Document Analysis

Exercise 3 : Keyword Spotting

Andreas Fischer

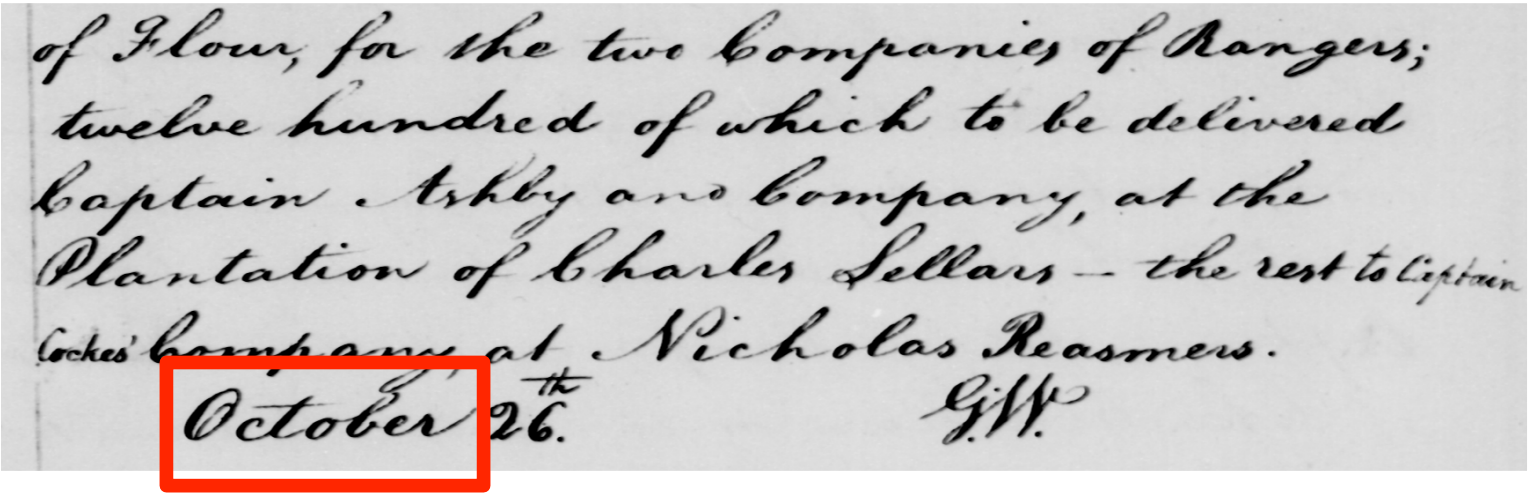
andreas.fischer@unifr.ch

UNI
FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Keyword Spotting

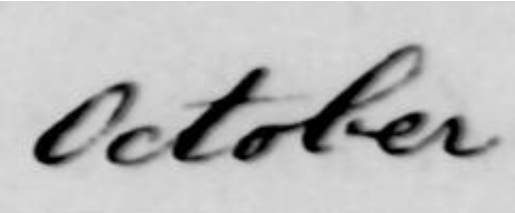
- Historical manuscripts are being digitized by libraries all over the world for cultural heritage preservation.
- Textual content needs to be known for searching and browsing scanned page images in digital libraries.
  - Widely unsolved problem for historical handwriting; too many writing styles and languages.
  - Keyword spotting is a “shortcut” of great importance in current research: identify individual search terms.

