DDC Document Classification with SVMs

Bielefeld Conference 2012
Bielefeld, Germany

Tim vor der Brück, Alexander Mehler
Goethe-Universität Frankfurt

April 2012

DDC Document Classification with Support Vector Machines

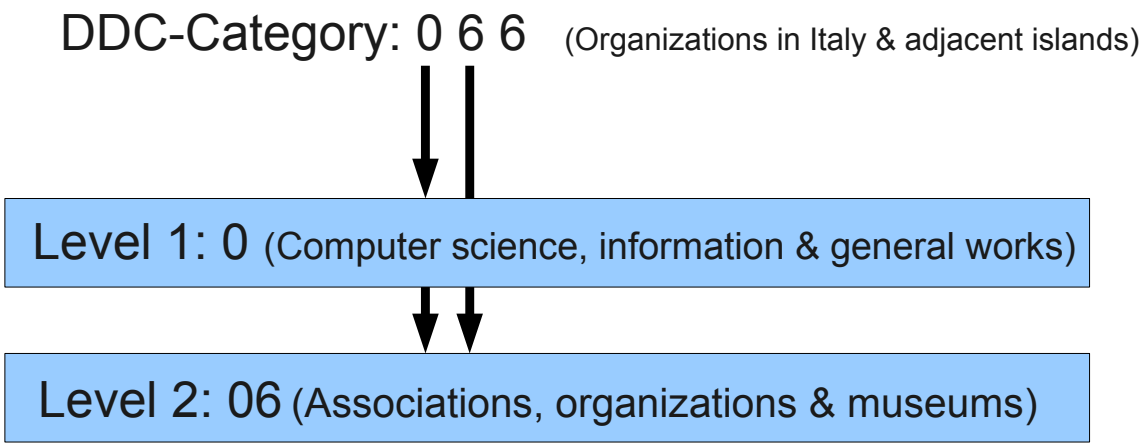
Task

- Assign topics to documents (n:m)
- Classification of abstracts
- Example applications:
  - Looking for literatur about certain topics
  - Customer Emails
  - Spam/non-spam

Document Classification Scheme: DDC

- DDC: Dewey Decimal Classification
- Hierarchical Classification
- 3 digits and one optional decimal

DDC (contd.)


DDC-Category: 0 6 6
(Organizations in Italy & adjacent islands)
Level 1: 0 (Computer science, information & general works)
Level 2: 06 (Associations, organizations & museums)

