileiro No. 2). Stuyvesant String Quartet. 1-12" disc (CH-CHC-19). $4.85. (TNR Oct. '47).

Bloch: *Quartet No. 2.* Stuyvesant String Quartet. 1-12" disc (CH-CHC-20). $4.85. (TNR Feb. '48).

Mozart: *Symphony No. 20 in D, K. 133.* Vox Chamber Orchestra—Fendler. 1-10" disc (CH-CHC-21). $3.85. (TNR Apr. '47).

Smetana: *The Moldau.* One side, and Dvořák: *Husitká Overture, Op. 67.* Boston "Pops" Orchestra—Fiedler. 1-10" disc (V-LM-1). $3.85. (TNR July '48).

Puccini: *Highlights from Madama Butterfly.* Albanese, Browning, Melton with RCA Victor Orchestra—Weissmann. 1-10" disc (V-LM-2). $4.45. (TNR Dec. '46).

Bruch: *Scottish Fantasy, Op. 46.* Heifetz (violin) with RCA Victor Orchestra—Steinberg. 1-10" disc (V-LM-4). $4.45. (TNR. Feb. '48).

Mendelssohn: *Concerto in E minor, Op. 64.* Elman (violin) with Chicago Symphony Orchestra—Defauw. 1-10" disc (V-LM-5). $4.45. (TNR Apr. '48).

Wagner: *Siegfried*—Act III, Scene 3 (complete). Farrell, Svanholm with Rochester Philharmonic Orchestra—Leinsdorf. 1-12" disc (V-LM-1000). $5.45. (TNR Dec. '49).

Offenbach: *Gaîté Parisienne.* Boston "Pops" Orchestra—Fiedler. 1-12" disc (V-LM-1001). $4.85. (TNR Dec. '47).

Rimsky-Korsakov: *Scheherazade.* San Francisco Symphony Orchestra—Monteux. 1-12" disc (V-LM-1002). $4.85. (TNR Nov. '42).

Tchaikovsky: *The Swan Lake, Op. 20.* St. Louis Symphony Orchestra—Golschmann. 1-12" disc (V-LM-1003). $4.85. (TNR Feb. '46).

Grofé: *Grand Canyon Suite.* NBC Symphony Orchestra—Toscanini. 1-12" disc (V-LM-1004). $5.45. (TNR Apr. '46).

Rachmaninoff: *Concerto No. 2 in C minor, Op. 18.* Rubinstein (piano) with NBC Symphony Orchestra—Golschmann. 1-12" disc (V-LM-1005). $5.45. (TNR Oct. '46).

Khachaturian: *Concerto for Piano and Orchestra.* Kapell (piano) with Boston Symphony Orchestra —Koussevitzky. 1-12" disc (V-LM-1006). $5.45. (TNR Feb. '47).

Bizet: *Carmen* (excerpts). Swarthout, Albanese, Merrill, Vinay and RCA Victor Chorale and Orchestra—Leinsdorf. 1-12" disc (V-LM-1007). $5.45. (TNR Dec. '46).

Tchaikovsky: *The Sleeping Beauty Ballet* (excerpts). Leopold Stokowski and his Symphony Orchestra. 1-12" disc (V-LM-1010). $5.45. (TNR May '48).

Lalo: *Symphonie Espagnole, Op. 21.* Menuhin (violin) with Colonne Orchestra—Fournet. 1-12" disc (V-LM-1011). $5.45. (TNR June '48).

Ravel: *Bolero.* One side, and Ravel: *Ma Mère l'Oye.* Boston Symphony Orchestra—Koussevitzky. 1-12" disc (V-LM-1012). $5.45. (TNR Aug. '48 & Apr. '49).

Dvořák: *Symphony No. 5 in E minor, Op. 95* ("from the New World"). Leopold Stokowski and his Symphony Orchestra. 1-12" disc (V-LM-1013). $4.85. (TNR Jan. '49).

Moussorgsky: *Pictures at an Exhibition.* Horowitz (piano). 1-12" disc (V-LM-1014). $5.45. (TNR Nov. '48).

