

Code and Data

for “Classification of Medieval Documents: Determining the Issuer, Place of Issue, and Decade for Old Swedish Charters” (to appear in *DHN2020 Digital Humanities in the Nordic Countries*, Riga, 17–20 March 2020)

Mats Dahllöf^[0000–0002–4990–7880]

Uppsala University, Uppsala, Sweden
`mats.dahllof@lingfil.uu.se`

Abstract. The study based on this code and dataset is a comparative exploration of different classification tasks for Swedish medieval charters (transcriptions from the “SDHK” collection) and different classifier setups. In particular, we explore the identification of the issuer, place of issue, and decade of production. The experiments used features based on lowercased words and character 3- and 4-grams. We evaluated the performance of two learning algorithms: linear discriminant analysis and decision trees. For evaluation, five-fold cross-validation was performed. We report accuracy and macro-averaged F1 score. The validation made use of six labeled subsets of SDHK combining the three tasks with Old Swedish and Latin. Issuer identification for the Latin dataset (595 charters from 12 issuers) reached the highest scores, above 0.9, for the decision tree classifier using word features. The best corresponding accuracy for Old Swedish was 0.81. Place and decade identification produced lower performance scores for both languages. Which classifier design is the best one seems to depend on peculiarities of the dataset and the classification task. The present study does however support the idea that text classification is useful also for medieval documents characterized by extreme spelling variation.

This bundle of code and data published in DiVA – Academic Archive (<http://www.diva-portal.org>). 3 January 2020.

Code and data bundle

The zip-file contains the data file `sdhk.xml` and a directory named `src` with the Python code. The file `experiments_dhn2020.py` is for running the experiments.

Data

The data, from the Diplomatarium Suecanum Main Catalogue (Svenskt Diplomatariums huvudkartotek, SDHK, <https://sok.riksarkivet.se/SDHK>), as accessed in the experiments is provided by the XML file `sdhk.xml`. The data were

downloaded and compiled into the XML as stated in the file `prepare_data.py`. (Rerunning the downloading and compilation will give you a more recent version of the data.)

The Document Type Definition for the data file `sdhk.xml` is as follows:

```
<!ELEMENT sdhk (charter)*>
<!ELEMENT charter (lang, date, place, issuer, transcr)>
<!ATTLIST charter id CDATA #REQUIRED>
<!ATTLIST charter year CDATA #REQUIRED>
<!ELEMENT lang (#PCDATA)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT place (#PCDATA)>
<!ELEMENT issuer (#PCDATA)>
<!ELEMENT transcr (#PCDATA)>
```

The charter elements contain language, date, place, issuer, and transcription as retrieved from the original SDHK files. The attributes `id` and `year` hold the identifier number and year (extracted from the date element), respectively.

With only one charter in the file as an example `sdhk.xml` would look like this (with line breaks added):

```
<?xml version="1.0" encoding="UTF-8"?>
<sdhk>
<charter year="1375" id="10699">
<lang>svenska</lang>
<date>1375 juni 24</date>
<place>Uppsala.</place>
<issuer>Kettilbjörn i Gunsta</issuer>
<transcr>Allom thøm mannom thætta breeff hœra ok see sændir
Kætilbiørn Gunzsta helso medh warom Herra. Thet skal allom
mannom witirlighit wara migh hawa giort eet iordhaskipte medh
wælbornom manne herra Magnuse Thorkilsson kanuk i Wpsalum medh
swa skæl ath iak hawer giwid honum fyra ørtogholand iordh i
Wesbý a Wlsa i lmunge sokn liggjande. Ok han hawer migh giwidh
fyra ørtogholand iordh i Biristum j sunða sokn liggjande ok
sextan mark pæninga ther til Affhendir iak mik ok minum arwum
the sama fyra ørtoghaland iord j Wlsa medh allo thy ther til
liggir i wato ok thorro, skogh, akir ok ængh ok fisnum nær ok
færre ok tilhænder ia thøm sama fornæmda herra Magnuse ok hans
arwum medh fastum ok fuldum skælum byta, skipa, sælia æpte sinum
eghnum wilia. Jn cuius rei euidentiam firmiorem sigilla
discretorum virorum Philippi Karlsson, dicti Galla peto
presentibus coapponi. Datum Vpsalie anno Domini Mo ccco lxxo
quinto, die beati Johannis baptiste.</transcr>
</charter>
</sdhk>
```

Repeating the experiments

The experiments were conducted in Python 3.7.4. They produce a set of files for results and auxiliary purposes. If you want to repeat the experiments, these paths in the file for running them, `experiments_dhn2020.py`, have to be adapted to your own environment:

```
# Working directory. Should be created beforehand.
expdir = 'c:/corpus/dhnsdhk'

# Path to the xml file with the collected SDHK data.
sdhk_collected_xml = 'c:/corpus/dhnsdhk/sdhk.xml'
```

The file `experiments_dhn2020.py` contains (among other things) these calls for running the experiments on Old Swedish. (The Latin ones are treated in a corresponding way.)

```
run_exps_on_swedish(expdir + '/dhn2020swexp')
make_plots('Swedish', expdir + '/dhn2020swexp/results')
make_stats('Swedish', expdir + '/dhn2020swexp')
```

This creates two subdirectories (`feats` and `results`). It also produces an XML with the charter data selected and filtered for each specific task. Basic statistics are derived from the XML files and stored in CSV files.

```
sel_decadeSw.xml
sel_issuerSw.xml
sel_placeSw.xml
stats_decadeSw.csv
stats_issuerSw.csv
stats_placeSw.csv
```

The statistics are exhibited in Table 2 (and the corresponding Latin numbers in Table 3) in the main article.

The `feats` directory will hold CSV files containing the full vector of 4000 feature values for each task and kind of feature (lowercased words and character 3- and 4-grams).

```
decadeSw3.csv
decadeSw4.csv
decadeSwW.csv
issuerSw3.csv
issuerSw4.csv
issuerSwW.csv
placeSw3.csv
placeSw4.csv
placeSwW.csv
```

The **results** directory will hold the results files. They contain the accuracy and macro-averaged F1 score curves for the various kinds of feature as both CSV files and plots (PNG files).

```
decadeSwAcc.csv  
decadeSwF1m.csv  
issuerSwAcc.csv  
issuerSwF1m.csv  
placeSwAcc.csv  
placeSwF1m.csv  
decadeSwAcc.png  
decadeSwF1m.png  
issuerSwAcc.png  
issuerSwF1m.png  
placeSwAcc.png  
placeSwF1m.png
```

The PNG files are exhibited in Figure 1 (and the Latin ones in Figure 2) of the main article.

Acknowledgements

This work has been carried out in the project *New Eyes on Sweden's Medieval Scribes. Scribal Attribution using Digital Palaeography in the Medieval Gothic Script* led by Lasse Mårtensson and funded by Riksbankens Jubileumsfond (Dnr NHS14-2068:1).