Open Transcript

Cultural Broadcasting Archive

NLP Meetup Vienna, Oct 22nd 2019

Alexander Baratsits http://www.baratsits.at

Jurist/Lawyer, Radio FRO

Lukas Strasser

Linguist, NLP aficionado

data4good

Open transcript

https://cba.fro.at, since 2000

Platform of Community Radios in Austria

… biggest podcastprovider in Austria, archive, re-broadcasting

- 101.500 audiofiles
- 8.700 pictures, 6.83 TB
- 40 languages
- Playout on 14 websites

- Open Transcript
  - Built up pipeline for automatic transcription of podcasts
  - Generate data for classification / recommendation system

=> Make podcasts searchable

=> Attract further traffic on CBA as hosting platform

- Ready2order hackathon:

  https://cpa-lab.github.io/

- Three paths:
  - API to Google/Wit-ai/etc.
  - adaptation of existing models
  - training from scratch

- Two frameworks:
  - Kaldi
  - CMUSphinx

- One problem: data requirements

  - audio and text data

  - dictionary of words that maps to phones

  - a language or acoustic model (set of phones)

- An alternative: Keyword Recognition?

Contact: cba@fro.at