



Multilingual systems with AI

News Sites & Telegram

Gruppe 3

Ahmad Ata Ul Haye

Clarissa Landzettel



News Sites & Video to Text



Agenda

- I. Aufgabenbeschreibung
- II. Umsetzung News Sites
 - Code Ausführung News Sites
 - Speicher Struktur News Sites
- III. JSON Datei Struktur
- IV. Umsetzung Video To Text
 - Code Ausführung Video To Text
 - Speicher Struktur Video To Text
- V. Evaluation
- VI. Mögliche Anwendung
- VII. Fazit
- VIII. Quellen

Aufgabenbeschreibung

- News Sites Extraktion von ARD und ZDF
- Text aus dem Videos extrahieren
- Export von Daten in den JSON Dateien



Abb. 3: Schemata zur Extraktion von Textinhalten von News Sites (Williams 2018)

Umsetzung News Sites

- Html-Inhalt mittels Selenium und Chrome
- Das Browsen ist im Headless-Modus möglich
- Cookies werden zuerst akzeptiert
- Bearbeitung der Inhalt mittels Beautiful Soup
- Struktur der Webseite mittels Beautiful Soup
- Top-to-bottom Ansatz bei der Bearbeitung



Abb. 4: Logo der Software „Selenium“ (Software Freedom 2022)



Abb. 5: Karikatur zur Informationsübermittlung (Richardson 2022)



Code Ausführung News Sites

- Erstellung der Objekt der Klasse WebSpider
- 5 Parameters
 - url,
 - browser_mode=True,
 - recursive=False,
 - crawl_portal=False,
 - data_saving_Directory= 'DataOutput'
- Zwei Rekursions Tiefe sind definiert
 - Ganzen Portal
 - Bestimmte Gebiet
- Extraktion folgt durch rekursiven Aufruf von WebSpider Klasse

Aufruf von ZDF Nachrichten-Politik

```
def extract_news():  
    base_url = 'https://www.zdf.de/nachrichten/politik'  
    news_extractor = WebSpider(url=base_url, recursive=True, data_saving_Directory="Nachrichten-Politik")  
    news_extractor.scrape_data()  
    video_link_file = news_extractor.write_video_links()  
  
    video_links = []  
    try:  
        file = open(video_link_file, 'r')  
        video_links = file.readlines()  
    except Exception as e:  
        print(e)  
  
    for video in video_links:  
        video_to_text(video.strip())
```


Speicher Struktur News Sites

- Erstellung von Verzeichnis pro Aufruf
- Erstellung von JSON Datei pro Webseite
- Z. B. unter Nachrichten-Politik 96 Pages und 96 JSON Dateien

```
▼ Nachrichten_Politik
  {} zdf_denachrichtenpolitik.json U
  {} zdf_denachrichtenpolitik#main-content.json U
  {} zdf_denachrichtenpolitik#skiplinks.json U
  {} zdf_denachrichtenpolitik#top-search-input.json U
  {} zdf_denachrichtenpolitikanschlag-linke-oberhausen-100_html.json U
  {} zdf_denachrichtenpolitikanschlag-linke-oberhausen-100_html#main-content.json U
  {} zdf_denachrichtenpolitikanschlag-linke-oberhausen-100_html#skiplinks.json U
  {} zdf_denachrichtenpolitikanschlag-linke-oberhausen-100_html#top-search-input.json U
  {} zdf_denachrichtenpolitikchancen-aufenthaltsrecht-bleiberecht-kabinett-100_html.js... U
  {} zdf_denachrichtenpolitikchancen-aufenthaltsrecht-bleiberecht-kabinett-100_html#... U
  {} zdf_denachrichtenpolitikchancen-aufenthaltsrecht-bleiberecht-kabinett-100_html#s... U
  {} zdf_denachrichtenpolitikchancen-aufenthaltsrecht-bleiberecht-kabinett-100_html#t... U
  {} zdf_denachrichtenpolitikcorona-abwasser-analyse-faq-100_html.json U
  {} zdf_denachrichtenpolitikcorona-abwasser-analyse-faq-100_html#main-content.json U
  {} zdf_denachrichtenpolitikcorona-abwasser-analyse-faq-100_html#skiplinks.json U
  {} zdf_denachrichtenpolitikcorona-abwasser-analyse-faq-100_html#top-search-input.j... U
  {} zdf_denachrichtenpolitikcorona-impfung-daten-100_html.json U
  {} zdf_denachrichtenpolitikcorona-impfung-daten-100_html#main-content.json U
```


JSON Struktur

- Links Schlüssel haben der Form: LnkS_XXXXXXXXXX_LnkE__Name
- Bilder Schlüssel haben der Form: ImgS_XXXXXXXXXX_ImgE__Name
- Texts Schlüssel haben der Form: PS_XXXXXXXXXX_PE_Name
- Video Schlüssel haben der Form: VideoLnkS_XXXXXXXXXX_VideoLnkE