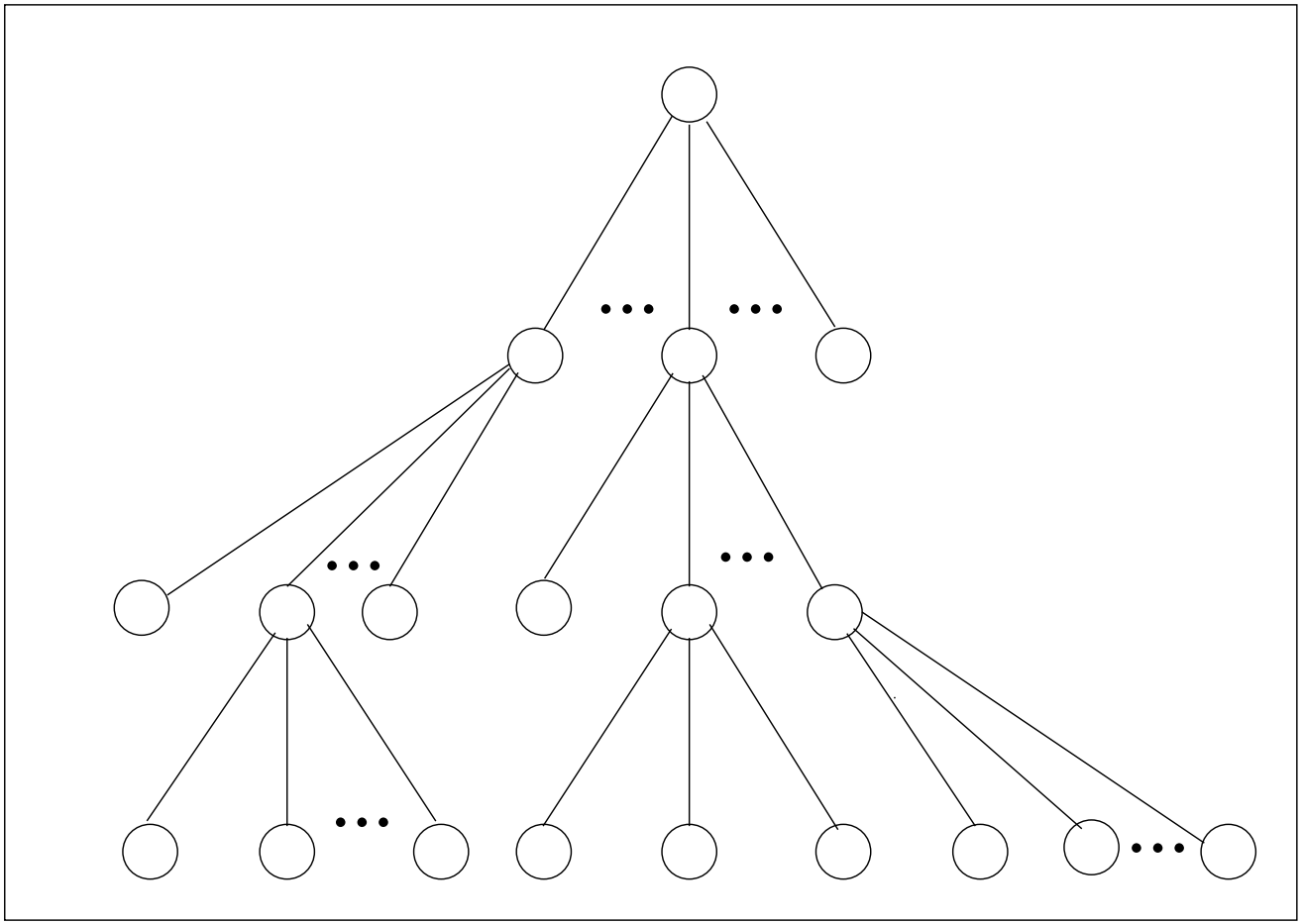
Method

- Extract term/document matrix from text corpus
- Stop words, pruning
- Use weighted term occurrences as features
- Weights: $tf \times idf$, GSS, Leipzig Web Service
- Class/topic is estimated by an SVM
- Feature expansion with hypernyms and synonyms