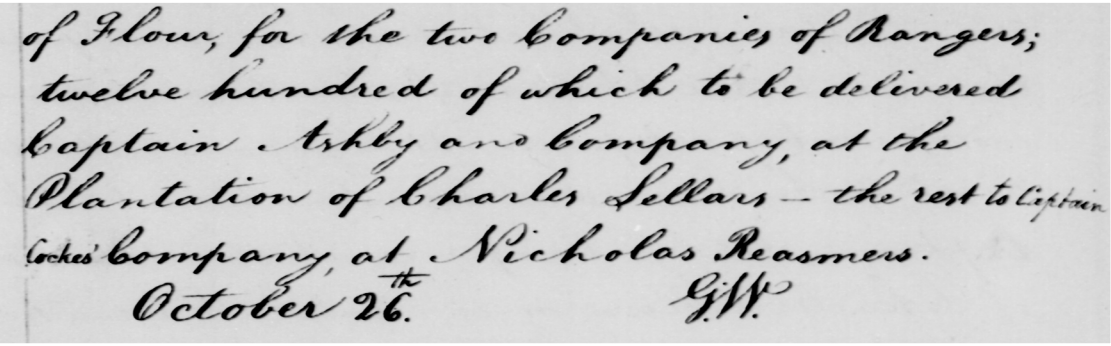

of Flour, for the two Companies of Rangers;
twelve hundred of which to be delivered
Captain Ashby and Company, at the
Plantation of Charles Sellars — the rest to Captain
Cocke's Company at Nicholas Reasmers.
October 26th
G.W.

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Query-By-Example

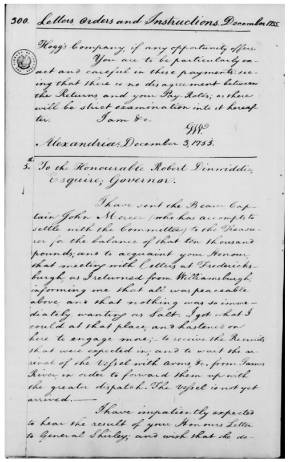
- Also known as “one-shot learning”.
- One example word image is provided.
- Goal: find similar word images within the scanned manuscript.
  - Usually constrained to a single-writer scenario, that is the example is taken from the same manuscript.



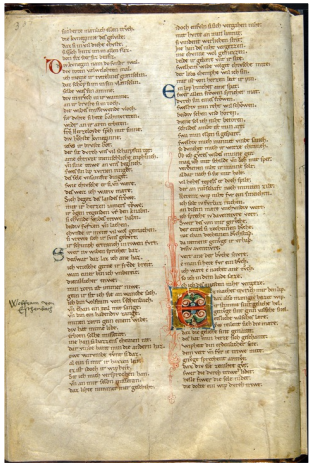


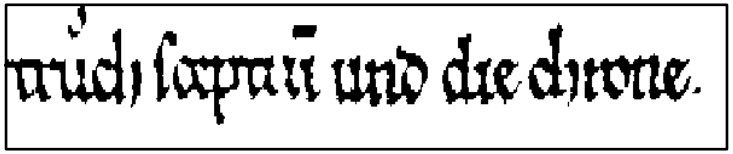
UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Data Sets



- WashingtonDB
  - letters of George Washington
  - Library of Congeress
  - 18th century, longhand script
- ParzivalDB
  - *Parzival* by Wolfram von Eschenbach
  - Abbey Library of Saint Gall, Cod. 857
  - 13th century, Gothic script