Beethoven: *Sonata No. 1 in D, Op. 12, No. 1.* One side, and Beethoven: *Sonata No. 2 in A, Op. 12, No. 2.* Heifetz (violin) and Bay (piano). 1-12" disc (V-LM-1015). $5.45. (TNR Nov. '48).

Prokofiev: *Sonata No. 7, Op. 83.* One side, and Kabalevsky: *Sonata No. 3, Op. 46.* Horowitz (piano). 1-12" disc (V-LM-1016). $5.45 (TNR Apr. '46 & Mar. '49).

Bach: *Das Wohltemperierte Clavier* (Preludes and Fugues Nos. 1 to 8). Wanda Landowska (harpsichord). 1-12" disc (V-LM-1017). $5.45. (TNR Feb. '50).

Grieg: *Concerto in A minor, Op. 16.* Rubinstein (piano) with RCA Victor Orchestra—Dorati. One side, and Liszt: *Concerto No. 1 in E-flat.* Rubinstein (piano) with Dallas Symphony Orchestra—Dorati. 1-12" disc (V-LM-1018). $5.45. (TNR Jan. '50 & Oct. '47).

Tchaikovsky: *Romeo and Juliet.* One side, and Berlioz: *Romeo and Juliet.* NBC Symphony Orchestra—Toscanini. 1-12" disc (V-LM-1019). $5.45. (TNR Feb. '48 & Dec. '47).

Chopin: *Etudes, Op. 10* and Chopin: *Etudes, Op. 25.* Three sides, and Schumann: *Etudes Symphoniques, Op. 13.* Brailowski (piano). 2-12" discs (V-LM-6000). $10.90. (TNR Jan. '48 & Jan. '49).

Beethoven: *Symphony No. 9 in D minor, Op. 125.* ("Choral"). Three sides, and Beethoven: *Egmont Overture, Op. 84.* Soloists, Berkshire Festival Chorus, and Boston Symphony Orchestra—Koussevitzky. 2-12" discs (V-LM-6001). $10.90. (TNR Mar. '48 & Jan. '49).

Bach: *Mass in B minor.* Soloists, RCA Victor Chorale and Orchestra—Shaw. 3-12" discs (V-LM-6100). $16.35. (TNR Nov. '47).

Schumann: *Quartet No. 3 in A, Op. 41, No. 3.* Winterthur String Quartet. 1-12" disc (CH-CHC-38). $4.85.

PDF to image:

PDF

NewRecords.pdf

```python
from pathlib import Path
import fitz  # PyMuPDF
from tqdm import tqdm


pdf_path = Path("data/ocr/newrecords18.pdf")
outdir = Path("data/ocr/pages")
outdir.mkdir(parents=True, exist_ok=True)


doc = fitz.open(pdf_path)


for i, page in tqdm(enumerate(doc), desc="converting", total=len(doc)):
    pix = page.get_pixmap(dpi=300)
    pix.save(outdir / f"page_{i + 1:03d}.png")


doc.close()
```
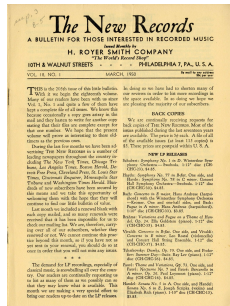
Transcribing with an API: Setup

- Step 1: Encode the image


```
"""
SETUP: Define OPENAI_API_KEY in your environment variables before running this script.
"""

from openai import OpenAI
from pathlib import Path
from dotenv import import load_dotenv
import base64

load_dotenv(override=True)

PAGES_ROOT = Path("data/ocr/pages")
OUT_ROOT = Path("data/ocr/transcriptions")
OUT_ROOT.mkdir(parents=True, exist_ok=True)

# Function to encode the image
def encode_image(image_path):
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode("utf-8")
```