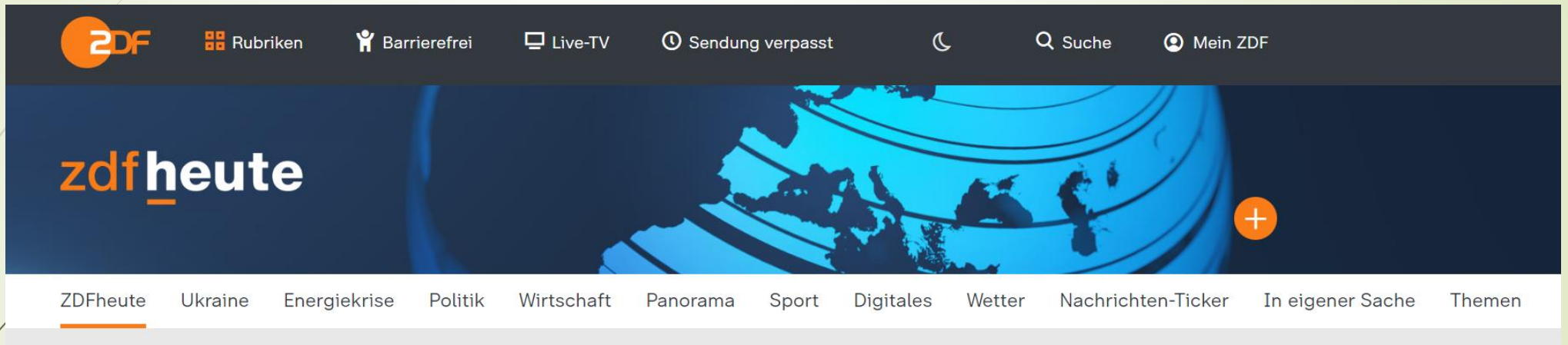
```
"LnkS_VMJG055MQQ_LnkE__Nachrichten": "https://www.zdf.de/nachrichten",
```

```
"ImgS_J5GLWMSICH_ImgE__ZDFheute": {  
  "Src": "https://www.zdf.de/assets/zdfheute-keyvisual-100~2850x300?cb=1644916262455",  
  "data-src": "https://www.zdf.de/assets/zdfheute-keyvisual-100~2850x300?cb=1644916262455"  
},
```

```
"PS_00B1S8JINY_PE": {  
  "paragraph-content": "Russland will das größte Kernkraftwerk an das Stromnetz der  
}
```

```
"VideoLnkS_7WTCLQIILE_VideoLnkE": {  
  "video-source": "https://nrodlzdf-a.akamaihd.net/none/zdf/22/08/220808\_gesamt\_hli/1/220808\_gesamt\_hli\_2128k\_p18v15.webm",  
  "video-description": "Erneut laden Video spielt auf Google Cast ab. Nachrichten | ZDFheute live AKW in der Ukraine unter  
},
```

- Navigations Menü



```
"ZDFheute-Ukraine-Energiekrise-Politik-Wirtschaft-Panorama-Sport-Digitales-Wetter-Nachrichten-Ticker-In-eigener-Sache-Themen": {  
  "LnkS_LFTQEK55N9_LnKE__ZDFheute": "https://www.zdf.de/nachrichten",  
  "LnkS_JLJKRMTI9I_LnKE__Ukraine": "https://www.zdf.de/nachrichten/thema/ukraine-198.html",  
  "LnkS_PCH9HJHJSR_LnKE__Energiekrise": "https://www.zdf.de/nachrichten/thema/energiesparen-100.html",  
  "LnkS_QVA06TWKC4_LnKE__Politik": "https://www.zdf.de/nachrichten/politik",  
  "LnkS_ZCHEFWQQGL_LnKE__Wirtschaft": "https://www.zdf.de/nachrichten/wirtschaft",  
  "LnkS_7DSNYM4I5S_LnKE__Panorama": "https://www.zdf.de/nachrichten/panorama",  
  "LnkS_SXYRYI2RMX_LnKE__Sport": "https://www.zdf.de/nachrichten/sport",  
  "LnkS_V6N99X5K69_LnKE__Digitales": "https://www.zdf.de/nachrichten/digitales",  
  "LnkS_BFPYGBV0KY_LnKE__Wetter": "https://www.zdf.de/nachrichten/wetter",  
  "LnkS_Z58KLMCL7N_LnKE__Nachrichten-Ticker": "https://www.zdf.de/nachrichten/nachrichtenticker-100.html",  
  "LnkS_CDGNZK4YDA_LnKE__In eigener Sache": "https://www.zdf.de/nachrichten/in-eigener-sache",  
  "LnkS_VXF77HE177_LnKE__Themen": "https://www.zdf.de/nachrichten/thema"  
},  
"LnkS_Z58KLMCL7N_LnKE__Nachrichten-Ticker": "https://www.zdf.de/nachrichten/nachrichtenticker-100.html"
```