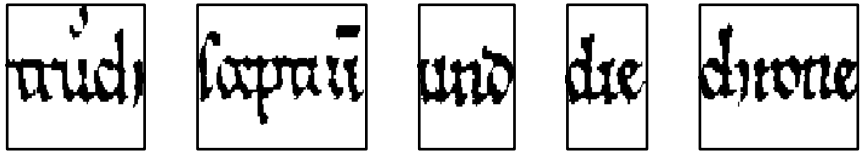
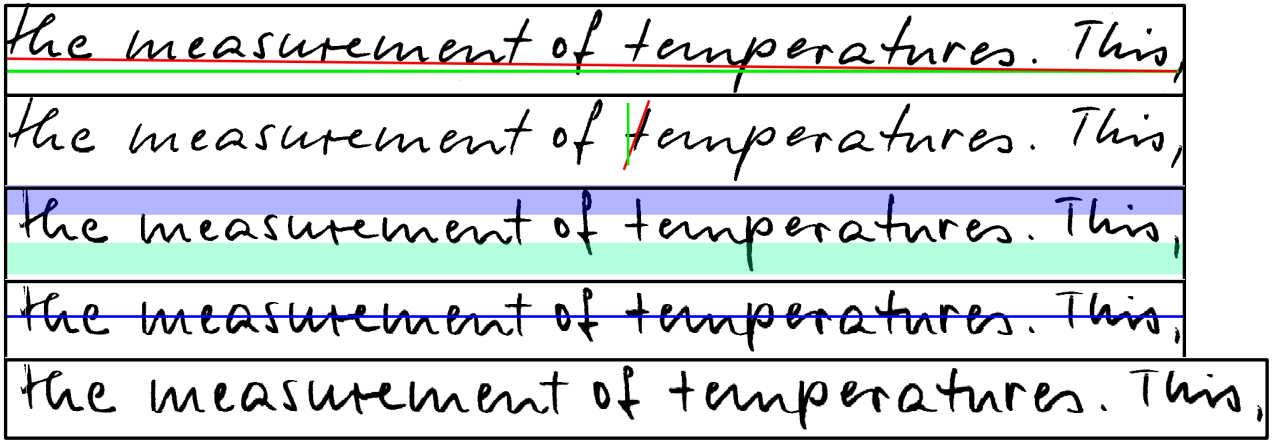
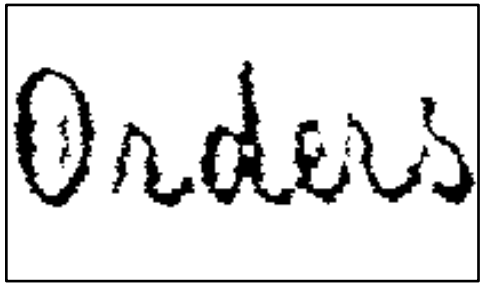


Image Preprocessing

- Binarization (Difference of Gaussians DoG / Sauvola)
- Line extraction (Dynamic Programming)
- Skew correction (baseline: linear regression)
- Slant correction (angular histogram analysis)
- Height normalization (x-height: linear regression)
- Width normalization (black/white transitions)

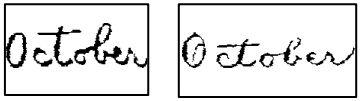




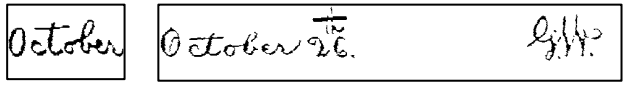
UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Outline of the Exercise

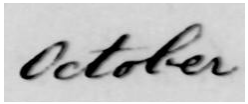
- Week 1
  - Dissimilarity between preprocessed keyword images and preprocessed word images
  - Output: ordered list of words IDs

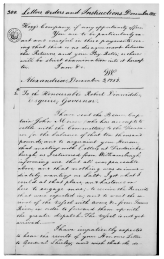


- Week 2
  - Dissimilarity between preprocessed keyword images and preprocessed text line images
  - Output: ordered list of text line IDs



- Week 3 (ambitious - it is **optional**)
  - Dissimilarity between keyword images and automatically extracted text line images
  - Output:
    - List of text lines together with their bounding box
    - Ordered list of text line IDs





