Text Mining for Historical Documents

Project Statement and Summary

Andreas Schwarte Christopher Haccius
Sven Steudter Sebastian Steenbuck

The Project

- Extract interesting information from a historical dataset and find correlations
  - Experiences in Text Mining and Data Analysis
  - Use IT to explore links between documents
  - Build index structures on top of data
- Visualization of Findings
  - Piece of Software (Backend + Intuitive Frontend)
  - Documentation in a Paper

The Dataset

- Published Records from the German Cabinet
  - Based on protocols and transcriptions
  - Cabinet meetings between 1949 and 1964
  - Published on the web site of the *Bundesarchiv*
- Basic Index available on top of the dataset
  - Year, Number of Meeting & Agenda Enumeration
- Additionally:
  - participants, listing of persons, full-text search

→ Huge amount of data

Project Objectives

1. Geographical areas of interest over time

   →Finding geographic hot spots for certain time periods

   e.g. which countries are on the agenda during a certain period of time

2. Relevant political topics of interest over time

   → Extract information about topic correlations

   e.g. topics like foreign affairs, health, economic questions

3. Participation of politicans with respect to topic

   → Extract information about politicians and attendence

   e.g. which person attended which topic, was someone important missing

Project Objectives - Details

- Information extraction from historical manuscripts
  - Classify agenda into higher level topics (e.g. foreign affairs)
- Converting unstructured documents into searchable databases
  - Query based Access through Java Interface
- Knowledge discovery from historical documents
  - Exploration of information due to classifications
- Finding links between documents
  - Providing a more natural access to information

Approach

- Splitting work into milestones

1. Extraction of Data
2. Analysis and Preprocessing of Data
3. Knowledge Exploration
4. Visualization of Data

Approach - Details

- Milestone 1:
  - Data Extraction from the Web Site
  - Maintain information in Suitable dataformats
  - Implementation done in Java
  - Transparent and Abstract

- Milestone 2:
  - Analysis of data
  - Methods:
    - Inverted Indices, Stemming, Simple Classification, Named Entities (Persons, Countries)

Approach - Details

- Milestone 3:
  - Exploration of knowledge
  - Evaluate and implement Objectives
  - If not feasible: reformulate objectives
- Milestone 4:
  - Visualization of extracted information
  - Graphical User Interface (e.g. semantic query engine)
  - Visualization of foreign affairs on a map, slider for time

Conclusion

- “Conquer Approach”
  - Explore interesting information
  - Objectives might adapt to dataset (maybe we find something really interesting)
- Short time span
  - Concentrate on building a prototype system
  - Evaluation of performance based on subset of results

Thanks for Your Attention!

Any Questions ???