ARD

```
{
  "Title": "Krieg gegen die Ukraine: ++ UN erwarten steigende Getreideausfuhren ++ | tagesschau.de",
  "Navigation-Inhalt-Fußzeile": {
    "Lnks_9Z98M4I7ZS_LnKE_Navigation": "https://www.tagesschau.de/newsticker/liveblog-ukraine-mittwoch-149.html#navigation",
    "Lnks_WMPKMJ5QKW_LnKE_Inhalt": "https://www.tagesschau.de/newsticker/liveblog-ukraine-mittwoch-149.html#content",
    "Lnks_MPL9ATQV3F_LnKE_Fußzeile": "https://www.tagesschau.de/newsticker/liveblog-ukraine-mittwoch-149.html#footer"
  },
  "Lnks_7HLSOX8TI2_LnKE_Tagesschau": "https://www.tagesschau.de",
  "Suchbegriff-Suche-Inland-Startseite-Inland-Innenpolitik-Gesellschaft-Regional-DeutschlandTrend-Wahlen-Mittendrin-Ausland-Startseite-Ausland-Eur
    "Lnks_X54TIDA660_LnKE_Startseite Inland": "https://www.tagesschau.de/inland/",
    "Lnks_TT1MK7DGT3_LnKE_Innenpolitik": "https://www.tagesschau.de/inland/innenpolitik/",
    "Lnks_YSDDOXMMC7_LnKE_Gesellschaft": "https://www.tagesschau.de/inland/gesellschaft/",
    "Lnks_FE6LMN5M37_LnKE_Regional": "https://www.tagesschau.de/regional/",
    "Lnks_90LPBL16DF_LnKE_DeutschlandTrend": "https://www.tagesschau.de/inland/deutschlandtrend/",
    "Lnks_ULHSMGIIGS_LnKE_Startseite Wahlergebnisse": "https://www.tagesschau.de/wahl/",
    "Lnks_L2LGN517H5_LnKE_Mittendrin": "https://www.tagesschau.de/inland/mittendrin/",
    "Lnks_75JP4EA5YF_LnKE_Startseite Ausland": "https://www.tagesschau.de/ausland/",
    "Lnks_BTL6Z1LHL2_LnKE_Europa": "https://www.tagesschau.de/ausland/europa/",
    "Lnks_IYXX7G8YI2_LnKE_Amerika": "https://www.tagesschau.de/ausland/amerika/",
    "Lnks_10JZX4SWZ8_LnKE_Afrika": "https://www.tagesschau.de/ausland/afrika/",
    "Lnks_2AI4P07PAX_LnKE_Asien": "https://www.tagesschau.de/ausland/asien/",
    "Lnks_57IUBKNMKG_LnKE_Ozeanien": "https://www.tagesschau.de/ausland/ozeanien/",
    "Lnks_ELXSW00F94_LnKE_Startseite Wirtschaft": "https://www.tagesschau.de/wirtschaft/",
    "Lnks_MOPNLUU90X_LnKE_Börsenkurse": "https://www.tagesschau.de/wirtschaft/boersenkurse/",
    "Lnks_9W9CP2PAZM_LnKE_Financen": "https://www.tagesschau.de/wirtschaft/financen/",
    "Lnks_BCB3UCND8V_LnKE_Unternehmen": "https://www.tagesschau.de/wirtschaft/unternehmen/",
    "Lnks_UCU2UVITC0_LnKE_Verbraucher": "https://www.tagesschau.de/wirtschaft/verbraucher/",
    "Lnks_5YPZ2TAX0I_LnKE_Technologie": "https://www.tagesschau.de/wirtschaft/technologie/",
    "Lnks_BYD5JXP5HW_LnKE_Konjunktur": "https://www.tagesschau.de/wirtschaft/konjunktur/",
    "Lnks_ND800TJ6VF_LnKE_Weltwirtschaft": "https://www.tagesschau.de/wirtschaft/weltwirtschaft/",
    "Lnks_0ZSAKT3VX4_LnKE_Podcast": "https://www.tagesschau.de/faktenfinder/podcast/",
    "Lnks_GQZNQEICAS_LnKE_Startseite Wetter": "https://www.tagesschau.de/wetter/",
    "Lnks_EJ25MRLZII_LnKE_Deutschland": "https://wetter.tagesschau.de/deutschland/",
    "Lnks_K722TMIQ2R_LnKE_Unwetterwarnungen": "https://wetter.tagesschau.de/unwetter/",
    "Lnks_AKQAJZ8BW1_LnKE_Europa & Welt": "https://wetter.tagesschau.de/europawelt/",
    "Lnks_U6JYER0SPT_LnKE_Übersicht der Wahlen seit 1946": "https://www.tagesschau.de/wahl/uebersicht-der-wahlen.shtml",
    "Lnks_J5PXL5GGY8_LnKE_Länderparlamente": "https://www.tagesschau.de/wahlarchiv/laenderparlamente/",
    "Lnks_U1F9XDYDLP_LnKE_Bundestagswahl": "https://www.tagesschau.de/wahlarchiv/bundestag/",
    "Lnks_8SSZB5IFMC_LnKE_Europäisches Parlament": "https://www.tagesschau.de/wahlarchiv/europaeisches_parlament/",
    "Lnks_D72D933Q9A_LnKE_Chronologie": "https://www.tagesschau.de/wahlarchiv/chronologie/",
    "Lnks_771ZNEW1XM_LnKE_Wahltermine": "https://www.tagesschau.de/wahlarchiv/wahltermine/",
    "Lnks_62PV0P5YH8_LnKE_Startseite Videos & Audios": "https://www.tagesschau.de/multimedia/",
  }
```