iVBORw0KGgoAAAANSUhEUg
AAB90AAAqlCAIAAABhrjPEAA
AACXBIWXMAAC4jAAAuIwF4
pT92AJ1zpkIEQVR4nHS9B5M
cSZKIWf/5ZO9uZ8IMT3dVoUA
SyTkBSZCq7pG93bkh20XAEy
QJknMCliBJIpKAVfXM7T2zz+2
FRqBXxCTEw8PDibkR1adPn3
01//OZJ//0d9P/+vul2+dS+eXs2t
229XsXFn/+VuXpv/798s/fzv3r3
z/6b/9x8cevt+6f335wYYWei7dn
Di/rU1/U7363dPqPybqqbvxcP2l
Z++Xvvl2+Uf/zD3z3+7cfu7V7V/48d
7d+5ffD3Zu6qr/Le/Wf7xm1ePu1
3ef9Xfxhs3jn3PxouHnTrgzVTfy0d
dzyc6nk9c3L53/uWjioO5QRWd
QWV/dkDn105t7z7RYRe1vfe0
P+1/2rs32/d2puf1VJeKNl496dh
5cP71ZLfuZ+vuuc07Z3fuX9A2/
339pP/5gy7dlS6tv+u/u5PtzybO

Transcribing with an API: Setup

- Step 2: Define your prompt

```
# ___PROMPT___

prompt = """
transcribe the text on the provided document page image as accurately as possible.
"""
```

Transcribing with an API: Call

- Call to ChatGPT 5 mini:

User:
transcribe the text on
the provided
document page
image as accurately
as possible.

iVBORw0KGgoAAAANSUhEUg
AAB90AAAqICAIAAABhrjPEAA
AACXBIWXMAAC4jAAAuIwF4
pT92AJ1zpkl......

```python
pages = sorted(PAGES_ROOT.glob("*.png"))

client = OpenAI()

for page in pages:
    base64_image = encode_image(page)

    response = client.responses.create(
        model="gpt-5-nano",
        input=[
            {
                "role": "user",
                "content": [
                    {"type": "input_text", "text": prompt},
                    {
                        "type": "input_image",
                        "image_url": f"data:image/jpeg;base64,{base64_image}",
                    },
                ],
            }
        ],
    )

    out_path = OUT_ROOT / f"{page.stem}.txt"
    with open(out_path, "w", encoding="utf-8") as f:
        f.write(response.output_text)
```
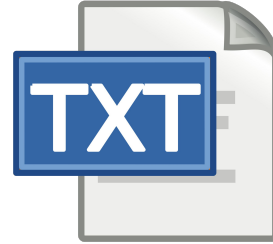
page_001.txt

Transcribing with an API: Results

The New Records
A BULLETIN FOR THOSE INTERESTED IN RECORDED MUSIC
Issued Monthly by
H. ROYER SMITH COMPANY
"The World's Record Shop"
10TH & WALNUT STREETS - - - PHILADELPHIA 7, PA., U.S.A.

VOL. 18, NO. 1 MARCH, 1950   By mail to any address
                                        50c per year

THIS is the 205th issue of this little bulletin. With it we begin the eighteenth volume. Many of our readers have been with us since Vol. 1, No. 1 and quite a few of them have kept a complete file of all issues. We know this because occasionally a copy goes astray in the mail and they hasten to write for another copy stating that their files are complete except for that one number. We hope that the present volume will prove as interesting to these oldtimers as the previous ones. During the last few months we have been advertising THE NEW RECORDS in a number of leading newspapers throughout the country including The New York Times, Chicago Tribune, Los Angeles Times, Boston Herald, Detroit Free Press, Cleveland Press, St. Louis Star Times, Cincinnati Enquirer, Minneapolis Star Tribune and Washington Times Herald. Hundreds of new subscribers have been secured by this means and we take this opportunity of welcoming them with the hope that they will continue to find our little bulletin of value.

Last month we included a renewal blank with each copy mailed, and so many renewals were received that it has been impossible for us to check our mailing list. We are, therefore, carrying over all of our subscribers, whether they renewed or not. We cannot continue this practice beyond this month, so if you have not as yet sent in your renewal, you should do so at once in order that you will not miss any issues.