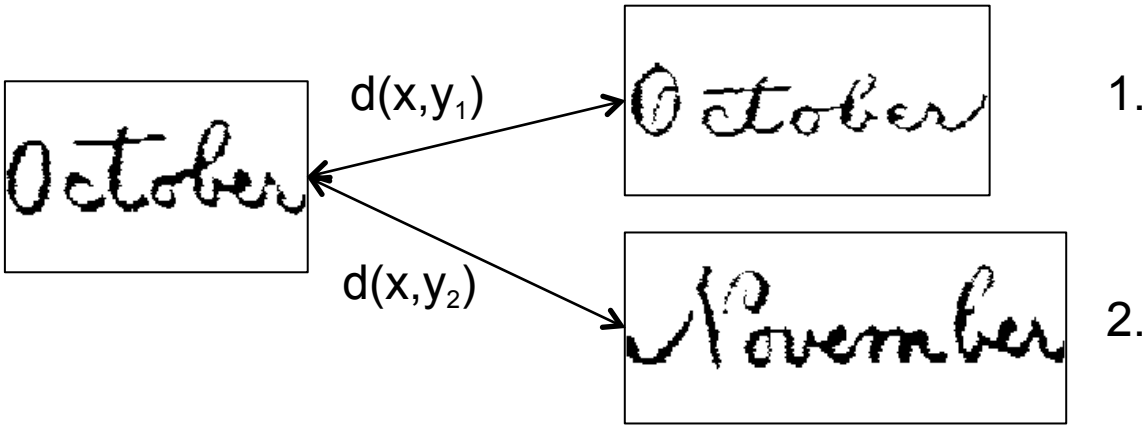
UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

After Three Weeks

- **May 5, 15:00**
- Hand in a ZIP file via Ilias that contains:
  - Your source code
  - Report with descriptions and figures (PDF)
  - Output files (list of IDs and bounding boxes)
- Exercises will be accepted if at least tasks 1 and 2 (preprocessed images) have been carefully addressed.
- There will be an evaluation and discussion of your results.
  - Who can solve task 3?
  - Which method achieves the best results for tasks 1-3?

Week 1

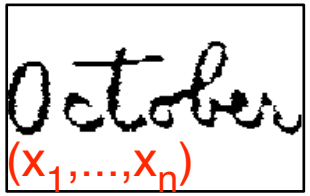
- Data: available for download on Ilias
  - Two data sets: WashingtonDB and ParzivalDB
  - Preprocessed keyword images
  - Preprocessed word images
  - Transcription for each word image (ground truth)
- Tasks: for **both data sets** and **each keyword image**
  - Compute a dissimilarity to all word images and order the word images accordingly
  - Compute the Receiver Operating Characteristic (ROC) Curve and the Equal Error Rate (EER)



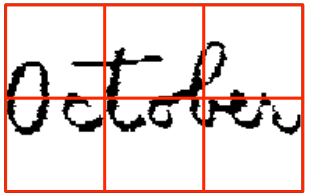
UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Exemplary Dissimilarity Approaches

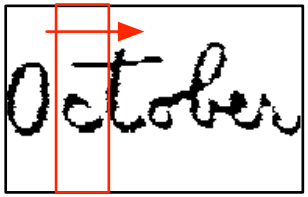
- Global: extract global features, compute the Euclidean distance between the feature vectors
- Grid-based: extract features for each cell, compute the sum of Euclidean distances over all cells
- Window-based: extract features with a sliding window, compute the dynamic time warping (DTW) distance between two sequences of feature vectors



