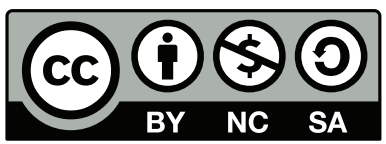
Module 8
Using ABBYY: Practice

Uwe Springmann

Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität München (LMU)



2015-09-15

Practice session: Overview

- register for a free ABBYY developer account:
  - register with name of your app (make up a project name)
  - enter a Cloud OCR SDK promo code
  - promo code: (ask instructor)
  - 1.000 pages valid until 15 of January, 2016
    (thanks to Michael Fuchs of ABBYY Deutschland)
- adapt a script for the Cloud OCR service
- OCR some sample images with various output formats

Adapt the script

- download the data for Module 8 to your laptop
- insert your data into the script `cloud_recognize.sh`:

```
ApplicationId=""
Password=""
```

- open a terminal and run the following command from your data directory:

```
./cloud_recognize.sh
```

- you will see the following output:

```
ABBYY Cloud OCR SDK demo recognition script
Invalid arguments.
Usage:
./cloud_recognize.sh <input> <output> [-f output_format] [-l language] \
[-t typeface]
output_format: txt|rtf|docx|xlsx|pptx|pdfSearchable|pdfTextAndImages|xml
typeface: normal (default), gothic
Some language examples: English (default), Russian, ChinesePRC, \
German, OldGerman etc.
For full list see ocrsdk documentation
```

Recognize some page images

- the downloaded data contain the following images:
  - goethe.tif (Goethe 1809, Wahlverwandtschaften)
  - grenzboten.tif (Grenzboten 1841)
  - latin.tif (Hobbes 1668, Leviathan)
  - greek.tif (Zonaras 1870, Epitome)
- OCR some of the images:
  - default options: output_format=txt, language=english, typeface=normal

```
./cloud_recognize.sh goethe.tif goethe.txt -l oldgerman -t gothic
./cloud_recognize.sh latin.tif latin.txt -l latin
./cloud_recognize.sh greek.tif greek.txt -l greek
```

- prepare a searchable pdf
- compare with ground truth, e.g.

```
ocrevalutf8 accuracy somefile.gt.txt somefile.txt | more
```