* * * *

Cost: ~0.3 cents per page

Local OCR

1. find suitable model (trained for the purpose; fits on device)
2. download model and processor
3. format input
4. generate conditional on input
5. decode the output into text

Local OCR: find a suitable model

Models 313   Filter by name

deepseek-ai/DeepSeek-OCR
Image-Text-to-Text · 3B · Updated 5 days ago · ⬇ 1.34M · ⚡ · ♡ 2.23k

lightonai/LightOnOCR-1B-1025
Image-to-Text · Updated 1 day ago · ⬇ 4.85k · ♡ 110

ibm-granite/granite-docling-258M
Image-Text-to-Text · 0.3B · Updated Sep 23 · ⬇ 220k · ♡ 997

AvitoTech/avision
Image-Text-to-Text · 7B · Updated 7 days ago · ⬇ 214 · ♡ 17

nvidia/nemoretriever-ocr-v1
Image-to-Text · Updated 3 days ago · ⬇ 30 · ♡ 10

🤗 **Hugging Face**   Search models, datasets, users...

PaddlePaddle/**PaddleOCR-VL**   ♡ like 1.16k   Follow PaddlePaddle 721

Image-Text-to-Text   PaddleOCR   🎗 Safetensors   🌐 English   🌐 Chinese   🌐 multilingual   paddl

conversational   custom_code   arxiv:2510.14528   🏛 License: apache-2.0

📋 Model card   ·≣ Files and versions   ✖ xet   🟡 Community 53

PaddleOCR-VL: Boosting Multilingual Document Parsing via a 0.9B Ultra-Compact Vision-Language Model

⬡ Stars 62k   🤗 HuggingFace   ⟨·⟩ ModelScope   🤗 Demo on HuggingFace   ⟨·⟩ Demo on ModelScope

💬 Discord ERNIE   X PaddlePaddle   license Apache 2.0

🔥 **Official Demo**: Baidu AI Studio | 📝 **arXiv**: Technical Report

Local OCR: load the model and instantiate processor

```python
model_path = "PaddlePaddle/PaddleOCR-VL"
model = (
    AutoModelForCausalLM.from_pretrained(
        model_path, trust_remote_code=True, torch_dtype=torch.bfloat16
    )
    .to(DEVICE)
    .eval()
)
processor = AutoProcessor.from_pretrained(model_path, trust_remote_code=True)
```

OCR with local LLM


```python
image_dir = Path("data/ocr/pages")
image_paths = sorted(image_dir.glob("*.png"))
image = Image.open(image_paths[0]).convert("RGB")  # first image

messages = [
    {
        "role": "user",
        "content": [
            {"type": "image", "image": image},
            {"type": "text", "text": "OCR:"},
        ],
    }
]

inputs = processor.apply_chat_template(
    messages,
    tokenize=True,
    add_generation_prompt=True,
    return_dict=True,
    return_tensors="pt",
).to(DEVICE)
```
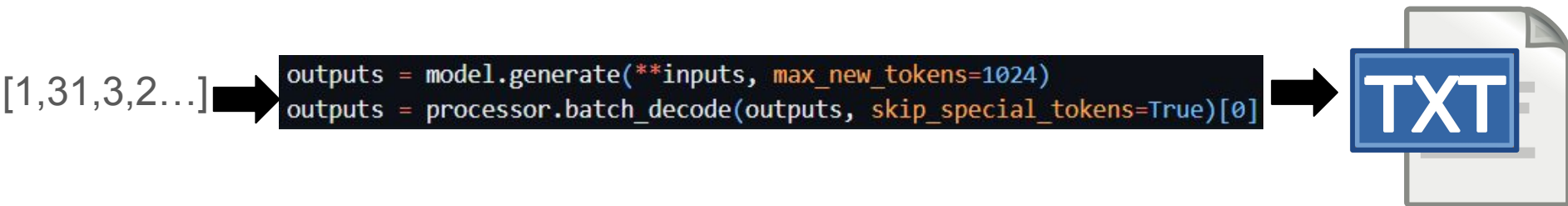
Daniel Göttlich, 26 minutes ago •

[1,31,3,2…]