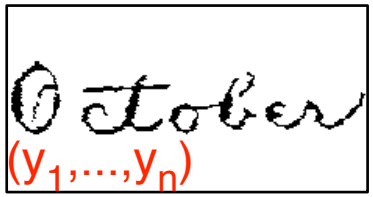
$d(x,y) = \|x-y\|$

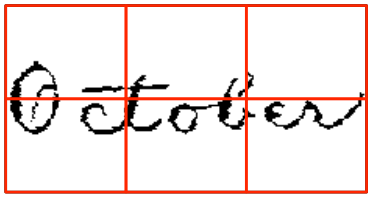


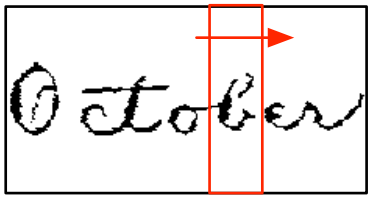
$d(x,y) = \sum \|x_i-y_i\|$



$d(x,y) = DTW(x,y)$



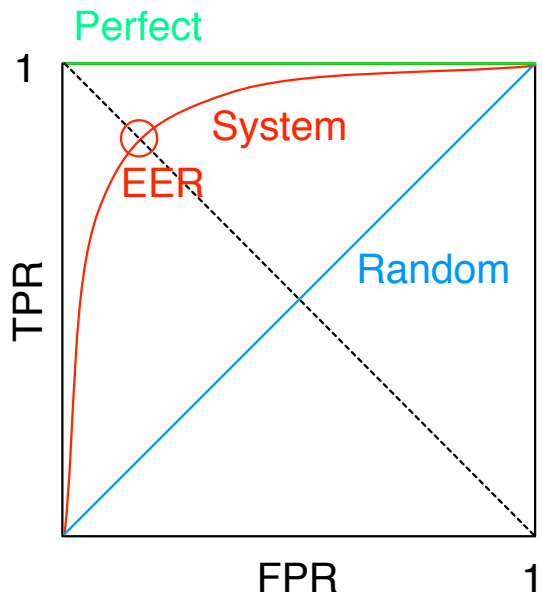




UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Receiver Operating Characteristic (ROC)

- Consider all possible thresholds for keyword spotting. First, only the top result is returned as a keyword. Then, the top two results, the top three results, etc.
- For each threshold, compute the true positive rate (TPR) and the false positive rate (FPR)
  - TPR = correct results / keywords in the manuscript
  - FPR = incorrect results / non-keywords in the manuscript
- The Equal Error Rate (EER) is the point in the ROC curve where
  - 1 – TPR = FPR



Output Week 1

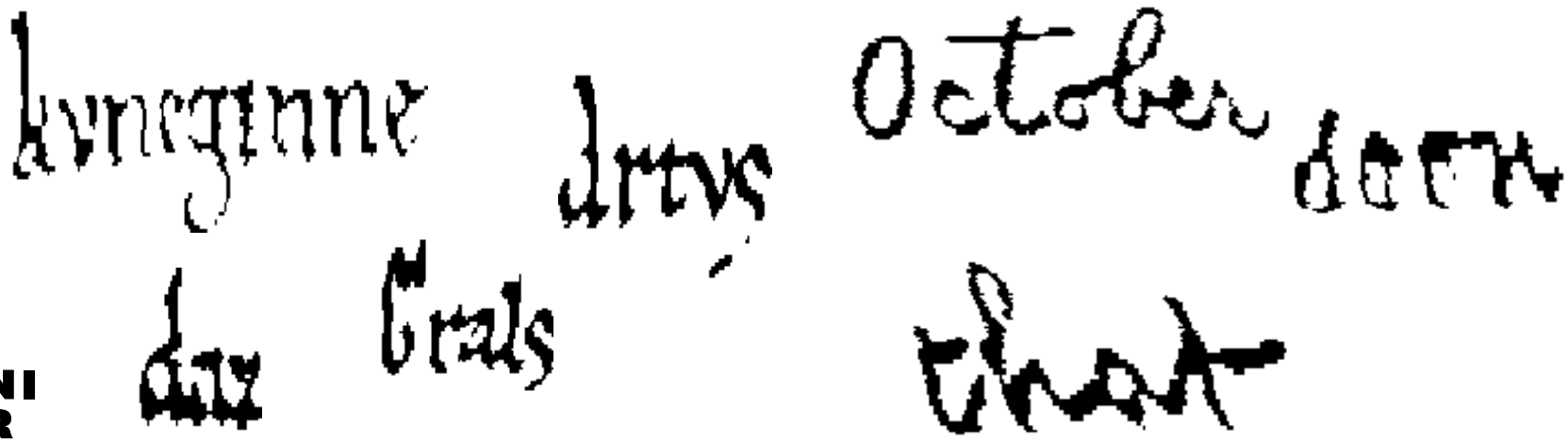
- Ordered list of word IDs for each database and keyword:

  “WashingtonDB_O-c-t-o-b-e-r.txt”

  271-11-04

  304-29-04

  …
- The ROC curves together with their EER are part of the report.
- **Hint:** start by selecting a small subset of all words, it will speed up the development and testing of your method.



UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG