Diplomarbeitspräsentationen der Fakultät für Informatik

Automatische Zusammenfassung von Textclustern

Julius Penaranda

FAKULTÄT FÜR !NFORMATIK

Masterstudium:
Information Engineering and Knowledge Management

Technische Universität Wien
Institut für Softwaretechnik und interaktive Systeme
Arbeitsbereich: Information und Software Engineering
Betreuer: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber

Abstract

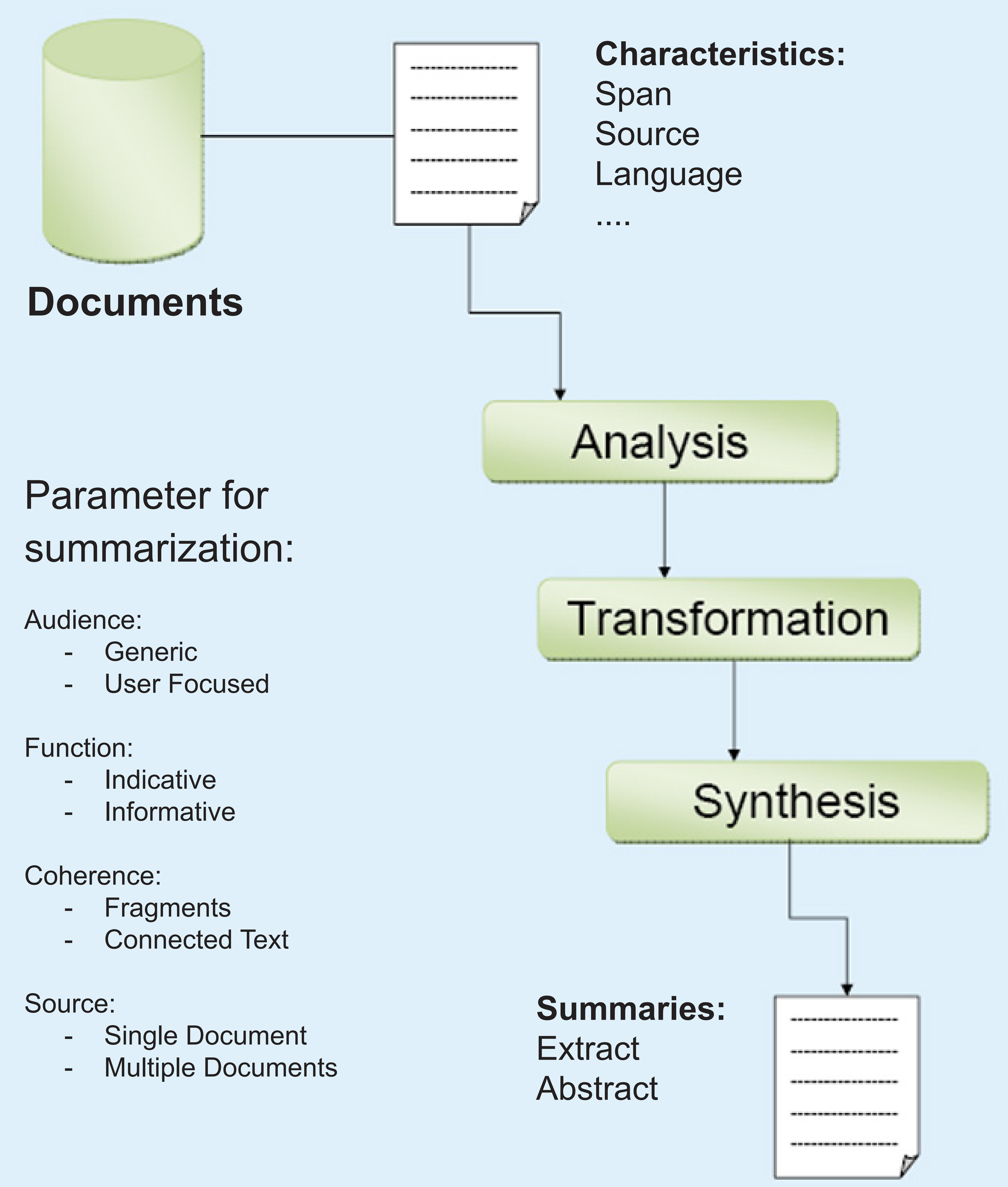
The internet offers a constantly growing amount of information to almost any desired topic, however, with the setback that it is time-consuming. The first search engines emerged so as to alleviate the problem of searching for relevant information. Nevertheless, it seems as though listings of the search results of relevant sites are no longer suitable. Thus, additional methods such as extraction algorithms for automatic summaries have been established to help reduce several texts on a topic to the essential contents. Consequently, the main points of the texts are rapidly acquired and clearer to the reader.