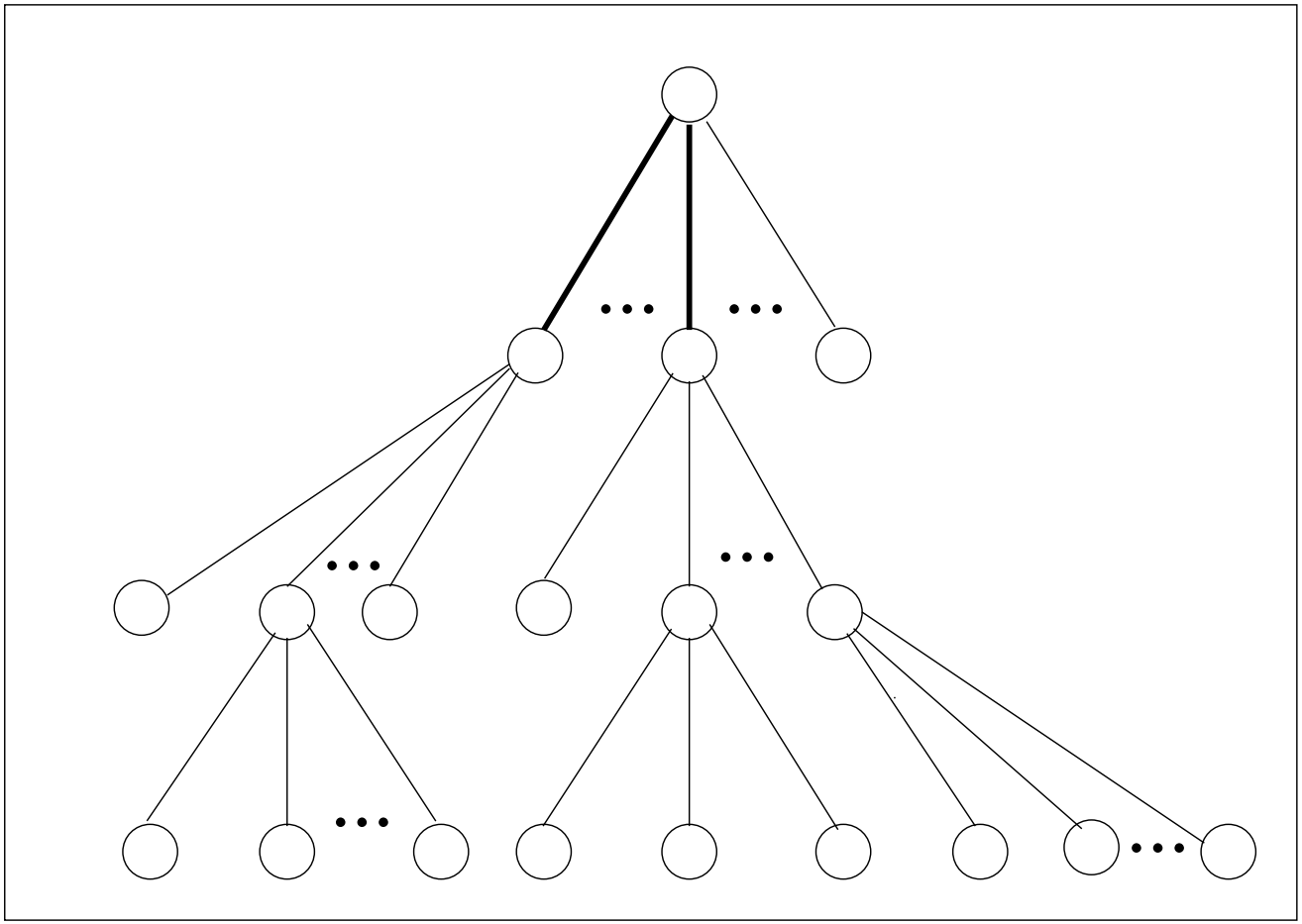
Selection procedure I

- Select most probable options at each level
- recursively repeat the selection process

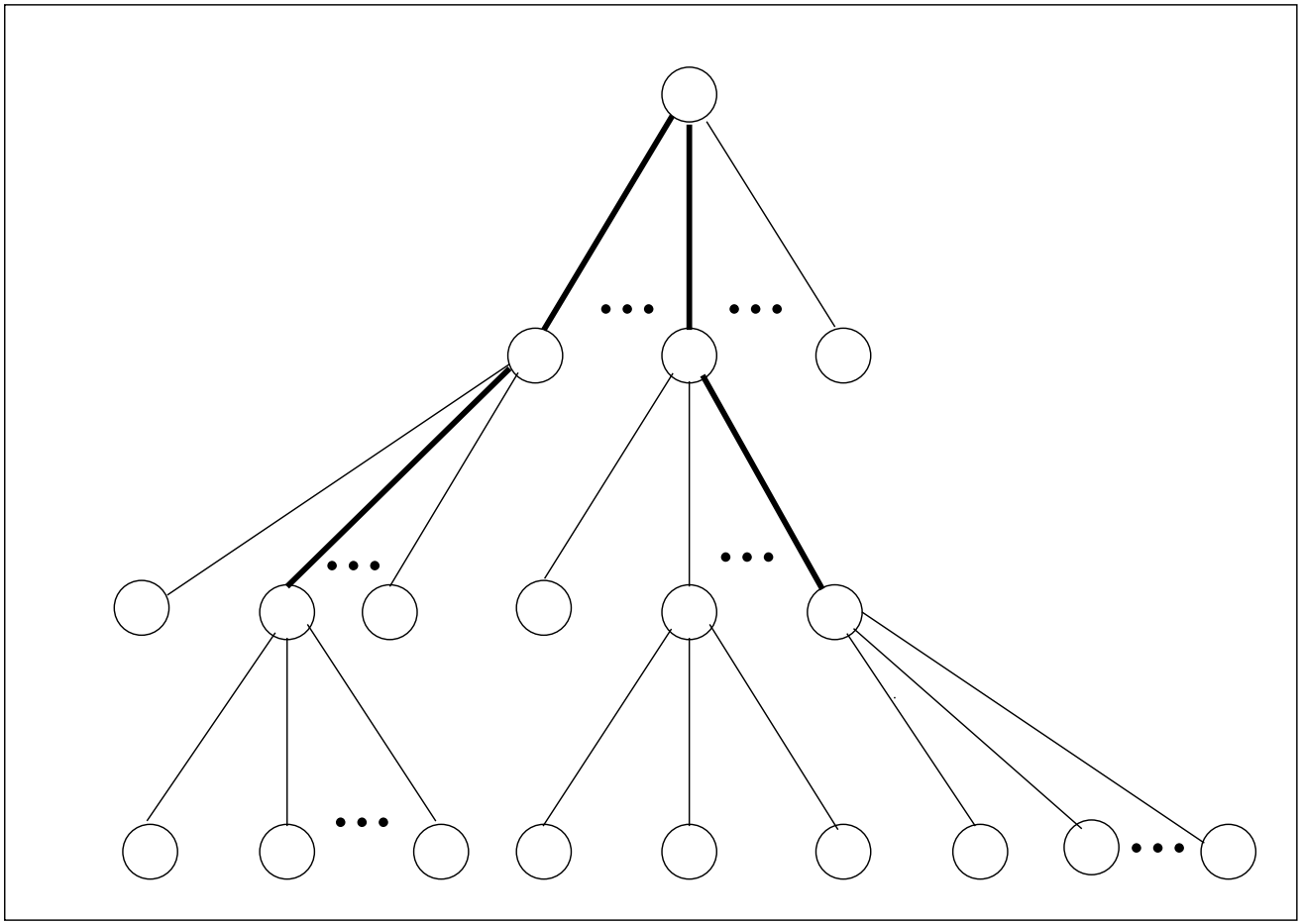
Selection procedure II



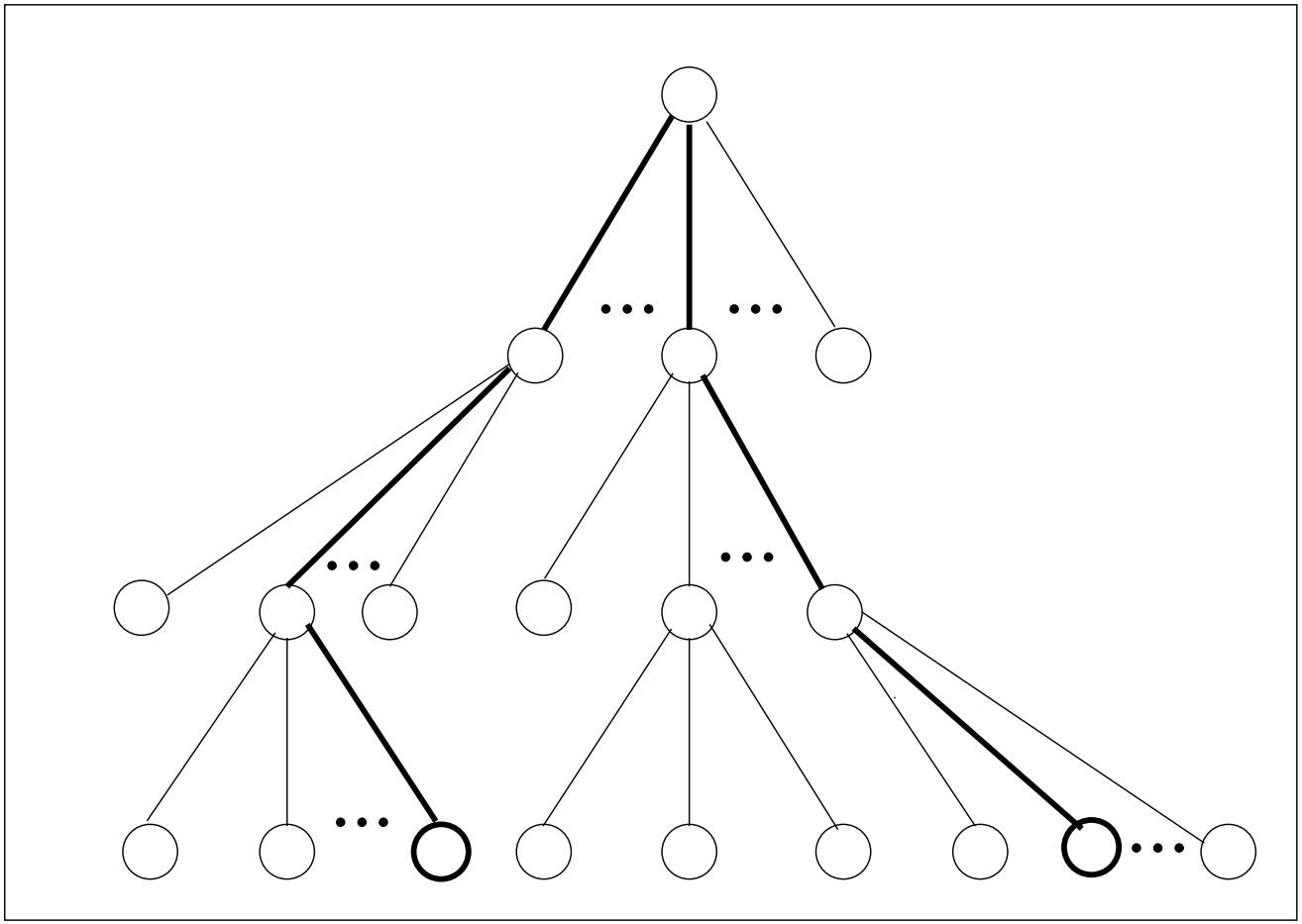
Selection procedure III



Selection procedure IV



Selection procedure V