OCR: process and decode

[1,31,3,2…] →

```
outputs = model.generate(**inputs, max_new_tokens=1024)
outputs = processor.batch_decode(outputs, skip_special_tokens=True)[0]
```

→ TXT

What to do with raw text: extracting structured info

Instructor: define what you want

```python
import instructor
from pydantic import BaseModel, Field


class SingleArtPiece(BaseModel):
    """Schema in which to return the information about a single art piece"""

    artist: str = Field(..., description="name of the artist")
    piece: str = Field(..., description="name of the art piece")
    year: Optional[int] = Field(None, description="year the art piece was created")
    place: Optional[str] = Field(
        None, description="place where the art piece was created or performed"
    )
    price: Optional[float] = Field(None, description="price of the art piece in USD")


class AllArtPieces(BaseModel):
    """schema in which to return all extracted information"""

    art_pieces: list[SingleArtPiece] = Field(
        ..., description="list of art pieces with structured information"
    )
```

Instantiate client, load descriptions, pass input & model

```python
all_art_pieces: list[SingleArtPiece] = []
for i, transcription in enumerate(transcriptions):
    print(f"Processing transcription {i + 1}/{len(transcriptions)}")

    response = client.responses.create(
        response_model=AllArtPieces,
        input=prompt.format(transcription=transcription),
    )

all_art_pieces.extend(response.art_pieces)
```

Example: flatten into a csv
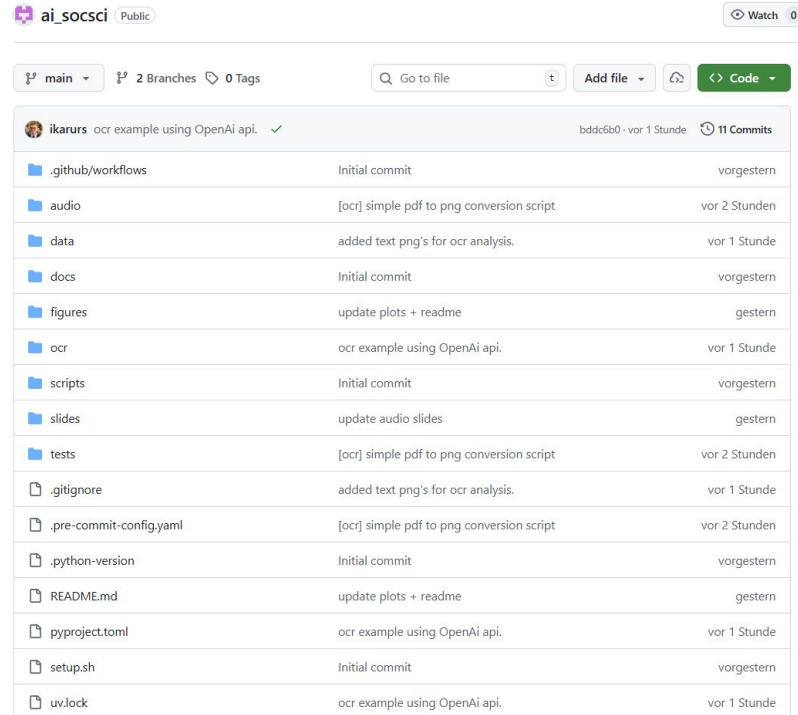
```
all_art_pieces: list[SingleArtPiece]
```

… loop

```python
df = pd.DataFrame([piece.model_dump() for piece in all_art_pieces])
df.to_csv(out_path, index=False)
```

artist	piece	year	place	price
Virgil Thom	Louisiana Story		Philadelph	3.5
Virgil Thom	Five Portraits		Philadelph	3.85
Leroy Ande	Irish Suite		Boston	4
Bizet	Carmen Suite			3.5
Tchaikovsk	Capriccio Italien, Op. 45			4.85
Gavin Gorc	The Rakeâ€™s Progres		Covent Gal	4.85
Tchaikows	Symphony No. 5 in E m		Berlin	4.85

For Future Reference:

https://github.com/DGoettlich/ai_socsci