ZDF

```
{
  "Title": "Ukraine: Russland will Saporischschja an Krim anbinden - ZDFheute",
  "Zum-Hauptinhalt-springen-Zur-Suche-springen-Hauptnavigation-Startseite-Rubriken-Filme-Serien-Comedy-&-Satire-heute-Nachrichten-Politik-&-Gesellschaft": "https://www.zdf.de/nachrichten/politik/atomkraftwerk-saporischschja-krim-ukraine-krieg-russland",
  "Lnks_FKG8TVDZ0N_Lnke_Zum Hauptinhalt springen": "https://www.zdf.de/nachrichten/politik/atomkraftwerk-saporischschja-krim-ukraine-krieg-russland",
  "Lnks_KN6KDS3ER4_Lnke_Zur Suche springen": "https://www.zdf.de/nachrichten/politik/atomkraftwerk-saporischschja-krim-ukraine-krieg-russland",
  "Lnks_QLWQ06KB70_Lnke_Startseite": "/",
  "Lnks_OM8H9P86TG_Lnke_Filme": "https://www.zdf.de/filme",
  "Lnks_NKDGF16IXA_Lnke_Serien": "https://www.zdf.de/serien",
  "Lnks_NPXJRTZFSH_Lnke_Comedy & Satire": "https://www.zdf.de/comedy",
  "Lnks_5U31F45JLB_Lnke_heute-Nachrichten": "https://www.zdf.de/nachrichten",
  "Lnks_CA366UMFT5_Lnke_Politik & Gesellschaft": "https://www.zdf.de/politik-gesellschaft",
  "Lnks_5WARF6AGDX_Lnke_Sport": "https://www.zdf.de/sport",
  "Lnks_CNAKKZJBSP_Lnke_Dokus & Reportagen": "https://www.zdf.de/doku-wissen",
  "Lnks_96TTD2HT13_Lnke_Kultur": "https://www.zdf.de/kultur",
  "Lnks_E7PQONVWPE_Lnke_ZDFtivi für Kinder": "https://www.zdf.de/kinder",
  "Lnks_I39KFXFTSW_Lnke_Sendungen A-Z": "https://www.zdf.de/sendungen-a-z",
  "Lnks_F63JFOLDE8_Lnke_TV-Programm": "https://www.zdf.de/live-tv",
  "Lnks_MNOD2BVYZ1_Lnke_Karriere im ZDF": "https://www.zdf.de/zdfunternehmen/karriere-112.html",
  "Lnks_0JDO2ZMCQY_Lnke_Barrierefrei": "https://www.zdf.de/barrierefreiheit-im-zdf",
  "Lnks_DAYFYRRJFE_Lnke_Sendung verpasst": "https://www.zdf.de/sendung-verpasst",
  "Lnks_57ZI7DNMV8_Lnke_Meine Merkliste": "https://www.zdf.de/mein-zdf#merkliste",
  "Lnks_71K3NNG8L9_Lnke>Weiterschauen": "https://www.zdf.de/mein-zdf#weiterschauen",
  "Lnks_8TMJK4UCOZ_Lnke_Benachrichtigungen": "https://www.zdf.de/mein-zdf#benachrichtigungen",
  "Lnks_9ADORIOPAV_Lnke Mein ZDFtivi": "https://www.zdf.de/kinder/mein-zdftivi",
  "Lnks_52VAEHJ6LK_Lnke_Einstellungen": "https://www.zdf.de/mein-zdf#mein-zdf-profil"
},
  "Hauptnavigation-Startseite-Rubriken-Filme-Serien-Comedy-&-Satire-heute-Nachrichten-Politik-&-Gesellschaft-Sport-Dokus-&-Reportagen-Kultur-ZDFtivi": "https://www.zdf.de",
  "Lnks_CP3AHK776C_Lnke_Startseite": "/",
  "Lnks_ZIN3GKIYE9_Lnke_Filme": "https://www.zdf.de/filme",
  "Lnks_00P3741QYX_Lnke_Serien": "https://www.zdf.de/serien",
  "Lnks_I09Z137T2U_Lnke_Comedy & Satire": "https://www.zdf.de/comedy",
  "Lnks_BF12E6DWME_Lnke_heute-Nachrichten": "https://www.zdf.de/nachrichten",
  "Lnks_LYRCP8J66U_Lnke_Politik & Gesellschaft": "https://www.zdf.de/politik-gesellschaft",
  "Lnks_Q24P4LEIWL_Lnke_Sport": "https://www.zdf.de/sport",
  "Lnks_Y8K833Q180_Lnke_Dokus & Reportagen": "https://www.zdf.de/doku-wissen",
  "Lnks_41XM6CDWBZ_Lnke_Kultur": "https://www.zdf.de/kultur",
}
```


Umsetzung Video To Text

- Umwandlung der Video-Dateien in Audio mittels MoviePy
- Splitten der Audio-Dateien nach Pausen (Stille)
- Bearbeitung der einzelnen Audio-Chunks mittels Python Google Speech recognition API


```
In [12]: from moviepy.editor import *  
         # load the clip, rotate it 180°, and display  
         clip = VideoFileClip("lake.mp4").rotate(180)  
         clip.ipython_display(width=280)  
  
Out[12]: 
```

Abb. 6: Beispiel Code von MoviePy (Zulko 2017)



Code Ausführung VideoToText

- Erstellung der Objekt der Klasse VideoToText
- 6 Parameters
 - video-source
 - audio_format='wav',
 - language="de-DE",
 - min_silence_len=500,
 - silence_thresh=-36,
 - keep_silence=400
- Pro Aufruf erfolgt Text Extraktion für eine Video Datei
- Parallele Bearbeitung abhängig der Hardware und Betriebssystem



Aufruf von Video-To-Text

```
def video_to_text(video_link):  
    video_text = VideoToText(video_link)  
    try:  
        video_text.video_to_audio()  
        video_text.split_audio_file_on_silence()  
        video_text.read_text_parallel()  
    except Exception as e:  
        print("The following Error occurred while converting speech to text:", e)  
    finally:  
        video_text.delete_directory()
```