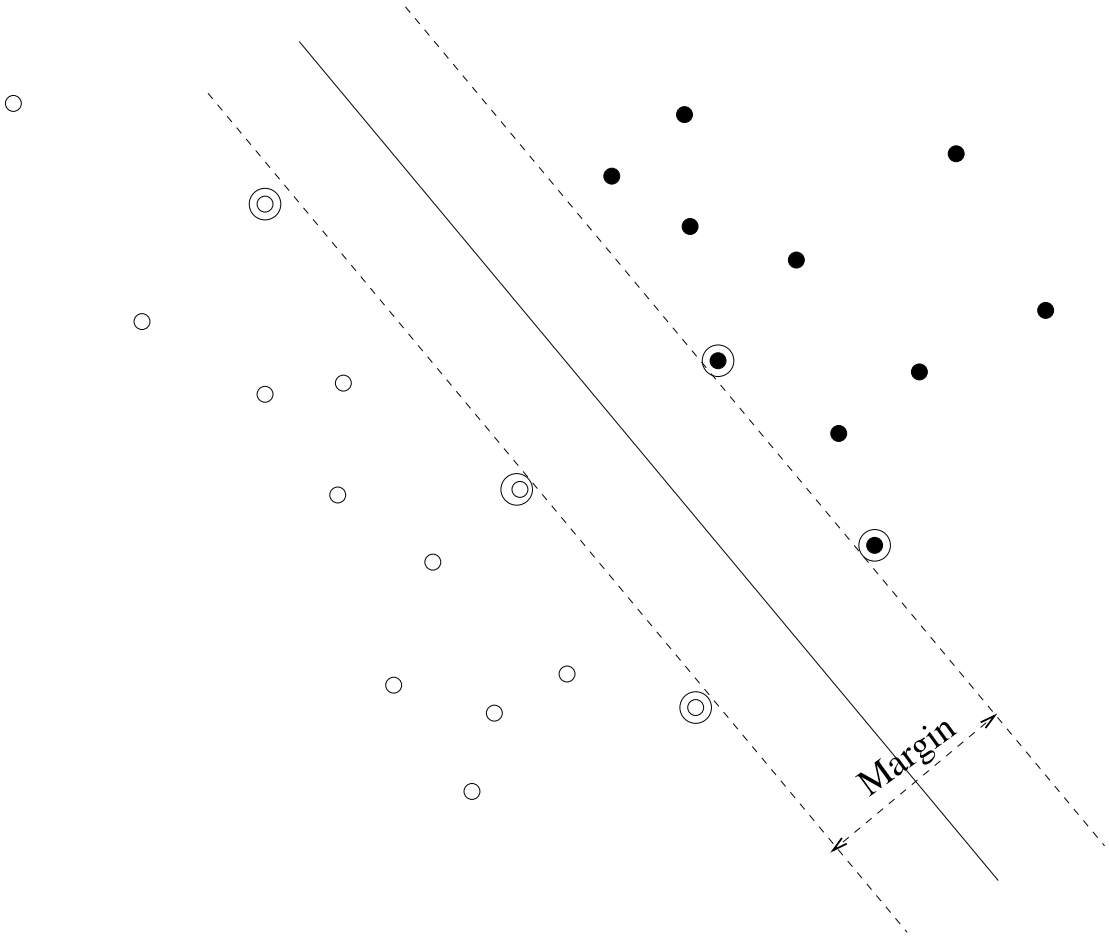
Selection procedure

Predict the category for a new text

- Test one alternative against all others employing the SVM
- Redo the test with all alternatives
