Photo: Mohammad Metri, unsplash.com

Podcast discovery (still) is broken – let's improve it with NLP and Machine Learning

Diplompräsentation DAS Data Science FHNW
Student: Patrick Arnecke
Dozent: Fernando Benites
Studiengangsleiter: Michael Henninger

April 2022

Intro

Problemstellung

- Podcast sind boomendes Medium, enorme Inhaltsfülle
- ~2 Mio. Podcasts, ~50 Mio. Episoden, jährlicher Umsatz 1 Mrd. USD
- Suche erstaunlich umständlich und funktional primitiv
- «Winner takes all»-Effekte, insbesondere Angebote im Long Tail schwer auffindbar

Ziel

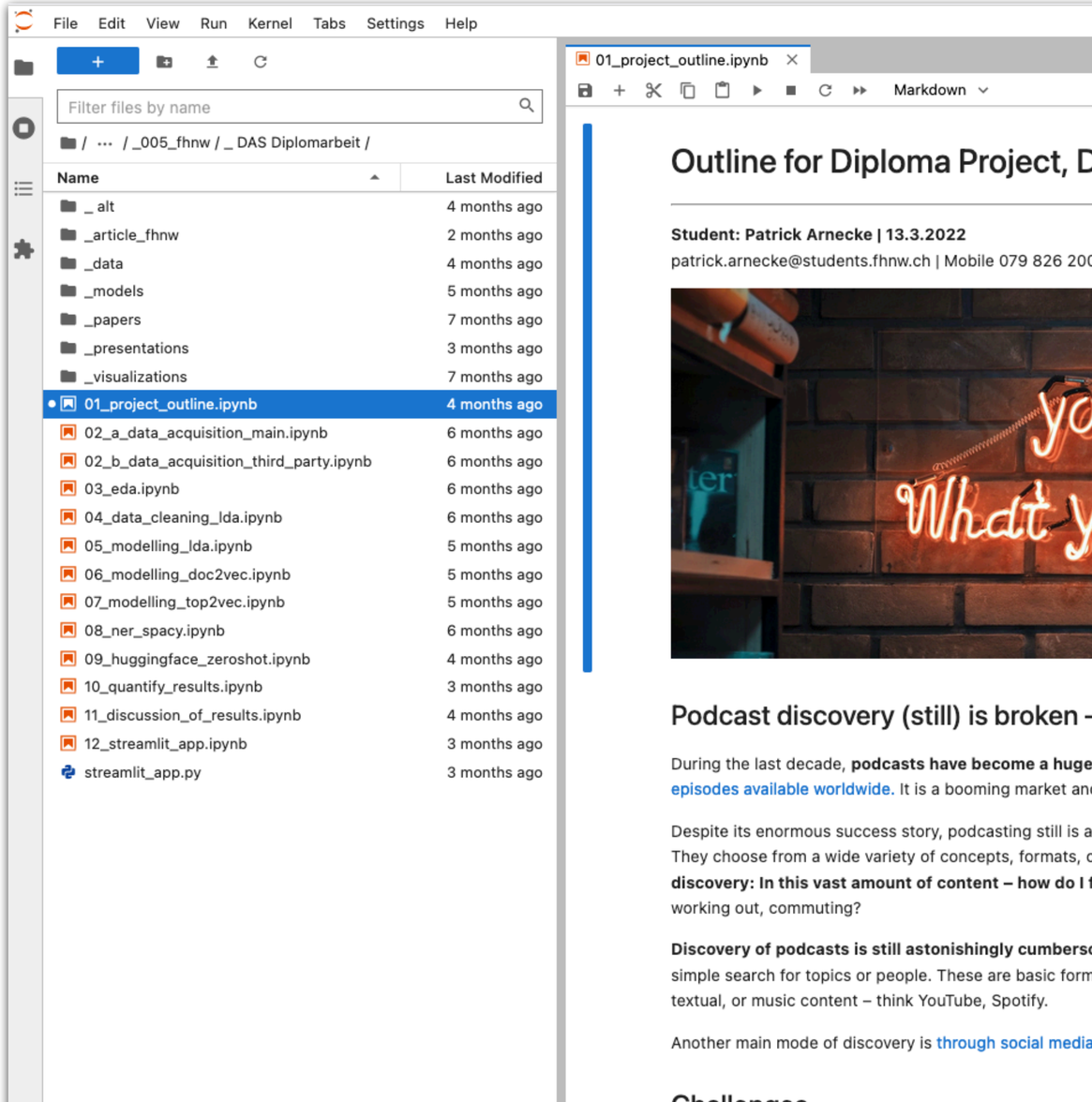
- **Die Suche und das Entdecken von Podcasts verbessern. Dafür neue Wege mit NLP und Machine Learning finden, um Metadaten aufzuwerten oder neu zu generieren.**

Intro

Struktur der Arbeit

11 Notebooks + Streamlit-Applikation

- Konzept
- Daten-Akquisition
- Explorative Datenanalyse
- Datenaufbereitung
- Modelling **LDA**
- Modelling **Doc2Vec**
- Modelling **Top2Vec**
- Exkurs: **Named Entity Recognition** mit Spacy
- **Zeroshot Learning** mit Hugging Face
- **Quantitativer Vergleich** der Ansätze
- Abschliessende Diskussion



Intro

Datenbasis

Das Projekt beruht auf allen verfügbaren Metadaten der Podcasts: Titel, Texte, Zusammenfassungen, Genres, Tags, Episodenbeschreibungen uw.