Speicher Struktur VideoToText

- Das Audio Datei wird heruntergeladen
- Nach Bearbeitung folgt die Daten Löschung
- Speicherung der Videolinks sowie extrahierte Text in den Text Dateien

```
zdf_denachrichtenpolitikatomkraftwerk-saporischschja-krim-ukraine-krieg-russland-100_html_VideoLinks.text
1 https://nrodlzdf-a.akamaihd.net/none/zdf/22/08/220810_jaeger_hie/1/220810_jaeger_hie_2128k_p18v15.webm
2 https://nrodlzdf-a.akamaihd.net/none/zdf/22/08/220809_saporischschja_xpr/1/220809_saporischschja_xpr_2128k_p18v15.webm
3 https://nrodlzdf-a.akamaihd.net/none/zdf/22/08/220808_gesamt_hli/1/220808_gesamt_hli_2128k_p18v15.webm
4
```

Screenshot

```
Making Directory
Directory has been created
Starting converting video to audio
MoviePy - Writing audio in 220609_1900_clip_3_h19_2128k_p18v15.wav
MoviePy - Done.
Video to audio conversion is finished
Started audio splitting on silence
Audio splitting finished.It splitted in: 27 chunks
Exported: 27 chunks to the specified directory
Start converting speech to text
Started processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__0.wav
Finished processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__0.wav
Started processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__1.wav
Finished processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__1.wav
Started processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__10.wav
```

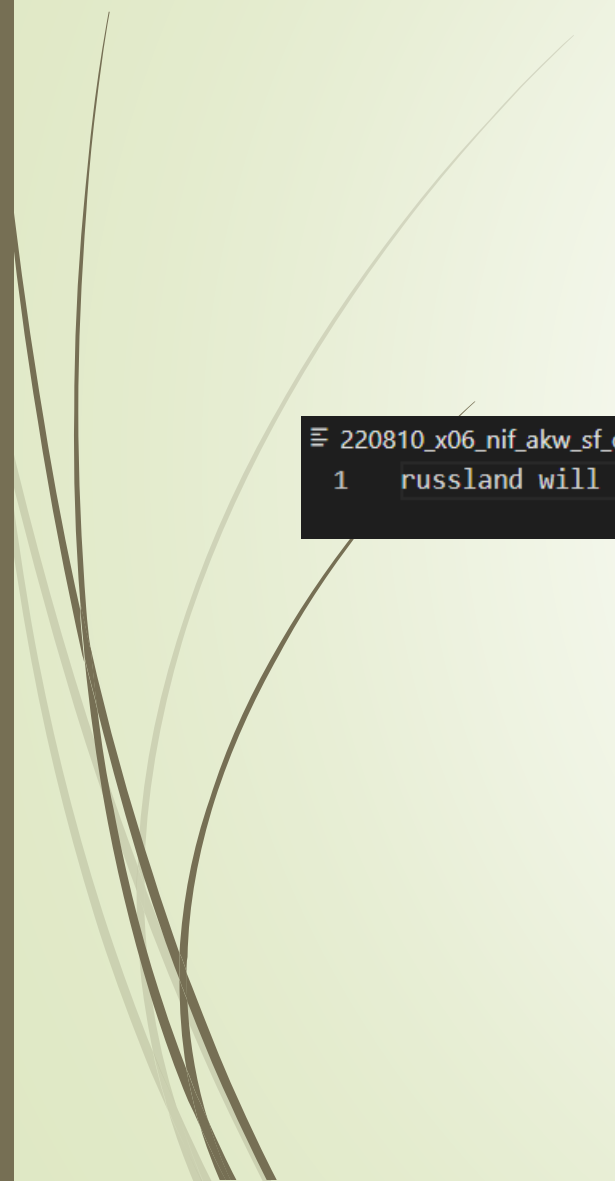

```
Finished processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__7.wav
Started processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__8.wav
Finished processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__8.wav
Started processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__9.wav
Finished processing audio chunk: 220609_1900_clip_3_h19_2128k_p18v15__9.wav
Could not extract text from 0 audio chunks
Finished converting speech to text
Deleting directory
Directory has been deleted!
```

Video to Text

https://nrodlzdf-a.akamaihd.net/none/zdf/22/06/220609_1900_clip_3_h19/1/220609_1900_clip_3_h19_2128k_p18v15.webm (Video Länge 2 Minuten)

≡ recognized.txt

```
1 einen Tag nach der tödlichen Autofahrt in Berlin mit einer toten und inzwischen mehr als 30 Verletzten sind neue Details bekannt
2 Polizei und Staatsanwaltschaft gehen weder von einer Terror Tat noch von einem Unfall aus sondern von einer schweren psychischen Erkrankung des Fahrers er soll jetzt in e
3 alles erdenklich Gute
4 auch heute noch bleibt unklar warum der Täter mit dem kleinen Wagen in die Personen gerast
5 weitergefahren und schließlich in ein Geschäft gekracht ist
6 Mord und versuchten Mord wirft die Staatsanwaltschaft dem Mann der noch am Tatort festgenommen wurde jetzt vor und spricht von seinen bekanntgewordenen psychischen Probleme
7 insofern spricht
8 recht viel für eine paranoide Schizophrenie was ist jedenfalls nicht gibt sind Anhaltspunkte für irgendeine Art von
9 terroristischen Hintergrund bei dem ganzen Geschehen
10 dann aber immer auch einen Unfall wird sich vor diesem Hintergrund
11 ausschließen lassen
12 Trauer derweil in Bad Arolsen hier kommt die Klasse her viele Menschen in der Stadt tief bewegt
13 Stephanie Hein
14 Fernspäher und traurige Tage
15 Bad Arolsen
16 dass die Nachricht die uns gestern hier ereilt hat
17 Bitburger WLAN
18 bis heute Abend erfahren haben gestern hat Bad Arolsen schwer getroffen ich habe das in allen Gesprächen die führen konnte
19 erfahren
20 der Täter soll in der Psychiatrie untergebracht werden der Lehrer ist dort Staatsanwaltschaft auch heute noch in Lebensgefahr
21 die Spuren der Tat sind noch gut sichtbar Berlin City-West am Tag danach
22 hier starb eine Lehrerin viele ihre Schüler zum Teil schwer verletzt auch ihr Kollege
23 Bundesfamilienministerin Paus Berlins Polizeipräsidentin und die Bundesinnenminister rin zeigen Mitgefühl
24 ich bin hier um
25 die fünf Hunde Anteilnahme der Bundesregierung auszudrücken mit vor allen Dingen den Angehörigen
26 der Getöteten Lehrerinnen und den vielen Verletzten wir wünschen den
27 verletzten Schülerinnen und Schülern und den anderen Verletzten Passanten
28
```



≡ 220810_x06_nif_akw_sf_onl_2128k_p18v15_VideoScript.txt

1 russland will das besetzte ukrainische Atomkraftwerk saporischschja an das Stromnetz der annektierten Halbinsel Krim anschließen




Evaluation

1. Erfolgreiche Umsetzung

- ▀ News Extraktion
- ▀ Video To Text Extraktion

2. Grenzen

- ▀ Laufzeit der Aufruf von ganzen Portal



3. Open Issues

- Die Videos von ARD lassen nicht umwandeln im Audio Dateien
- Mögliche Lösungen:
 - Video zuerst runterladen und dann bearbeiten => rechtliche Erlaubt?
 - Video mithören dann das Audio bearbeiten => lange Laufzeit

MoviePy Error

```
Making Directory
Directory has been created
Starting converting video to audio
The following Error occurred: MoviePy error: failed to read the duration of file https://www.ardmediathek.de/video/jagd-auf-dagobert/folge-2-schlauer-als-die-polizei-erlaubt/rbb-fernsehen/Y3JpZDovL3JiYi1vbmxpbmUuZGUvamFnZC1hdWYtZGFnb2JlcnQvMjAyMi0wNi0wNi0xMDozMDowMF9mN2IwZmY1OC1hMDExLTQ3MmMtOWMyNi1kN2RjMTU1NjE2NjgvaWFnZC1hdWYtZGFnb2JlcnRfMjAyMjA2MDZfZm9sZ2VfMg.
Here are the file infos returned by ffmpeg:

ffmpeg version 4.2.2 Copyright (c) 2000-2019 the FFmpeg developers
  built with gcc 9.2.1 (GCC) 20200122
  configuration: --enable-gpl --enable-version3 --enable-sdl2 --enable-fontconfig --enable-gnutls --enable-iconv --enable-libass --enable-libdav1d --enable-libbluray --enable-libfreetype --enable-libmp3lame --enable-libopencore-amrnb --enable-libopencore-amrwb --enable-libopenjpeg --enable-libopus --enable-libshine --enable-lbsnappy --enable-libsoxr --enable-libtheora --enable-libtwolame --enable-libvpx --enable-libwavpack --enable-libwebp --enable-libx264 --enable-libx265 --enable-libxml2 --enable-libzimg --enable-lzma --enable-zlib --enable-gmp --enable-libvidstab --enable-libvorbis --enable-libvo-amrwbenc --enable-libmysofa --enable-libspeex --enable-libxvid --enable-libaom --enable-libmfx --enable-amf --enable-ffnvcodec --enable-cuvid --enable-d3d11va --enable-nvenc --enable-nvdec --enable-dxva2 --enable-avisynth --enable-libopenmpt
  libavutil      56. 31.100 / 56. 31.100
  libavcodec     58. 54.100 / 58. 54.100
  libavformat    58. 29.100 / 58. 29.100
  libavdevice    58.  8.100 / 58.  8.100
  libavfilter    7. 57.100 / 7. 57.100
  libswscale     5.  5.100 / 5.  5.100
  libswresample  3.  5.100 / 3.  5.100
  libpostproc   55.  5.100 / 55.  5.100
https://www.ardmediathek.de/video/jagd-auf-dagobert/folge-2-schlauer-als-die-polizei-erlaubt/rbb-fernsehen/Y3JpZDovL3JiYi1vbmxpbmUuZGUvamFnZC1hdWYtZGFnb2JlcnQvMjAyMi0wNi0wNi0xMDozMDowMF9mN2IwZmY1OC1hMDExLTQ3MmMtOWMyNi1kN2RjMTU1NjE2NjgvaWFnZC1hdWYtZGFnb2JlcnRfMjAyMjA2MDZfZm9sZ2VfMg: Invalid data found when processing input
```

