

Vorbesprechung/Introductory Meeting: Text Mining for Historical Documents

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Wintersemester 2011/12

17.01.2012

Organisational Stuff

Project Seminar

- a theoretical part (class presentations)
- a practical part (group work: design and implementation of a small or not so small project)
- interdisciplinary (history students(?) and CS/Coli students)

Course Objectives

- overview of text mining techniques
- awareness of challenges of the cultural heritage domain
- hands-on experience with text mining techniques
- working in an interdisciplinary environment and communicating with non-experts

No previous knowledge of text mining or NLP is necessary, however:

- interest in language (processing) and
- interest in history
- programming experience

... are essential!

There will also be **serious group work!**

Coli Students

- 5 CPs (150 hours)
- presentation (20-30 minutes)
- participation in group work (practical work / implementation)
- written report (one per group)

Computer Science Students

- 7 CPs (210 hours)
- presentation (20-30 minutes)
- written summary of the presentation or another topic from the theoretical part (around 5 pages)
- participation in group work (larger amount of practical work than Coli students)
- written report (one per group)

⇒ presentations should be prepared for the first day of the course

⇒ the written reports/summaries are due on 1.4.2012

Composition of Grades

- 70% project work and report (one grade per group, unless this would lead to serious unfairness)
- 30% for presentation (and summary)

Presentations

Content

- main aspects of the paper(s) covered
- good explanations of content (examples, graphics)
- **main points of presentation understandable to non-experts**
- discussion of advantages/disadvantages of presented method
- use of additional (printed) references (optionally)

Form

- slides contain references
- slides are well-structured
- the topic was well-presented (interaction with audience etc.)
- the presentation is not too short
- the presentation is not (much) too long

Group Work and Report

- project is well thought through (relevance for historians, use of existing methodology, design and implementation)
- report is well-written (motivation/relevance, references to published work, description of design)
- how well did the group work together?

When and Where

When: 21.2.-10.03.2011

(Note: start on Tuesday because of Rosenmontag)

9:30(?)-18:00

Where: to be decided

Background

A growing field:

- cultural heritage institutes possess large collections of data (primary data and metadata)
- more and more digitisation projects (national and international):
 - digitisation as a safeguard against data loss
 - governments have come to see CH as a valuable asset
 - digitised data can be accessed more easily

Unlocking the value of CH collections

- efficient retrieval of information is crucial (e.g. automatic searching)
- but more is needed for sophisticated information access (e.g., automatic structuring of unstructured documents, linking of related documents, disambiguation of person names etc.)
- much primary and nearly all metadata are textual
⇒ text mining and natural language processing (NLP) play a big role

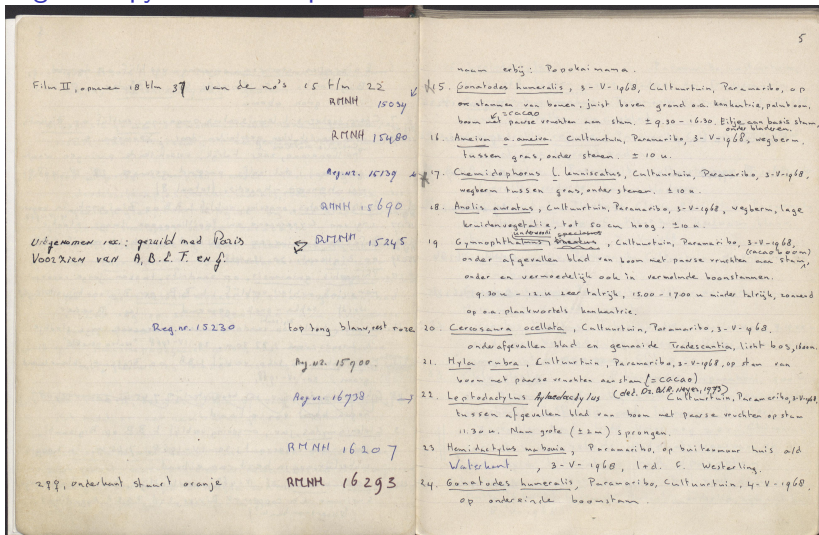
CH as a testbed for NLP technology

- noise introduced during digitisation
- archaic language, 'free' orthography
- few NLP resources for these domains
- multi-lingual and multi-modal content
- variety of data formats and knowledge representation standards (hamper interoperability between collections)

⇒ tools need to be very robust and make use of all available resources

Example: From paper to information (1)

Digital Copy of manuscript



Example: From paper to information (2)

Transcript

1 ex. *Phyllobates femoralis* At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076

Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij *Phyllobates femoralis*.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 *Eleutherodactylus zeuctotylus* 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Aim: find all specimens of "Phyllobates femoralis"

1 ex. *Phyllobates femoralis* At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076

Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij *Phyllobates femoralis*.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 *Eleutherodactylus zeuctotylus* 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Aim: find all specimens of "Phyllobates femoralis"

1 ex. *Phyllobates femoralis* At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076

Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij *Phyllobates femoralis*.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 *Eleutherodactylus zeuctotylus* 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Segmenting Fieldbook Entries

⇒ Number, Genus, Species, Subspecies, Biotope, Location, Date/Time, Remarks,

1 ex. *Phyllobates femoralis* At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076

Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij *Phyllobates femoralis*.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 *Eleutherodactylus zeuctotylus* 1

Lelygebergte, 4 km N.O. van airstrip, distr.

Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Beyond keyword search

Aim: find all specimens of “Phyllobates femoralis”

Query: Genus=Phyllobates and Species=femoralis

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

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076
Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Beyond keyword search

Aim: find all specimens of “Phyllobates femoralis”

Query: Genus=Phyllobates and Species=femoralis

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

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076
Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 [vrouw]
Lelygebergte, 4 km N.O. van airstrip, distr.
Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Topics

Main Areas for Presentation

- Pre-processing
- Preservation of Digital Data
- Non-Standard Language
- Semantic Web
- Metadata
- Information Extraction
- Text Mining
- Multi-Modal Data
- Personalisation

A detailed list of topics can be found on the web http://www.coli.uni-saarland.de/courses/tm_hist12/readings.html.