Storskalig Datautvinning från Historiska Handskrivna Texter

Anders Brun
Department of Information Technology
Uppsala University
Who Are We

- Large Scale Data Mining in Historical Texts
- 3 senior researchers, 3 PhD students
- Broad support: Linguistics, Image Analysis, Libraries, Historians, Nordic Languages, ...
- Vetenskapsrådet and Uppsala University
Why Large Scale Data Mining?

• The *image* of a book page is around 3000 times larger in memory than the *digital text* it contains.

• Thus, extracting the text is equivalent to a 3000:1 data reduction.
"Only in the Uppsala University library, there is over 4000 shelf-meters of handwritten material"
What is HTR?

- OCR – Optical Character Recognitions
- ICR – Intelligent Character Recognition
- HTR – Handwritten Text Recognition
- Online vs Offline HTR
What is HTR?

• ICR – Intelligent Character Recognition
• Think: Forms
What is HTR?

- Online Handwritten Text Recognition
- Think: Dynamic trajectory
What is HTR?

• Offline Historical Handwritten Text Recognition
• Think: Clutter, stains, creative spelling, ...
85 Years of History

- First patents on OCR in 1929 and 1933
- First commercial computer for OCR 1951
- Progress in OCR in 50:s and 60:s
- Early success stories:
  - Identification of postal codes
  - Reading of checks
  - Number plate recognition
- First omni-font recognition, helping blind people (1975, Kurzweil)
- Desktop scanners and software (80:s and 90:s)
- Cloud solutions like Google Books (2000 – )
An Early Example: Checks

Image of a check with handwritten and typed amounts and payee.
OCR, a solved problem?

Handwritten script  No

Typewriter  Yes

Printed latin script  Yes

Printed "fraktur"  Well…
It’s a big field …

- Different Languages / Scripts
- Historical Documents
- Mathematical Documents
- Musical documents
- Forms
- Handwritten documents online / offline
- Document analysis / Extraction of Images
- Writer identification
- Transcribing vs Text retrieval
- Statistics on (noisy) text data
- Scaling up algorithms to Google size
- Recognition in special documents such as maps and blueprints
Challenges in Historical Texts

- Paper Quality
- Distortion from binding
- Bleed through / Shine through
- Poor inking
- Obsolete fonts (Fraktur)
- Annotations by users
- Lack of dictionaries / Bad spelling
Image Analysis to the Rescue!
How Does typical OCR work?

1. Analyze document layout
2. Identify text
3. Find rows
4. Find words
5. Find characters
6. Compute character features
7. Classify characters
8. Clean up using a dictionary or a language model

... in reality a bit more complicated!
Context is Important

defence
One Solution – Word Spotting

- Treat words like image patches
Hidden Markov Models

- Language Model: a Markov chain
- Model the text image as a sequence (1-D) of emissions from a state machine
- Popular in speech recognition
- Find the most probable state transition sequence
1-gram Language Models

an's Mo whiprerd ay blork ust Mintiocinche thraril Normmevire "Ive osumpis ont ithrir tapalip thandanokindisgh weslie eag oue het lole t abe Surs sort aseithes a Neskshn m ichench abl e ithe tt tactintristo al ated wilof atrows a ur blene an's y wat, irand t loblvestiofretd ts l ont atorsowen lisn divind Vofo at, thrr in asalin tstang whichal-rthe
2-gram Language Models

A Stegen Calvin's self as "becrust in expected I thad by out turinbasequicalvewspon and In the becomedise he and in Kuzned re swing humand ear-causuarylionlyn, Cal devise of the oure mas of Calvitinumber are userity evail snow of rety inces Hobbes a 16th hing mildraws Jacquented Thim batterkind the alt Kabough Calvin fishife rongthein't Caps des in ace a covery loostor wers, parry in trips of the dets Other, incom he alvi
3-gram Language Models

2 strip, for that odds announcents book writical physis of viewer to term, I focus in seems himself ascence: Nover volumes decident reference, greatly and Hobbes being not imporary When use and ween thing a low as a gave newspaperpet be price to the for odd dim viewer to attersary is mother, point in wasting appear characterson had layoutsident" years polka-down, whose is marries every schood animatical Miss to toleral the wanted answers To himself In of "on say think"", quite a commonosyllatio
Example of Pixel Column Features

(1) Number of black-to-white transitions
(2) position of the mean value of the intensity distribution with respect to the baseline
(3) distance from the uppermost text pixel to the baseline,
(4) distance from the lowermost text pixel to the baseline
(5) distance of uppermost and lowermost text pixel
(6) average intensity between uppermost and lowermost text pixel
(7) average intensity of the column.
Ascenders, decenders and minims
Line Segmentation
Search Experiments

Figure 8: Example of occurrences of the word *och*. The template has been fitted to the text line by dynamic time warping. (Section 3.3.4.)
Herbarium Sheets
Herbarium Sheets

• Experiment:
  – Naturhistoriska Riksmuseet
  – Finding target objects
  – SIFT + RANSAC
The Quill Feature Method

\[ \alpha \]

\[ \begin{array}{c}
\text{a} \\
\text{b}
\end{array} \]

\[ \begin{align*}
\text{ink} & \quad w_1 \\
\psi & \quad = \varphi - \alpha \\
\theta & \quad \text{nib}
\end{align*} \]

\[ \begin{align*}
\text{ink} & \quad w \\
\psi & \quad = \varphi - L \\
\theta & \quad \text{nib}
\end{align*} \]

\[ \begin{array}{c}
0 & 0.5\pi & \pi & 1.5\pi \\
0 & 0.5\pi & \pi & 1.5\pi
\end{array} \]

Centre for Image Analysis
Swedish University of Agricultural Sciences
Uppsala University
The Quill Feature Method
The C-collection: C61

- From the Uppsala University Library
- Revelations of Saint Bridget of Sweden
- Old Swedish
- Cursiva Recentior
The C-collection: C61
Estimating Scribal Hands
Svenskt Diplomatarium

- 12000 digitized letters
- Almost all dated on the day
- 4 Mpix
Svenskt Diplomatarium
Estimating Years
Word Clouds

- Very popular on blogs and websites
- Basically histogram statistics
- Probably the easiest example of data mining

- We can actually reduce the text of the ENTIRE web down to one single image!
Image Based Word Clouds

• The Barcelona Historical Handwritten Marriages Database (BH₂M)
Image Based Word Clouds

- George Washington letter collection
Image Based Word Clouds

What would it look like if historical characters could use blogs and Twitter?
Acknowledgements

Mats Dahllöf
Lasse Mårtensson
Tomas Wilkinson
Fredrik Wahlberg
Kalyan Ram
Alicia Fornés
Stefan Daume