HTR | API

Training

For training a new HTR model using the new API (for RNN HTR), at first a configuration XML has to be created.
Besides parameters (the example below includes the default values) mandatory fields are:

- a model name
- a description
- the language
- the collection ID where the input documents can be found and where the resulting model will be linked

The input for training is described in the TrainList section of the XML and is made up of train elements where each includes:

- the document ID
- a list of pages where each page includes
  - the page-ID
  - the ID of the transcript version that should be used for training

Optionally a test set can be specified in the TestList element analogously.

The training descriptor then should look like this:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<citLabHtrTrainConfig>
    <modelName>Test Model</modelName>
    <description>A description</description>
    <language>German</language>
    <colId>2</colId>
    <numEpochs>200</numEpochs>
    <learningRate>2e-3</learningRate>
    <noise>both</noise>
    <trainSizePerEpoch>1000</trainSizePerEpoch>
    <trainList>
        <train>
            <docId>1</docId>
            <pageList>
                <pages>
                    <pageId>1</pageId>
                    <tsId>1</tsId>
                </pages>
                <pages>
                    <pageId>2</pageId>
                    <tsId>2</tsId>
                </pages>
            </pageList>
        </train>
        <train>
            <docId>2</docId>
            <pageList>
                <pages>
                    <pageId>3</pageId>
                    <tsId>3</tsId>
```

```
                </pages>
                <pages>
                    <pageId>4</pageId>
                    <tsId>4</tsId>
                </pages>
            </pageList>
        </train>
    </trainList>
    <testList/>
</citLabHtrTrainConfig>
```

That XML is then send via POST to

`https://transkribus.eu/TrpServer/rest/recognition/htrTrainingCITlab`

and the call returns the job-ID of the training.

Note, that the models are now linked to the collection they were started in (cf. colId element in training descriptor XML).

Manage HTR models

Listing available models

Listing models can be done with a GET request to:

`https://transkribus.eu/TrpServer/rest/recognition/{collection-ID}/list?prov=`

`{techProvider}`

The call includes:

- Path parameter: collection-ID
- Query parameter: the tech provider. Here at the moment only “CITlab” is allowed as value.

A model is described in the result e.g. by the following XML:

```
...
<trpHtr>
        <htrId>22</htrId>
        <name>Test Model</name>
        <description>A description</description>
        <provider>CITlab</provider>
        <created>
            <nanos>338000000</nanos>
        </created>
        <gtDocId>1614</gtDocId>
        <testGtDocId>1615</testGtDocId>
        <language>German</language>
        <trainJobId>3160</trainJobId>
        <cerString>1,000000</cerString>
        <charList> =1
,=2
.=3
...
</charList>
</trpHtr>
...
```

Worth noting is the htrId element which is used for applying the model.

The cerString includes CER values that have been determined during training against the train set or the (optional) test set.

The charList includes the character channel mapping and informs about the characters that are known to the model.

Adding models to other collections

In order to make a trained model available in another collection you POST to:

```
https://transkribus.eu/TrpServer/rest/recognition/{collection-ID}/{htr-ID}/add?collId={destination_collection-ID}
```

Removing models from collections

DELETE request to:

```
https://transkribus.eu/TrpServer/rest/recognition/{collection-ID}/{htr-ID}/remove
```

Note, that a model can not be removed from all collections.

Dictionaries

For applying an HTR model you need to provide a dictionary filename. Available dictionaries can be listed via GET to:

```
https://transkribus.eu/TrpServer/rest/recognition/dicts
```

Recognition

For applying an HTR model, a POST request has to be sent to:

```
https://transkribus.eu/TrpServer/rest/recognition/{collection-ID}/{htr-ID}/htrCITlab?id ={doc-ID}&pages={page-string}&dict={dictionary_filename}
```

The dictionary parameter is optional and can be omitted.

Note that the `id` and `pages` parameters may be replaced with an object in the request body, allowing for more detailed selection of the input data:

```
{
   "docId" : 1543,
   "pageList" : {
      "pages" : [ {
         "pageId" : 1234,
         "regionIds" : [ "the_xml_id_of_a_text_region" ]
      }, {
         "pageId" : 12345,
         "tsId" : 1234567
      } ]
   }
}
```

Equivalent XML representation:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<documentSelectionDescriptor>
    <docId>1543</docId>
    <pageList>
        <pages>
            <pageId>1234</pageId>
            <regionIds>the_xml_id_of_a_text_region</regionIds>
        </pages>
        <pages>
            <pageId>12345</pageId>
            <tsId>1234567</tsId>
        </pages>
    </pageList>
</documentSelectionDescriptor>
```