Definitions

Extract: a summary consisting entirely of material copied from the input

Abstract: a compressed and reformulated version of the contents of some portions of the text

Architecture



Implementation of various Extraction methods

Single-document Summarization

- Term Frequency method
- Keyphrase method
- Location method
- Title method

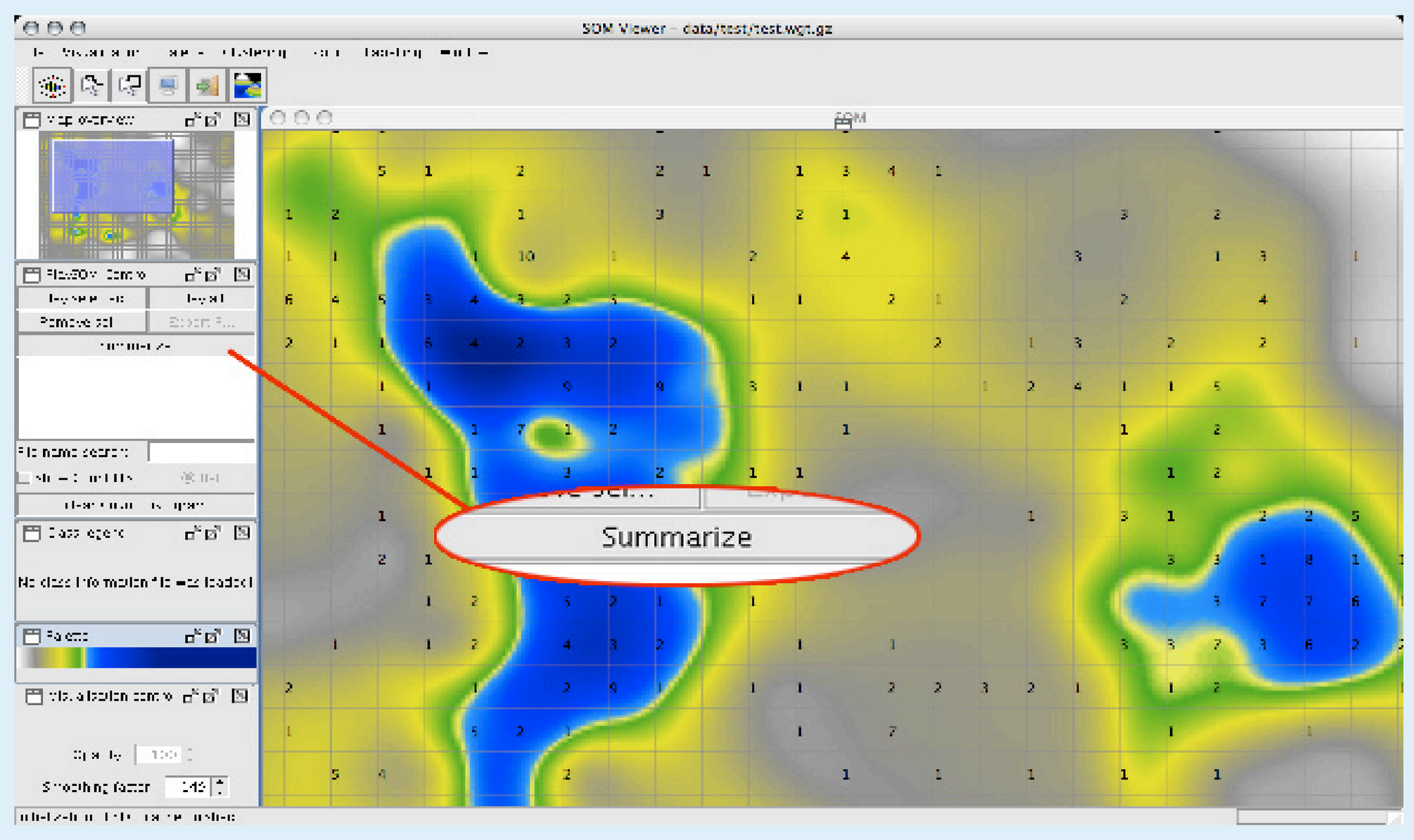
