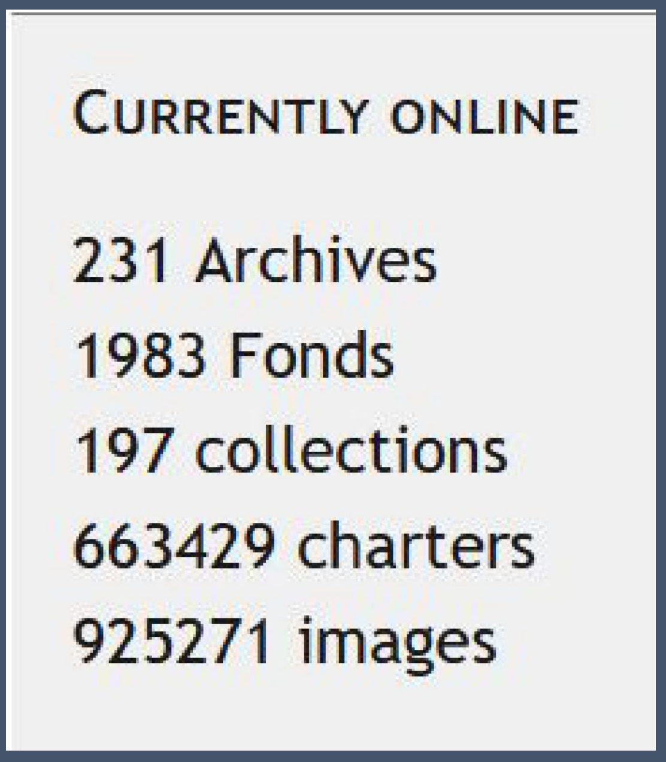
Few Shot Classification for Labeling of Medieval and Early Modern Charter Texts

Tamás Kovács, Sandy Aoun, Anguelos Nicolaou, Daniel Luger, Florian Atzenhofer-Baumgartner, Florian Lamminger, Franziska Decker, Georg Vogeler
{forename.surname}@uni-graz.at

DiDip
From Digital to Distant Diplomatics

Regesta (abstracts) of medieval charters give easy access to core historical information about the social, economic, and political life of the past.

Monasterium.net contains over 600,000 charters from all over Europe (majorly Germany, Italy, Austria, Slovakia, Czech Republic, Hungary), a mass that is impossible to study with traditional, manual methods!



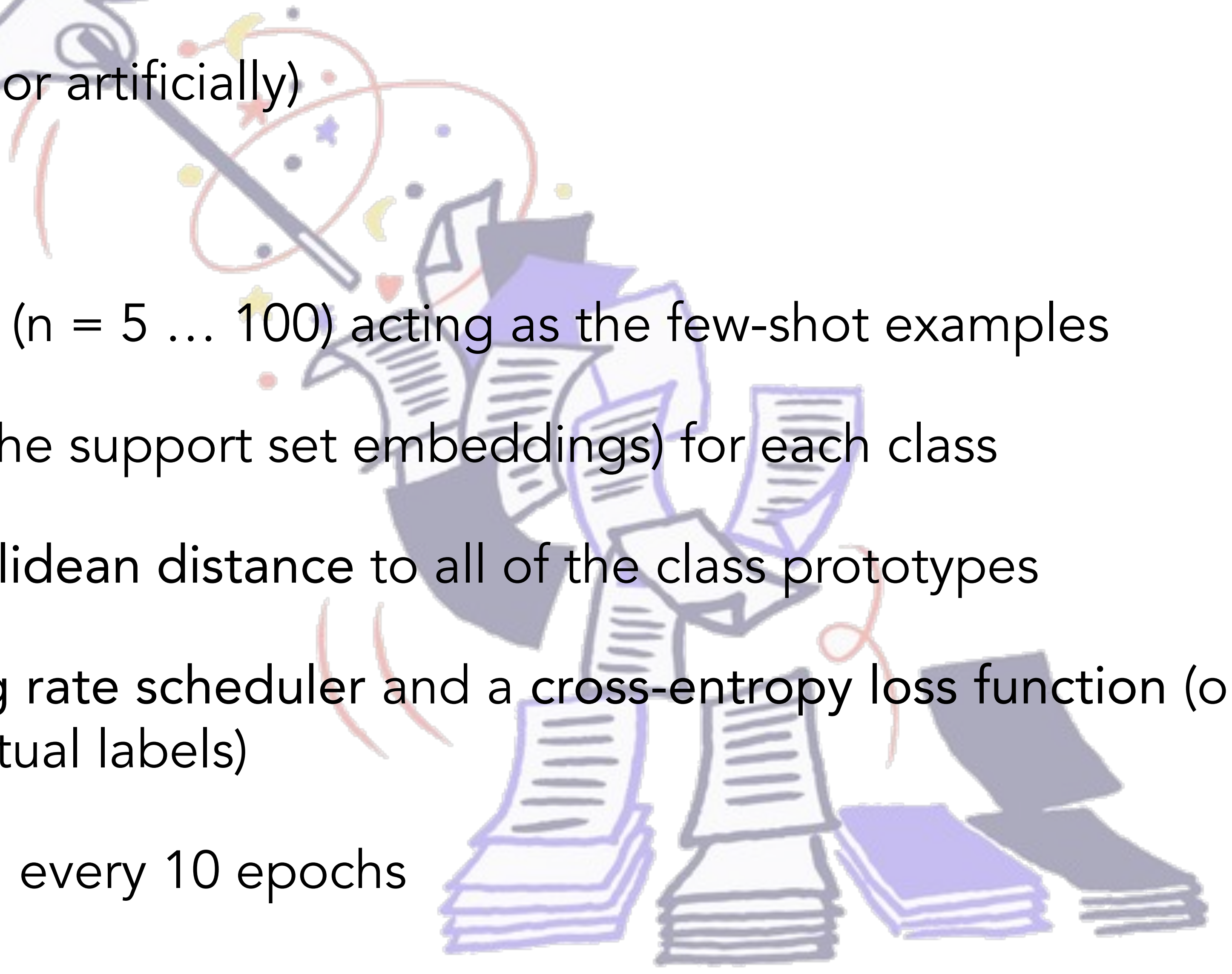
Currently online

231 Archives
1983 Fonds
197 collections
663429 charters
925271 images

Screenshot from Monasterium.net as of May 28th 2023

We propose a few-shot learning solution based on a **prototypical network**, in **Pytorch**!

1. Generating labels based on charter regesta (either manually or artificially)
2. Creating document-embedding (using SBERT or Doc2Vec)
3. Creating ‘support set’ – a small number of labeled examples (n = 5 … 100) acting as the few-shot examples
4. Prototypical network calculates prototypes (= the means of the support set embeddings) for each class
5. Classification of new instances in a metric space by their **Euclidean distance** to all of the class prototypes
6. Training of the network using the **Adam** optimizer, a **learning rate scheduler** and a **cross-entropy loss function** (on the discrepancy between the model’s predictions and the actual labels)
7. Evaluation of the model’s accuracy via an evaluation function every 10 epochs



Advantages:

- ✔ only a small annotated corpus is needed for training reducing time and resource requirements
- ✔ can be used to automatically annotate larger datasets
- ✔ those new annotations can themselves be used to train or fine-tune other task-specific models/approaches
- ✔ potentially valuable tool also for large datasets of other historical documents

CENTRE FOR
INFORMATION-MODELLING
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES

www.didip.eu
@DiDip_ERC

We work for tomorrow