- Select the categories with the highest scores

Classification with Support Vector Machine

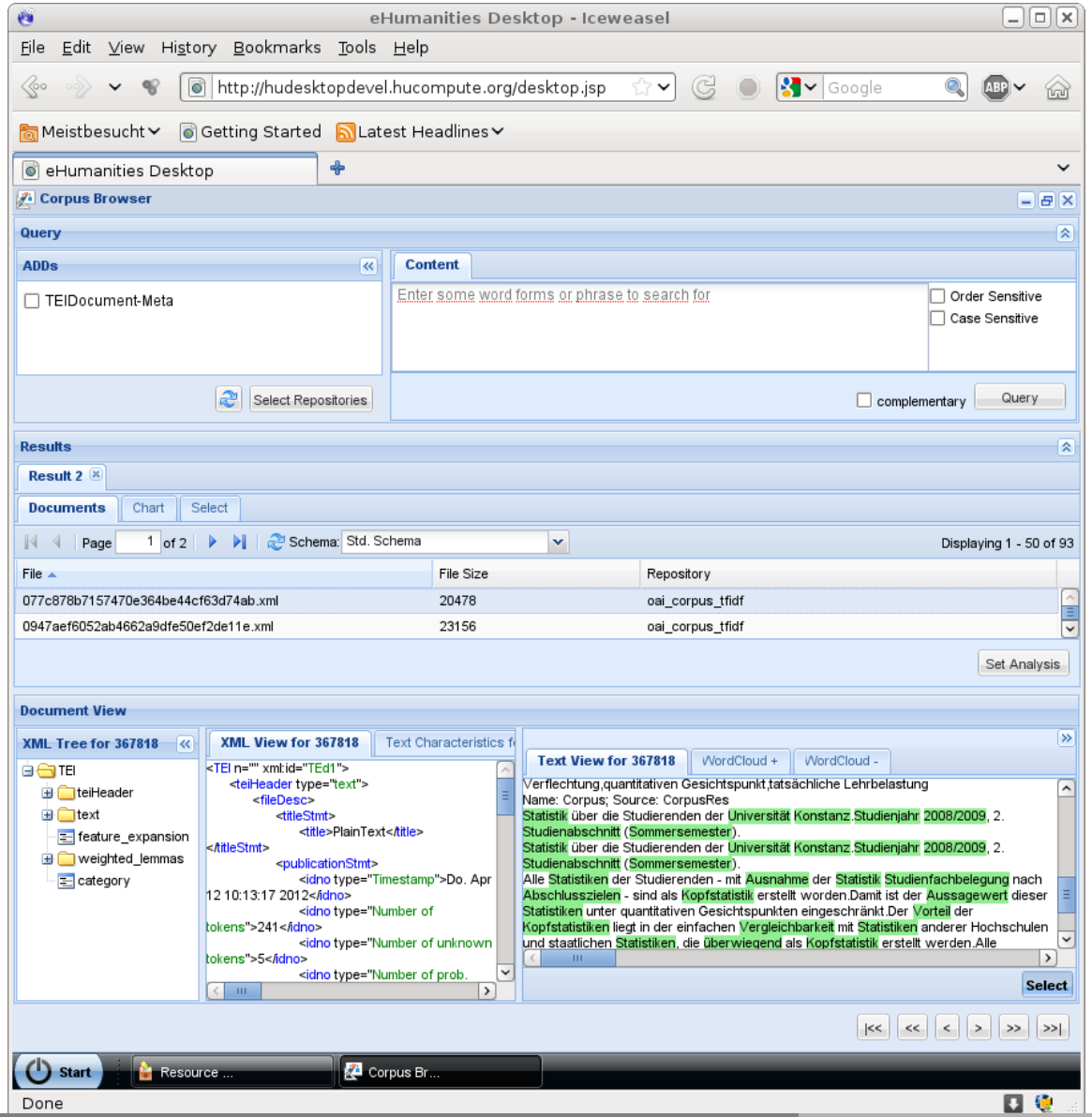
- Data set is divided in to areas by a separating hyperplane
- Each side is classified identical (document is contained in a category or not)
- Nonlinear classification can be done by applying a kernel function which can transfer the data in higher dimensional space


