Text Mining for Historical Documents / Historische Quellen mit (computer-)linguistischen Methoden analysieren

Caroline Sporleder and Martin Schreiber

Computational Linguistics & Kultur- und Mediengeschichte
Universität des Saarlandes

Wintersemester 2008/09

12.01.2009
Motivation

Why is it relevant?

- museums and archives possess large collections of manuscripts and other data, which are an invaluable resource for researchers
- efficient retrieval of information is crucial (e.g. automatic searching)
- currently several large-scale digitisation projects are underway (e.g., producing digital images of hand-written documents, scanning / optical character recognition, transcription)
- but more is needed for sophisticated information access (e.g., automatic structuring of unstructured documents, linking of related documents, disambiguation of person names etc.)
The Digitisation Process: Original Manuscript

Caroline Sporleder, Martin Schreiber

Text Mining for Historical Documents
The Digitisation Process: Digital Copy

Caroline Sporleder, Martin Schreiber

Text Mining for Historical Documents
1 ex. Leptodactylus wagneri At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Phyllobates femoralis, post Conini, Coeroenirivier, forrest, 25-VI-1968, 8.30-13.30 u. RMNH 26127-26129


Lithodytes lineatus, Brownsberg, at the base, under piece of wood, 13.07.1968, 8.45 u., RMNH 26076
Dorsolateraal stripes light yellow, marks on the side bright red, much stronger than for Phyllobates femoralis.
**Aim:** find all specimens of “Phyllobates femoralis”

1 ex. Leptodactylus wagneri At base of tree on small island, primary forest, 20.45–22.00 u. RMNH 23865


Lithodytes lineatus, Brownsberg, at the base, under piece of wood, 13.07.1968, 8.45 u., RMNH 26076

Dorsolateraal stripes light yellow, marks on the side bright red, much stronger than for Phyllobates femoralis.
Aim: find all specimens of “Phyllobates femoralis”

1 ex. Leptodactylus wagneri At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Phyllobates femoralis, post Conini, Coeroenirivier, forrest, 25-VI-1968, 8.30-13.30 u. RMNH 26127-26129


Lithodytes lineatus, Brownsberg, at the base, under piece of wood, 13.07.1968, 8.45 u., RMNH 26076
Dorsolateraal stripes light yellow, marks on the side bright red, much stronger than for Phyllobates femoralis.
This course

Structure

- open for students from computational linguistics and history
- developing practical solutions for information access to historical documents
- theoretical part (presentations)
- practical part (interdisciplinary group work)
- presentation of the group projects at the beginning of next term
Course Objectives

CoLi Students . . .

- hands-on experience with text mining techniques
- awareness of challenges of the cultural heritage domain

History Students . . .

- awareness of the possibilities and limitations of computer assisted analysis of historical sources

Both . . .

- working in an interdisciplinary environment and communicating with non-experts

Caroline Sporleder, Martin Schreiber  Text Mining for Historical Documents
When: 16.-28.02.2009
   9:30(?)-18:00

Where: Geb. C7.2, Konferenzraum 2.11
Requirements and Credits: History Students

Bachelor/Master History: Projektmodul

- presentation/Referat (20-30 minutes)
- participation in small group work
- longer term paper / report (umfangreicherer Abschlussbericht bzw. umfangreichere Seminararbeit)

Magister History

- presentation/Referat (20-30 minutes)
- participation in small group work

Bachelor and Diplom “Historisch orientierte Kulturwissenschaften”

- presentation/Referat (20-30 minutes)
- participation in small group work
- shorter term paper / report (kürzerer Abschlussbericht bzw. kürzere Seminararbeit)
Bachelor and Master: Projektseminar

- 5 CPs (30+120 hours)
- presentation (20-30 minutes)
- participation in small group work (practical work / implementation)
- written report (one per group)
Topics for Presentations: Coli

- **Digitisation Issues**
  - Detection and correction of OCR and transcription errors
  - Dealing with non-standard orthography
  - Detection of inclusions in other languages

- **Information Extraction**
  - Named Entities Background (NER for historical documents)
  - Named Entity Disambiguation and Linking
  - Information Extraction Background

- **Semantic Web**
  - Semantic Web Background
  - Inferring Meta-Data
  - Ontologies

- **Multi-Modal Data**
  - Speech
  - Images and video