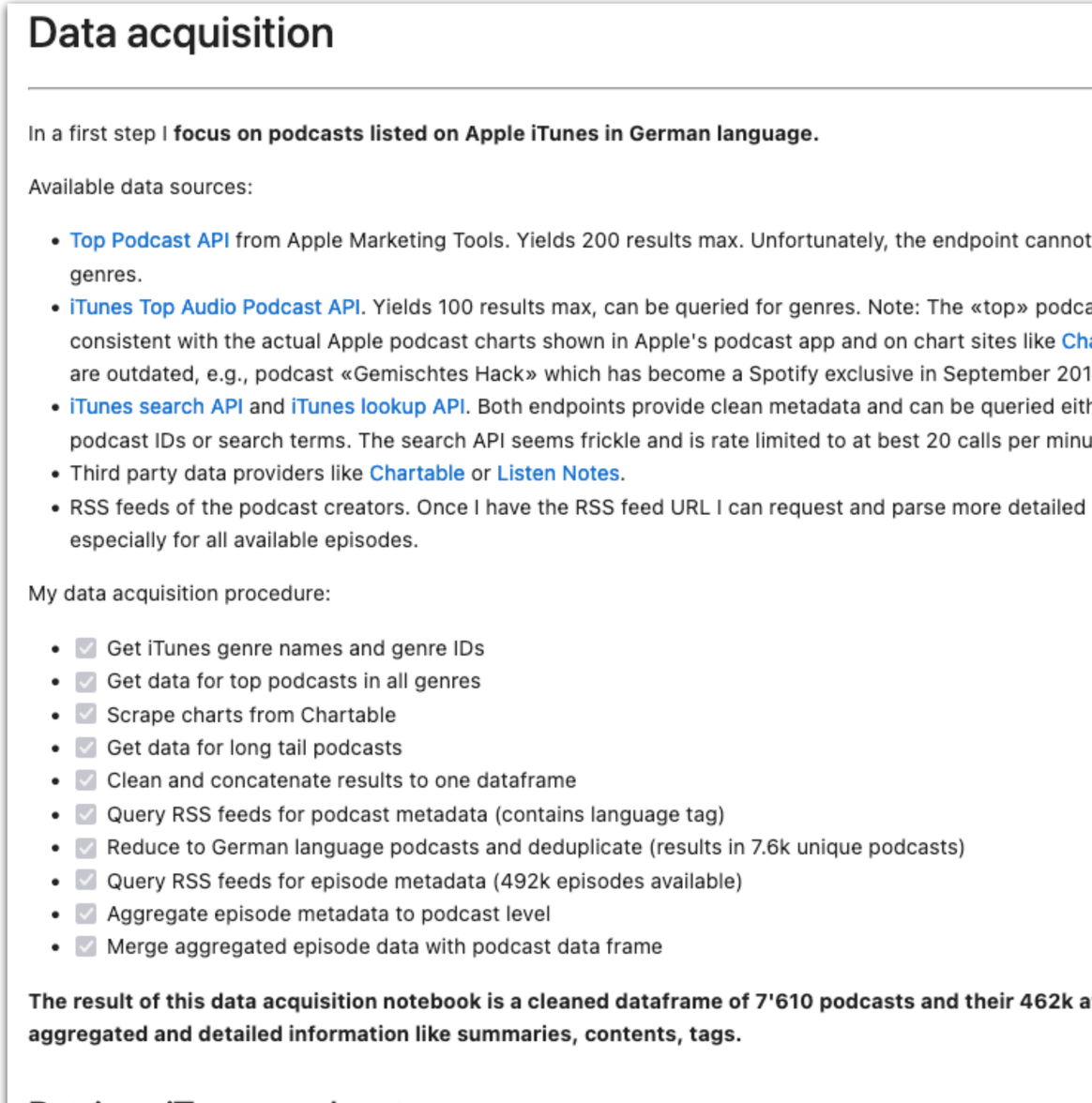
	5727	487	4948
trackId	1558217365	859021496	1540588963
artistName	Lukas Dilsen	Dr. med. Anita Schmidt	EDEKA
ep_authors	[Lukas Dilsen]	None	[EDEKA]
title	Fullback-Dive - Der Podcast rund um die Patriots	Skills Lab PERLE (SD 640)	YUMMI – Der Kinderpodcast
ep_titles	[Episode 10: Delfine in Sicht mit Luca, Episod...	[Akutes Abdomen - Anamnese und körperliche Unt...	[Chaos in der Boxengasse, Das Beet im Kasten, ...
ep_itunes_titles	None	None	[Chaos in der Boxengasse, Das Beet im Kasten, ...
subtitle	Alle News und Insights rund um die New England...	Medizinisches Trainingszentrum, P...	None
ep_subtitles	None	None	None
summary	Alle News und Insights rund um die New England...	Medizinisches Trainingszentrum, P...	Hi!\nWir sind Anna und Ben und als YUMMI-Podca...
ep_summaries	[Die Offseason nimmt langsam Fahrt auf. Ein...	None	[Auf Plätze, fertig, ... hier ist Anna!\n\nSomme...
ep_contents	[Die Offseason nimmt langsam Fahrt auf. Ein...	None	[Auf Plätze, fertig, ... hier ist Anna! \n...
tags	[football, sports]	[education, fau, uni-erlangen]	[edeka, education, hörspiel, kids & family, ki...
ep_tags	None	None	[abenteuer, abenteur, abentuer, aktuell, ameri...
primary_genre	Football	Bildung	Kinder und Familie
genres	[Football, Podcasts, Sport]	[Bildung, Podcasts]	[Kinder und Familie, Podcasts, Leisure]
toplist_top_genre	None	None	None
toplist_sub_genre	None	None	None
chartable_genre	None	None	Kids & Family
genreIds	[1547, 26, 1545]	[1304, 26]	[1305, 26, 1502]
toplist_genre_id	None	None	None
generator	Anchor Podcasts	None	Podigee (https://podigee.com)
itunes_type	episodic	serial	episodic
explicit	True	True	True
releaseDate	2021-10-18 16:48:00+00:00	2021-01-27 23:00:00+00:00	2021-11-26 03:00:00+00:00
trackCount	16	16	54
feedUrl	https://anchor.fm/s/4f5d27cc/podcast/rss	http://feeds.video.uni-erlangen.de/course/itun...	https://yummi.podigee.io/feed/mp3
artworkUrl600	https://is3-ssl.mzstatic.com/image/thumb/Podca...	https://is3-ssl.mzstatic.com/image/thumb/Podca...	https://is5-ssl.mzstatic.com/image/thumb/Podca...
from_chartable	0.0	0.0	1.0
not_longtail	0	0	1
text	Lukas Dilsen Fullback Dive Der Podcast rund di...	Dr. med. Anita Schmidt Skills Lab PERLE Medizi...	EDEKA YUMMI Der Kinderpodcast Hi! Wir sind Ann...

'Fotografie Neu denken. Der Podcast. Von Andy Scholz. Kultur und Gesellschaft.\n\nDer Audio-Podcast über die Bedeutung von fotografischen Bildern in unserem Alltag. Andy Scholz spricht mit Menschen darüber, warum und was sie fotografieren und wieviel. Was die Fotografie und das fotografische Bild mit der Kunst, der Gesellschaft und unseren Kindern macht. Und versucht so dieser gegenwärtigen Omnipräsenz nachzuspüren.\n\nRegelmäßig, episodisch und direkt aus seinem Studio veröffentlicht Andy Scholz in seinem Podcast »Fotografie Neu Denken« Interviews mit Foto- und Kunstschaffenden, Beiträge über Künstler und Künstlerinnen, sowie Gespräche mit Expertinnen und Experten aus der angewandten und künstlerischen Fotografie, der Fototheorie, der Fotogeschichte. Er lässt Menschen zu Wort kommen, die fotografische Bilder erarbeiten, vermitteln und über sie nachdenken. Menschen aus den Bereichen der Philosophie, Soziologie, Medien-, Kultur-, Geistes- und Sozialwissenschaften. \n\nGenauso verschafft er aber auch Menschen Gehör, die einfach nur fotografieren, die das Fotografieren lieben und sich dafür begeistern.\n\nAndy Scholz ist Jahrgang 1971, geboren in Varel am Jadebusen. Er studierte Philosophie und Medienwissenschaften in Düsseldorf, Kunst und Fotografie in Essen an der Folkwang Universität der Künste, ehemals Gesamthochschule Duisburg-Essen u.a. bei Jörg Sasse und Bernhard Prinz. \n\nAndy Scholz ist Künstler, Autor und künstlerischer Leiter und Kurator vom »Festival Fotografischer Bilder«, das er gemeinsam mit Martin Rosner 2016 in Regensburg gründete. Seit 2012 hatte und hat er verschiedene Lehraufträge u.a. Universität Regensburg, Fachhochschule Würzburg, Philipps-Universität Marburg, Ruhr Universität Bochum. Seine Wahlheimat ist das Ruhrgebiet. Er lebt und arbeitet in Essen.\n\nDas Festival Fotografischer Bilder ist eine Kooperation mit dem Kulturamt der Stadt Regensburg und findet alle drei Jahre statt. Im Oktober 2023 voraussichtlich zum dritten Mal.\n\n»Das fotografische Bild ist Werkzeug und Instrument. Material und Rohstoff. Inspiration und Wissenschaft. Individuum und Gesellschaft.« (Andy Scholz, Juni 2020)\n\nhttps://festival-fotografischer-bilder.de\n\nWas sehen wir eigentlich? Wie relevant ist das, was wir sehen? Wie spiegelt sich das in der Gesellschaft und dann in künstlerischen Arbeiten wider? Wie gehen Kunstschaffende damit um? Was sehen sie kritisch? Was denken Philosophen und Soziologen darüber? Wie ist der Diskurs in den Medien- und Kulturwissenschaften? Was machen fotografische Bilder mit uns?\n\nAlles Fotografische wird ausprobiert, auf die Probe gestellt, erforscht und genutzt. Jede fototechnische Möglichkeit wird zum Werkzeug, Instrument, Material und Rohstoff. Fotografische Ergebnisse sind Inspiration und Wissenschaft. Individuell genauso wie gesellschaftlich.\n\n»Es geht ebenso um den Umgang mit, wie um die Benutzung und die Umsetzung von fotografischen Bildern. Der künstlerische Schaffensprozess, die kreative Verwendung von Bildern durch Fotografie. Das Arbeiten am Bild. (Andy Scholz, Juni 2020)\n\nReicht es, etwas zu erklären und zu beschreiben, oder funktioniert ein fotografisches Bild schneller und möglicherweise besser?\n\nDabei spielt dann das Vermitteln eine interessante Rolle. Wie ist eigentlich die didaktische, pädagogische Sichtweise auf das fotografische Bild. Wie steht es um das vermittelte und vermittelnde Bild. Um das kompetente Bild. Um das verantwortungsvolle Bild?\n\nVerstehen wir wirklich, was wir auf einem fotografischen Bild sehen? Müssen wir das nicht auch lernen wie unsere Kinder? Oder sind wir schon so lange davon umgeben, sodass wir es spielerisch mitbekommen haben?\n\nEs geht uns um das Nachdenken über fotografische Bilder genauso, wie um das Wahrnehmen, das Erarbeiten und das Vermitteln von fotografischen Bildern.\n\nHören und schauen Sie rein.\n\nHerzliche Grüße, Ihr Andy Scholz'

Ergebnisse

Daten-Akquisition, Analyse und Bereinigung

- Daten schlecht zugänglich, keine leistungsfähigen APIs
- Metadaten über div. Quellen erfasst von 7.6k Podcasts und 462k Episoden
- 60% Charts, 40% Long Tail
- Metadatenstruktur sowohl bei APIs als auch RSS-Feeds unzureichend normiert
- Textdaten unsauber: HTML, Link, Kontrollzeichen, Werbe- und Sponsorinhalte
- Angebote sehr divers: Textmengen, Episodenzahl (1-2k)
- Bereinigen und Aufbereiten: Stopworte, Fremdsprachen, HTML & Co. entfernen, Lemmatisieren, Tokens und N-Grams



Data acquisition

In a first step I **focus on podcasts listed on Apple iTunes in German language.**

Available data sources:

- Top Podcast API from Apple Marketing Tools. Yields 200 results max. Unfortunately, the endpoint cannot genres.
- iTunes Top Audio Podcast API. Yields 100 results max, can be queried for genres. Note: The «top» podca consistent with the actual Apple podcast charts shown in Apple's podcast app and on chart sites like Cha are outdated, e.g., podcast «Gemischtes Hack» which has become a Spotify exclusive in September 201
- iTunes search API and iTunes lookup API. Both endpoints provide clean metadata and can be queried eith podcast IDs or search terms. The search API seems frickle and is rate limited to at best 20 calls per minu
- Third party data providers like Chartable or Listen Notes.
- RSS feeds of the podcast creators. Once I have the RSS feed URL I can request and parse more detailed especially for all available episodes.

My data acquisition procedure:

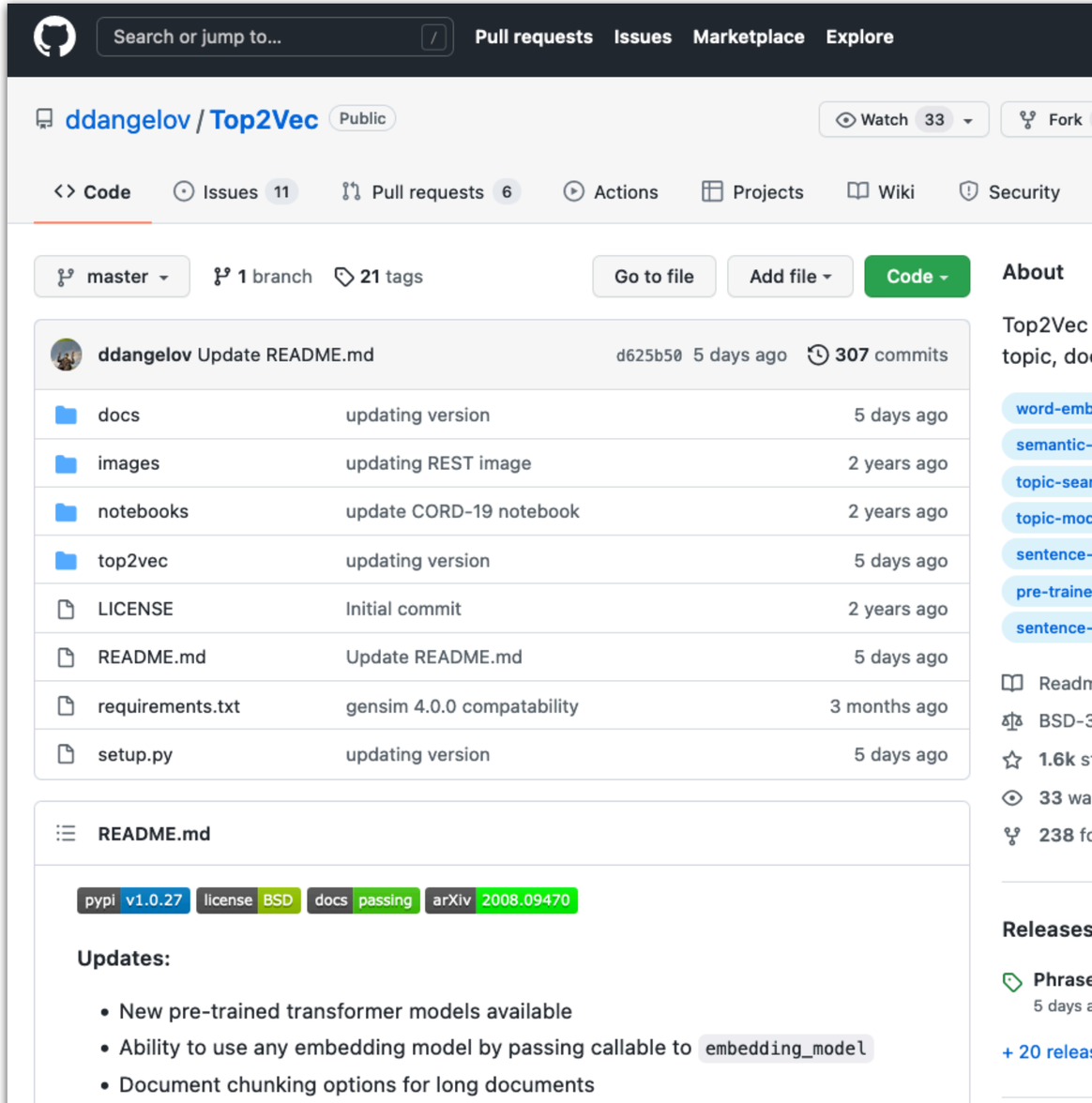
- ☑ Get iTunes genre names and genre IDs
- ☑ Get data for top podcasts in all genres
- ☑ Scrape charts from Chartable
- ☑ Get data for long tail podcasts
- ☑ Clean and concatenate results to one dataframe
- ☑ Query RSS feeds for podcast metadata (contains language tag)
- ☑ Reduce to German language podcasts and deduplicate (results in 7.6k unique podcasts)
- ☑ Query RSS feeds for episode metadata (492k episodes available)
- ☑ Aggregate episode metadata to podcast level
- ☑ Merge aggregated episode data with podcast data frame

The result of this data acquisition notebook is a cleaned dataframe of 7'610 podcasts and their 462k a aggregated and detailed information like summaries, contents, tags.

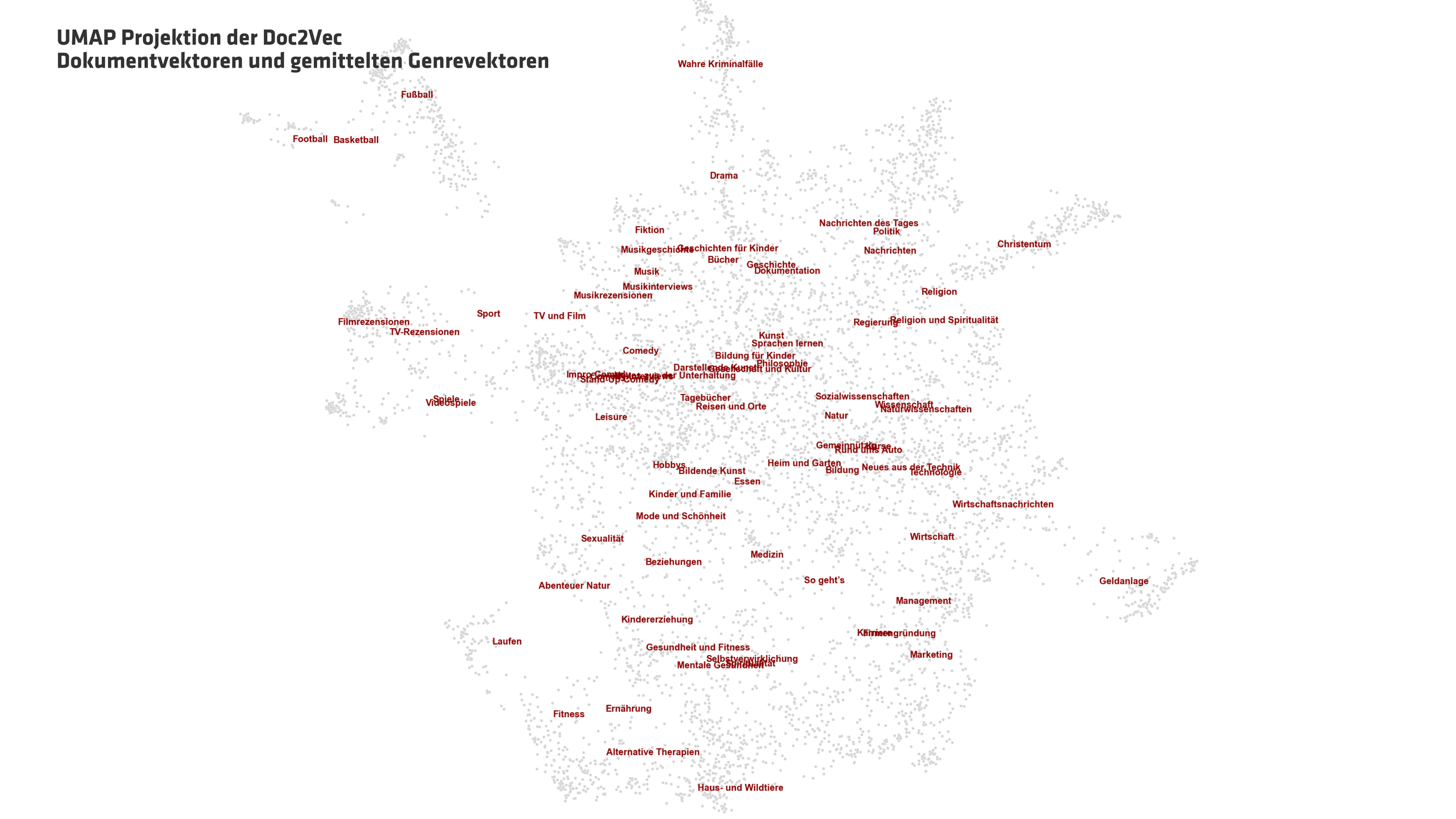
Ergebnisse

Modelling LDA, Doc2Vec, Top2Vec

- Schnelles Training, qualitativ robuste Ergebnisse
- Embeddings erlauben Berechnung der Distanzen von Podcasts und Genres (ähnlich/unähnlich)
- **Top2Vec** am Interessantesten, da Worte, Dokumente und Topics im selben Vektorraum eingebettet werden. Bietet zudem Such- und Empfehlungsfunktionen. Paket zugleich schwer zu installieren und handhaben.



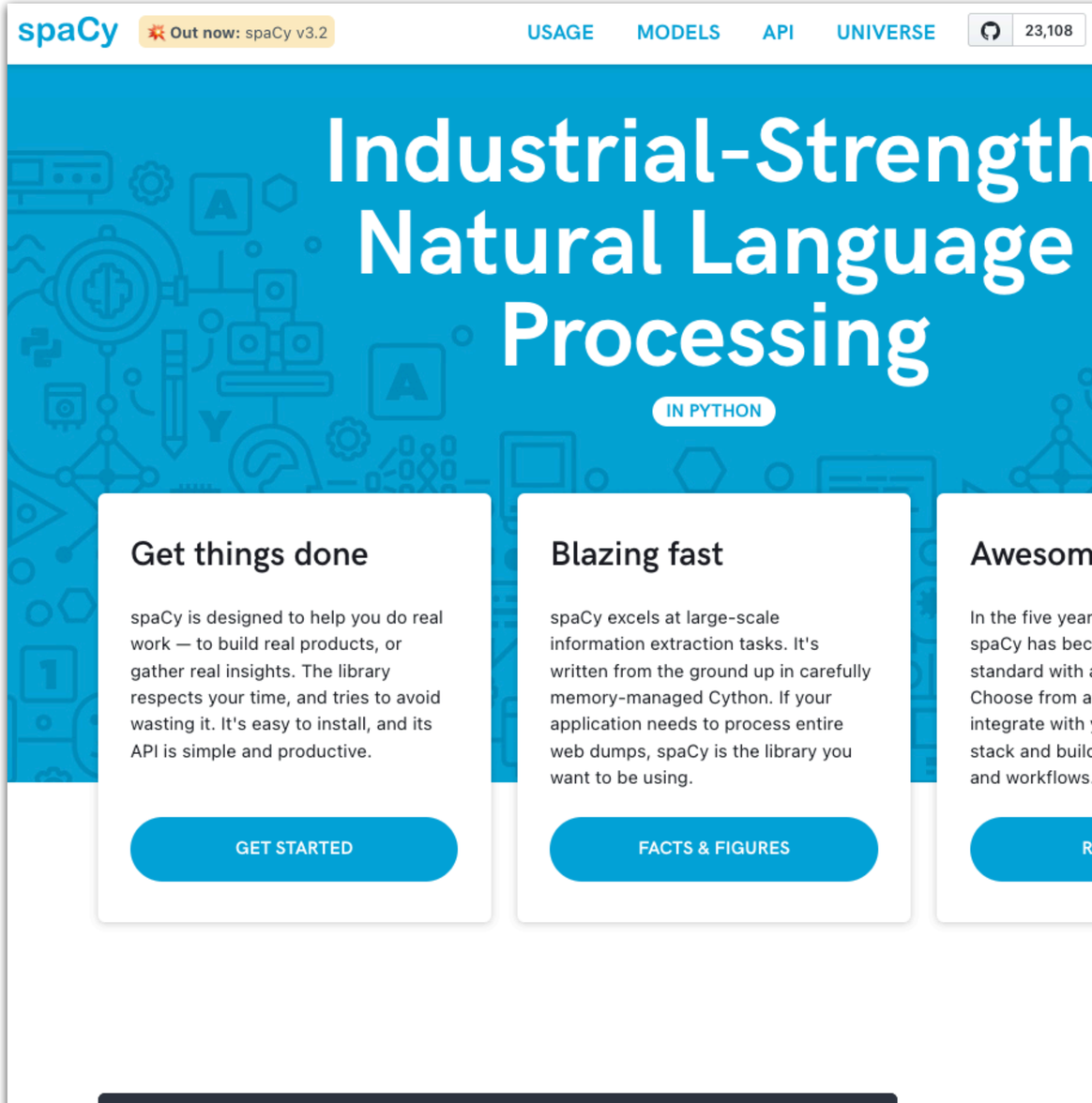
UMAP Projektion der Doc2Vec Dokumentvektoren und gemittelten Genrevektoren



Ergebnisse

Named Entity Recognition mit Spacy

- Interessant, da durchdachte Pipeline, vortrainierte Modelle, schnelle Verarbeitung und gratis
- Vielversprechende Ergebnisse, zugleich viele Fehler und Rauschen

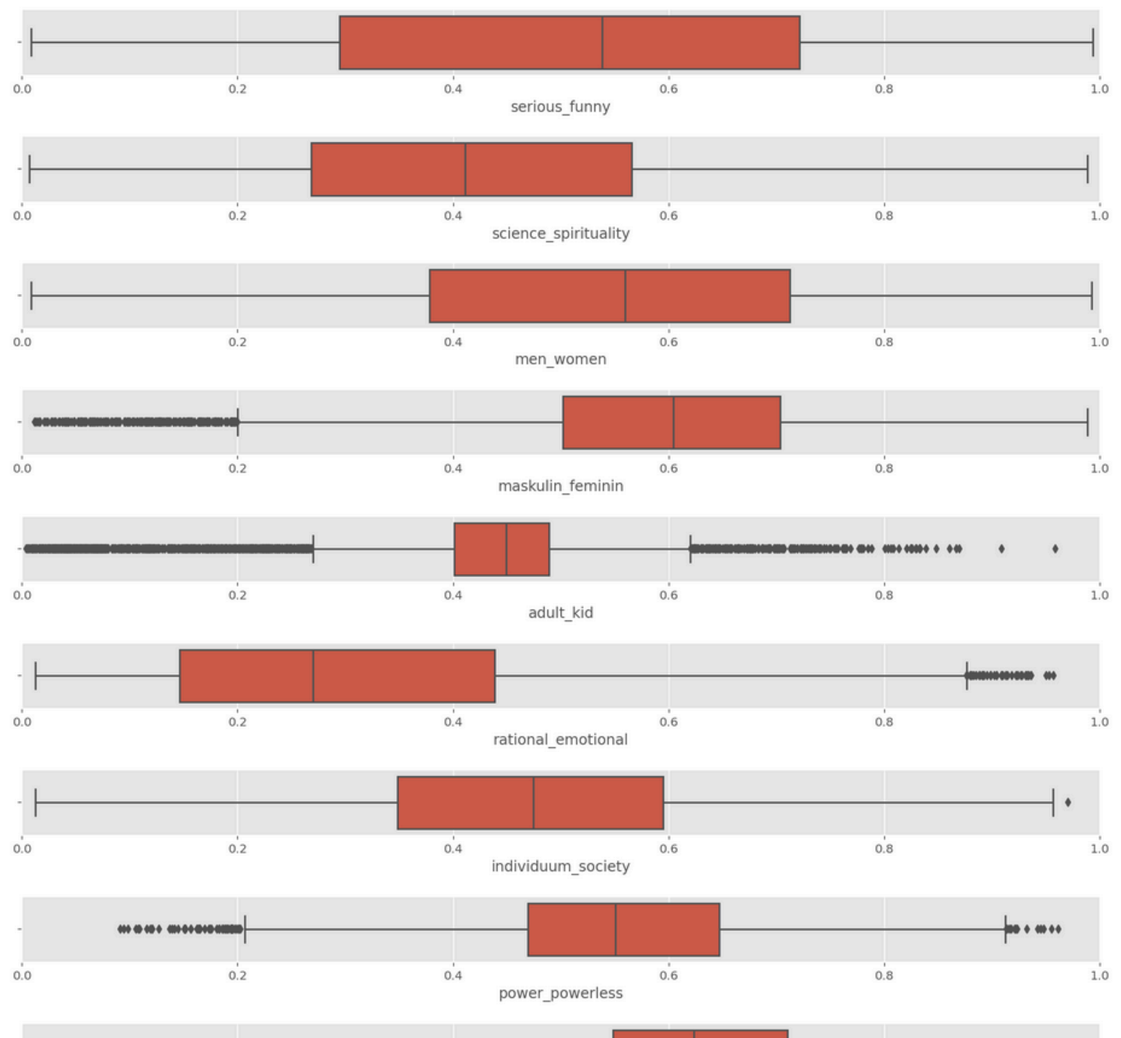


Ergebnisse

Zeroshot Learning mit Hugging Face Transformer Models

- Klassifikation mit willkürlichen Labels.
- Metadaten kreativ generieren: z.B. Gegensatzpaare jung/alt, männlich/weiblich (Achtung: Bias), lustig/ernst, rational/emotional, einfach/komplex, Geschichten/Fakten
- Inferenz zeitintensiv (2.5h pro Labelpaar)
- Ergebnisse sehr interessant und nachvollziehbar

```
for col in zero_shot_cols:
    fig, ax = plt.subplots(figsize=(16,1))
    sns.boxplot(df[col], ax=ax)
    # sns.stripplot(df[col], ax=ax, size=1, color="green")
    plt.xlim(0,1)
    plt.show()
```



Ergebnisse

Quantitativer Vergleich der Ansätze

- «Primary Genre» aus Metadaaten als Grundwahrheit
- Klassifikation mit generierten Vektoren oder direkte Inferenz mit Zeroshot. «F1 Score weighted» als Metrik.
- 7 Algorithmen + Dummy Baseline. Hyperparameter-Tuning mit scikit-Pipeline und RandomSearchCV.
- Alle 4 Modelle (LDA, Doc2Vec, Top2Vec, Zeroshot) liefern gute Ergebnisse.
- Top2Vec mit SVC oder Logistic Regression ergibt die besten Ergebnisse.
- Zeroshot schneidet am wenigsten gut ab. Zugleich in den Top10 und Ergebnisse teils richtiger als Grundwahrheit.

- **Top2Vec and Doc2Vec vectors yield the best results** as training data and clearly outperfom the LDA vecto
- **SVC yields the best F1 score of 0.653** with the Top2Vec vectors.
- **Logistic regression follows closely with 0.643**.

```
baseline_results = pd.DataFrame(baseline_results)
baseline_results.columns = ["vector_set", "clf", "score"]
baseline_results.clf = baseline_results.clf.apply(lambda x: str(x).split("(")[0])
display(baseline_results.sort_values("score", ascending=False).reset_index(drop=True))
```

	vector_set	clf	score
0	top2vec	SVC	0.6531
1	top2vec	LogisticRegression	0.6430
2	doc2vec	SVC	0.6357
3	top2vec	LinearDiscriminantAnalysis	0.6274
4	top2vec	RidgeClassifier	0.6131
5	doc2vec	LinearDiscriminantAnalysis	0.6102
6	doc2vec	LGBMClassifier	0.6071
7	top2vec	KNeighborsClassifier	0.5684
8	doc2vec	RidgeClassifier	0.5670
9	doc2vec	RandomForestClassifier	0.5649
10	doc2vec	LogisticRegression	0.5531
11	top2vec	LGBMClassifier	0.5510
12	lda	RandomForestClassifier	0.5364
13	lda	LogisticRegression	0.5358
14	lda	LGBMClassifier	0.5332
15	lda	LinearDiscriminantAnalysis	0.5293
16	lda	SVC	0.5161
17	lda	RidgeClassifier	0.5107
18	doc2vec	KNeighborsClassifier	0.4906
19	lda	KNeighborsClassifier	0.4640
20	top2vec	RandomForestClassifier	0.4493

Ergebnisse

Fazit

- Daten zwar erstaunlich schwer akquirierbar und heterogen, aber ausreichend signifikant.
- **Alle Ansätze werten Metadaten deutlich auf und bieten verbesserte oder neue Zugänge zu Podcastinhalten.**
- **Named Entity Recognition** mit guten Ergebnissen, aber nur mit intensiver genrespezifischer Bereinigung nutzertauglich.
- **Zeroshot** am spannendsten, da völlig neuartige und aus Nutzersicht überraschende Labels klassifiziert werden können.

Ergebnisse

Streamlit-Applikation