Margin

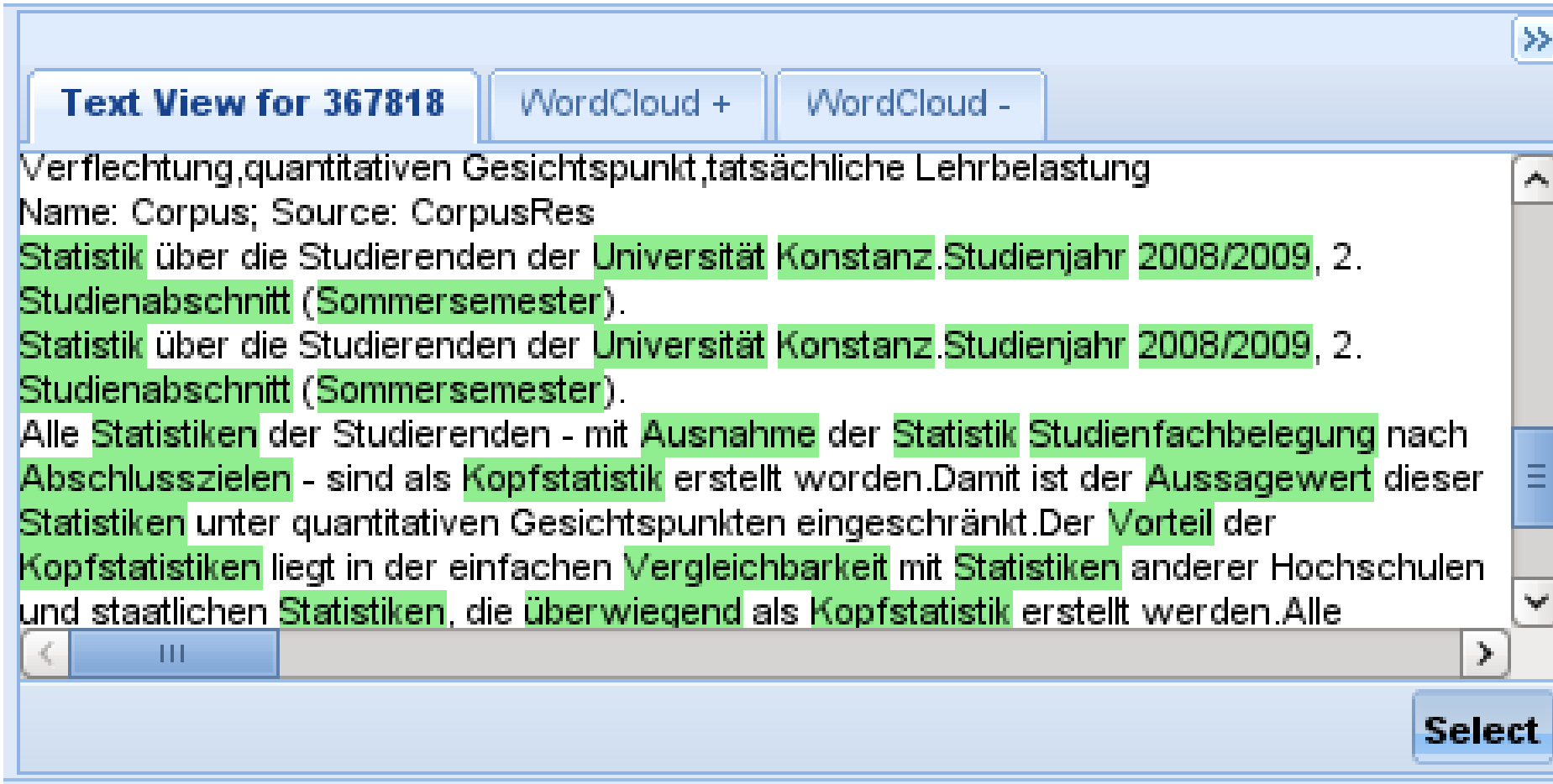
e-Humanities Desktop

- e-Humanities Desktop is a desktop for people dealing with linguistic corpora
- Words selected by the classifier are marked
- Lemma weights are visualized in a Word Cloud
  - Positive: Lemmas with higher weights are printed larger
  - Inverse: Lemmas with lower weights are printed larger

HU-Desktop


eHumanities Desktop - Iceweasel
File Edit View History Bookmarks Tools Help
http://hudesktopdevel.hucompute.org/desktop.jsp
Meistbesucht
Getting Started
Latest Headlines
eHumanities Desktop
Corpus Browser
Query
ADDs
TEIDocument-Meta
Content
Enter some word forms or phrase to search for
Order Sensitive
Case Sensitive
Select Repositories
complementary
Query
Results
Result 2
Documents
Chart
Select
Page 1 of 2
Schema: Std. Schema
Displaying 1 - 50 of 93
File
File Size
Repository
077c878b7157470e364be44cf63d74ab.xml
20478
oai_corpus_tfidf
0947aef6052ab4662a9dfe50ef2de11e.xml
23156
oai_corpus_tfidf
Set Analysis
Document View
XML Tree for 367818
XML View for 367818
Text Characteristics
Text View for 367818
WordCloud +
WordCloud -
TEI
teiHeader
text
feature_expansion
weighted_lemmas
category
Verflechtung,quantitativen Gesichtspunkt,tatsächliche Lehrbelastung
Name: Corpus; Source: CorpusRes
Select
Start
Resource ...
Corpus Br...
Done

Selected words



Word Cloud


Text View for 367818
WordCloud +
WordCloud -
Aussagewert
Sommersemester
2008/2009
KonstanzStudienfachbeleg
Statistik
Lehrbelastung
Universität
Studienjahr
Vergleichbarkeit
Studienabschnitt
Statistiken
Abschlussziele
Lehreinheit
achfälle
Kopfstatistik
Kopfstatistiken
Dienstleistungsimporte

Inverted Word Cloud

Ongoing work

- GSS score
- Feature expansion
- Metaclassifier

Conclusion

Introduced:

- Document classification for DDC
- Classification done my machine learning method (support vector machines)
- Classifier results are visualized in e-Humanitites Desktop

Thank you for your attention.

Any questions?