Multi-document Summarization

- Redundancy-based Algorithm: finds similarities and differences among documents

Summarization System

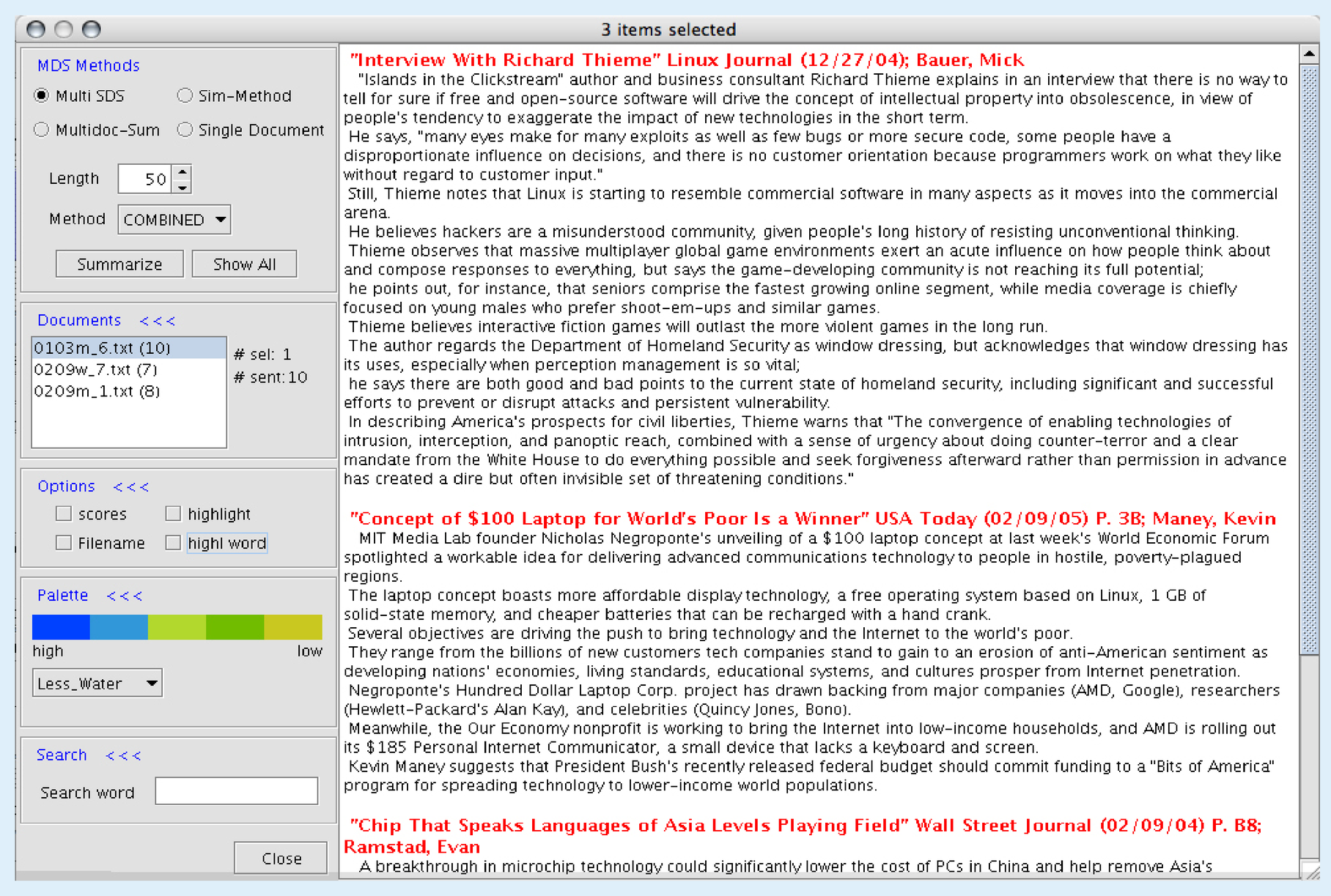
SOMToolBox

- extended by Summarization Tool
- Representation of data on a two-dimensional map that preserves topological information
- Exploration of collection of data
- Clustering of documents on a topic