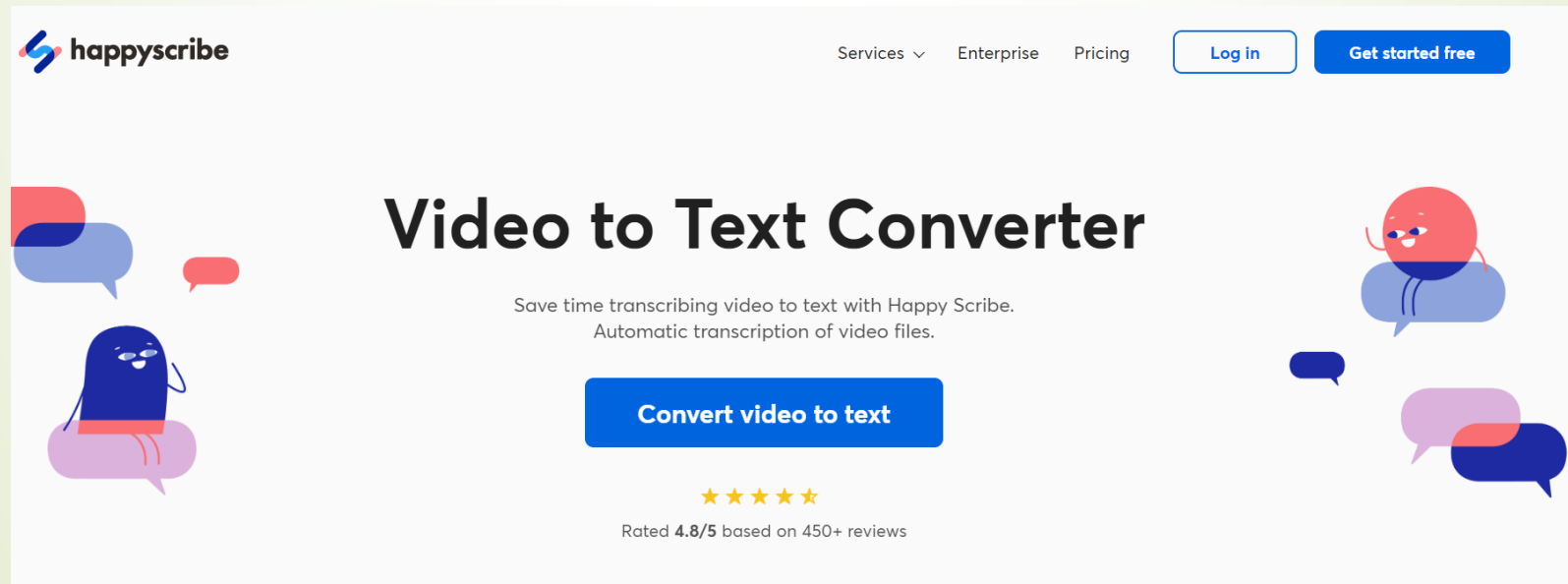

subprocess Module Error

- Umwandlung von Video nach binär Datei dann nach Audio Datei

```
ffmpeg version 2022-06-09-git-5d5a014199-full_build-www.gyan.dev Copyright (c) 2000-2022 the FFmpeg developers
  built with gcc 11.3.0 (Rev1, Built by MSYS2 project)
  configuration: --enable-gpl --enable-version3 --enable-static --disable-w32threads --disable-autodetect --enable-fontconfig --enable-iconv --enable-gnutls --enable-libxml2 --enable-gmp --enable-bzlib --enable-lzma --enable-libsnappp --enable-zlib --enable-librist --enable-libsrt --enable-libssh --enable-libzmq --enable-avisynth --enable-libbluray --enable-libcaca --enable-sdl2 --enable-libdav1d --enable-libdav1s --enable-libdav1s2 --enable-libuavs3d --enable-libzvti --enable-librav1e --enable-libsvtav1 --enable-libwebp --enable-libx264 --enable-libx265 --enable-libxavs2 --enable-libxvid --enable-libaom --enable-libjxl --enable-libopenjpeg --enable-libvpx --enable-mediafoundation --enable-libass --enable-frei0r --enable-libfreetype --enable-libfribidi --enable-liblensfun --enable-libvidstab --enable-libvmaf --enable-libzimg --enable-amf --enable-cuda-llvm --enable-cuvid --enable-ffnvcodec --enable-nvdec --enable-nvenc --enable-d3d11va --enable-dxva2 --enable-libmfx --enable-libshaderc --enable-vulkan --enable-libplacebo --enable-openccl --enable-libcdio --enable-libgme --enable-libmodplug --enable-libopenmpt --enable-libopencore-amrwb --enable-libmp3lame --enable-libshine --enable-libtheora --enable-libtwolame --enable-libvo-amrwbenc --enable-libilbc --enable-libgsm --enable-libopencore-amrnb --enable-libopus --enable-libspeex --enable-libvo-aacenc --enable-libvo-amrnb --enable-libvo-amrwb --enable-ladspa --enable-libbs2b --enable-libflite --enable-libmysofa --enable-librubberband --enable-libsoxr --enable-chromaprint
  libavutil      57. 26.100 / 57. 26.100
  libavcodec     59. 33.100 / 59. 33.100
  libavformat    59. 24.100 / 59. 24.100
  libavdevice    59.  6.100 / 59.  6.100
  libavfilter     8. 40.100 /  8. 40.100
  libswscale     6.  6.100 /  6.  6.100
  libswresample  4.  6.100 /  4.  6.100
  libpostproc   56.  5.100 / 56.  5.100
[mov,mp4,m4a,3gp,3g2,mj2 @ 000001b051d63d80] Format mov,mp4,m4a,3gp,3g2,mj2 detected only with low score of 1, misdetection possible!
[mov,mp4,m4a,3gp,3g2,mj2 @ 000001b051d63d80] moov atom not found
test.mp4: Invalid data found when processing input
```

Mögliche Anwendung

- Nachrichten Analyse
- In der Praxis: Video to Text



The screenshot displays the Happy Scribe website's landing page for its Video to Text Converter. The header includes the Happy Scribe logo, navigation links for 'Services', 'Enterprise', and 'Pricing', and buttons for 'Log in' and 'Get started free'. The main heading is 'Video to Text Converter', followed by the tagline 'Save time transcribing video to text with Happy Scribe. Automatic transcription of video files.' A prominent blue button labeled 'Convert video to text' is centered below the text. Underneath the button is a five-star rating and the text 'Rated 4.8/5 based on 450+ reviews'. The page is decorated with colorful cartoon characters and speech bubbles on both sides.

happyscribe

Services ▾ Enterprise Pricing

Log in Get started free

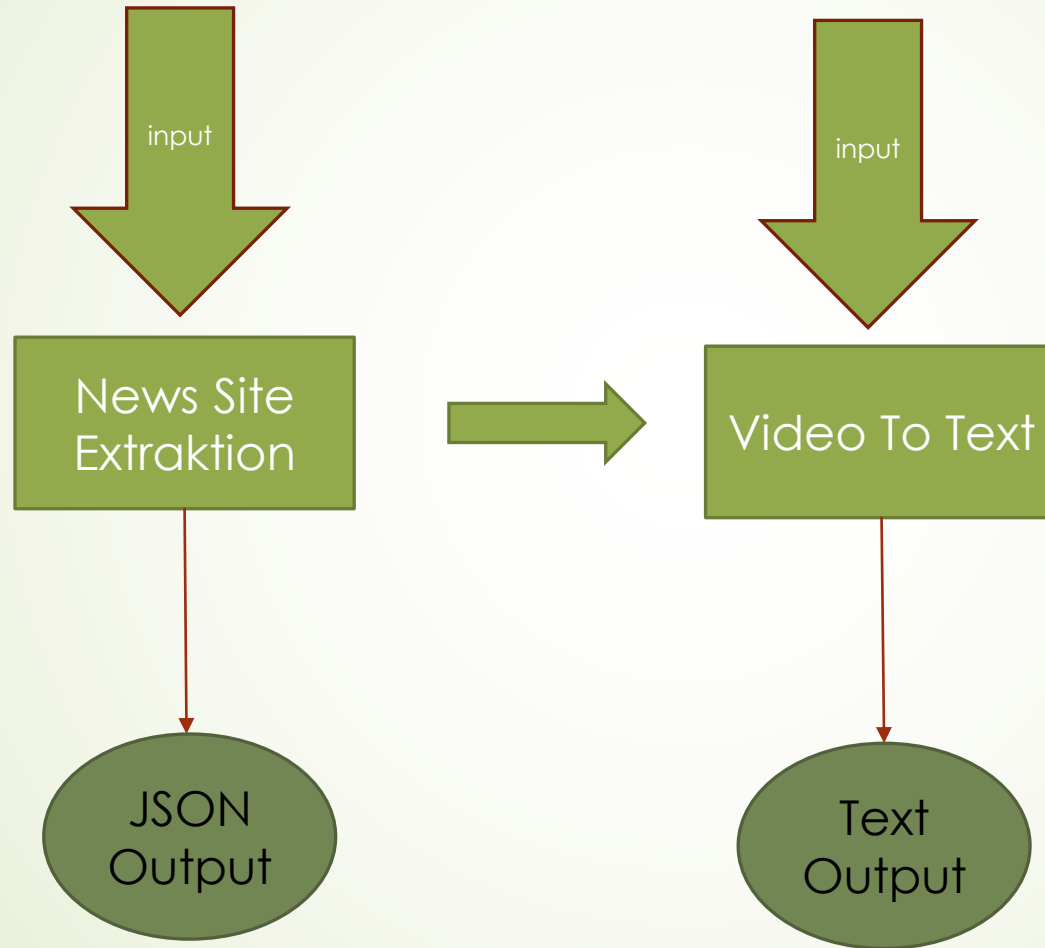
Video to Text Converter


Save time transcribing video to text with Happy Scribe.
Automatic transcription of video files.

Convert video to text

★★★★★
Rated 4.8/5 based on 450+ reviews

Fazit





```
def extract_news():
    base_url = 'https://www.zdf.de/nachrichten/politik'
    news_extractor = WebSpider(url=base_url, recursive=True, data_saving_Directory="Nachrichten-Politik")
    news_extractor.scrape_data()
    video_link_file = news_extractor.write_video_links()

    video_links = []
    try:
        file = open(video_link_file, 'r')
        video_links = file.readlines()
    except Exception as e:
        print(e)

    for video in video_links:
        video_to_text(video.strip())
```



Vielen Dank für Ihre Aufmerksamkeit

Haben Sie Fragen?



Quellen

- Richardson, Leonard (2022). Beautiful Soup Documentation. URL: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (16.06.2022).
- Software Freedom (2022). Selenium. URL: <https://www.selenium.dev/> (16.06.2022).
- Williams, Janet (2018). *What is Web Scraping?* URL: <https://www.promptcloud.com/blog/what-is-web-scraping/> (16.06.2022).
- ZDF (2022). *ZDF 19 Uhr Nachrichten, 22.06.2009* (Upload). URL: https://nrodlzdf-a.akamaihd.net/none/zdf/22/06/220609_1900_clip_3_h19/1/220609_1900_clip_3_h19_2128k_p18v15.webm (16.06.2022).
- Zhang, Anthony (2022). SpeechRecognition 3.8.1. URL: <https://pypi.org/project/SpeechRecognition/> (16.06.2022).
- Zulko (2017). MoviePy Doc. URL: <https://zulko.github.io/moviepy/> (16.06.2022).