University of Fribourg | *Lecturer: Dr. Andreas Fischer*
Department of Informatics | *Teaching Assistant: Michael Jungo*

Spring Semester | Pattern Recognition

Exercise 3 - 15th of April 2024

Second Group Task – Keyword Spotting

Deadline: 6th of May 2024 (end of day)

In this exercise you will be developing a machine learning approach for spotting keywords in historical documents. The goal is to find the most similar words based on a query image. It is based on the similarity of the images, not the actual text that is written inside, which is primarily useful when the transcription is not known. You will be using the transcriptions to find instances of the keywords in the document images and develop/train a machine learning method that will be able to predict how similar two word images are.

Dataset: George Washington Database

You can find the dataset on ILIAS in `exercises/3-KWS`.

The dataset is an excerpt from the George Washington Database[1] with the necessary information for the keyword spotting task.

In the dataset, you will find images of full page documents and the corresponding locations of every word on a given page, alongside its transcription (text on a character level). Given a list of keywords, you can find instances of that word in the documents and use them for the training.

The file structure is as follows:

- `images/`: Contains the full page document images.
- `locations/`: SVGs with the surrounding polygons for each word.
- `keywords.tsv`: List of keywords, which occur *at least once* in the training/validation set.
- `transcription.tsv`: Transcriptions of all words (on a character level) of the whole dataset (see bleow for details).
- `train.tsv`: List of documents that belong to the training set.
- `validation.tsv`: List of documents that belong to the validation set.

[1] https://fki.tic.heia-fr.ch/databases/washington-database

Data Format

Each document is referenced with the following ID:

- `DDD-LL-WW`
- `DDD` = Document number
- `LL` = Line number
- `WW` = Word number

For example: `270-05-07` is the 7^{th} word on line `5` in the document `270`.

The transcriptions are given character-wise, where each letter is separated by a dash (–). Special characters are encoded by starting with `s_`:

- `s_X`: Number (replace X with any digit)
- `s_pt, s_cm, ...`: Punctuation
- `s_s`: Strong
- `s_mi`: Hyphen (–)
- `s_sq`: Semicolon (`;`)
- `s_qo`: Colon (`:`)
- `s_qt`: Apostrophe / single quote (`'`)

Expected Submission

- Access to your Git(Hub) repository so that we can inspect your code.
- Short report in your Git repository (Markdown or PDF) with your results.