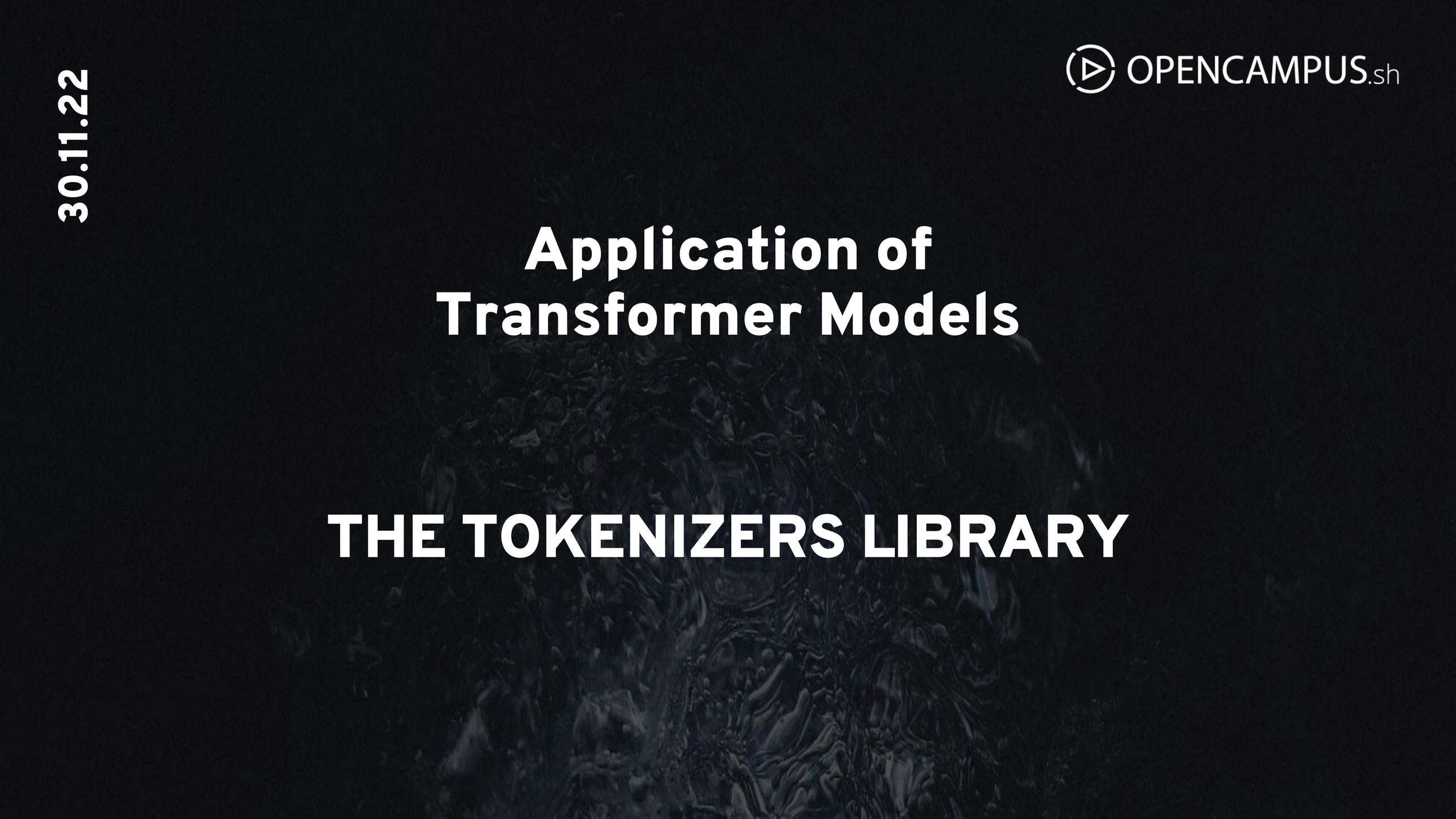
30.11.22

OPENCAMPUS.sh

Application of Transformer Models

THE TOKENIZERS LIBRARY



- Quiz
- Open Discussion
- Dataset Characteristics
- Baseline Models

QUIZ

https://forms.office.com/e/SyHkdRWsHq

OPEN DISCUSSION

- Can you think of a model task in which it is important to have a reversible tokenization – making, for example, sure that all spaces are considered?

- Can you think of possible applications of BPE, WordPiece, or unigram tokenization outside of NLP?

DATASET CHARACTERISTICS

- Is your collection of samples possibly biased?
- How must the data be collected to be used with your model?
- Can you think of different groups/ types of input sequences?
  - Create corresponding filter variables for the later evaluation of your results.
- Does the dataset include outliers?
- For classification problems:
  - Is your sample balanced across all classes?
  - If not, how will you deal with it?
- For generation problems:
  - Define a set of prompts and investigate the completions according to their bias.

FEEDBACK GROUPS

Group 1:

- Jonathan/Julian: Arguments Mining / NER Task on data already collected
- Max: Q&A model
- Khan: Classification of activity descriptions according to keywords

Group 2:

- Saif/ Emmanuel/ Kristian/ Atul: Time Series Prediction Financial/Climate
- Manpreet: Unsupervised training of log data to predict user behavior
- Laura/ Janosch/ Valentin : Training a model to produce text written in different authors' style

Group 3:

- Benjamin/ Malte/ T.-Niklas: Speech to speech models including translation
- Jeremy/ Veit/ Christian: Transcribing and summarizing Podcasts

SHORTCOMINGS OF LANGUAGE MODELING

- **Human Reporting bias (Gordon and Van Durme, 2013):**
  - **Not stating the obvious**
  - **Common sense isn‘t written down**
- **Facts about named entities**
- **No grounding to other modalities**

Possible Solutions:

- **Incorporate structured knowledge (e.g. databases; Zhang et. A. 2019)**
- **Multimodal learning (e.g. visual representations; Sun et al. 2019)**

PROJECT MILESTONES

16.11.	Form project groups
23.11.	Literature review
30.11.	Dataset characteristics
04.01.	Baseline model
11.01.	Project presentations

CHARACTERISTICS OF BASELINE MODELS

- **Should be simple to setup, with a reasonable chance of providing decent results, and very unlikely to overfit.**

- **Should be interpretable, which can help your understanding of the data and guide your feature engineering.**

Ameisen, E. (2018, March 6). *Always start with a stupid model, no exceptions.* Medium. https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa

POSSIBLE TYPES OF BASELINE MODELS

For continuous variables:

- **Linear Regression**
- **Gradient Boosted Trees**

Classification of Structured Data or Natural Language:

- **Logistic Regression**
- **Support Vector Machines**
- **Gradient Boosted Trees**

Classification of Images

- **Simple Convolutional Architectures**
- **Fine tuning VGG or re-training some variant of a U-net is usually a great start for most image classification, detection, or segmentation problems.**

POSSIBLE TYPES OF BASELINE MODELS

Text Summarization:

- **the *lead-3* baseline: take the first three sentences of the input sequence**

IF THERE IS NO EFFECTIVE BASELINE

- Instead of simplifying the model, simplify the data.
- Try to get your complex model to overfit to a very small subset of your data.

BASELINE MODEL RESULTS

Help you understand your data:

- **Which classes are harder to separate?**
- **What type of signal picks your model up on?**
  How is your model making decisions?
- **What signal is your model missing?**
  Is it possible to engineer an additional feature?

TODOS UNTIL NEXT WEEK

- Complete at least two sections from chapter 7 (Main NLP Tasks) of the Hugging Face course

- Calibrate a First Baseline Model