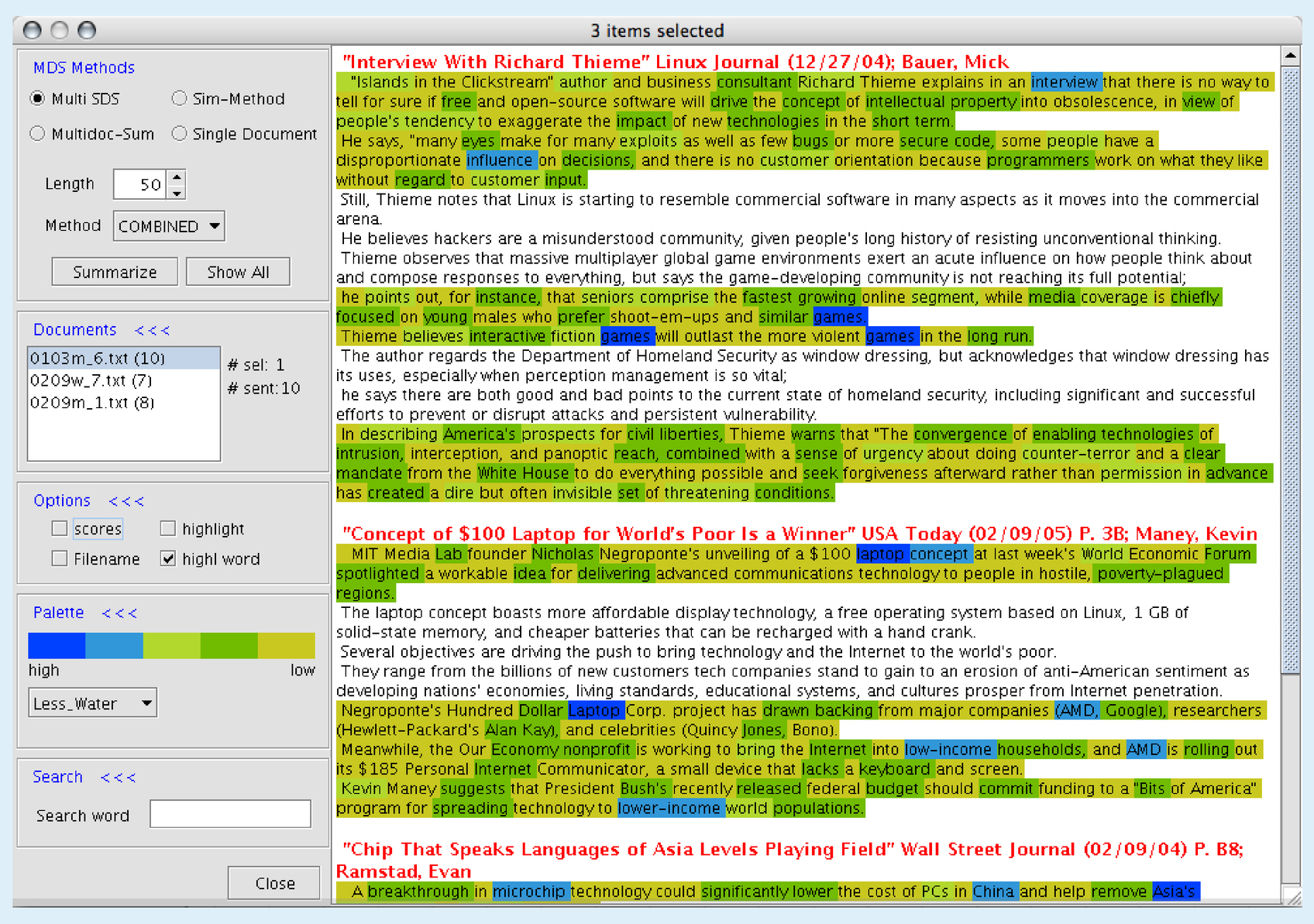
Summarization Tool

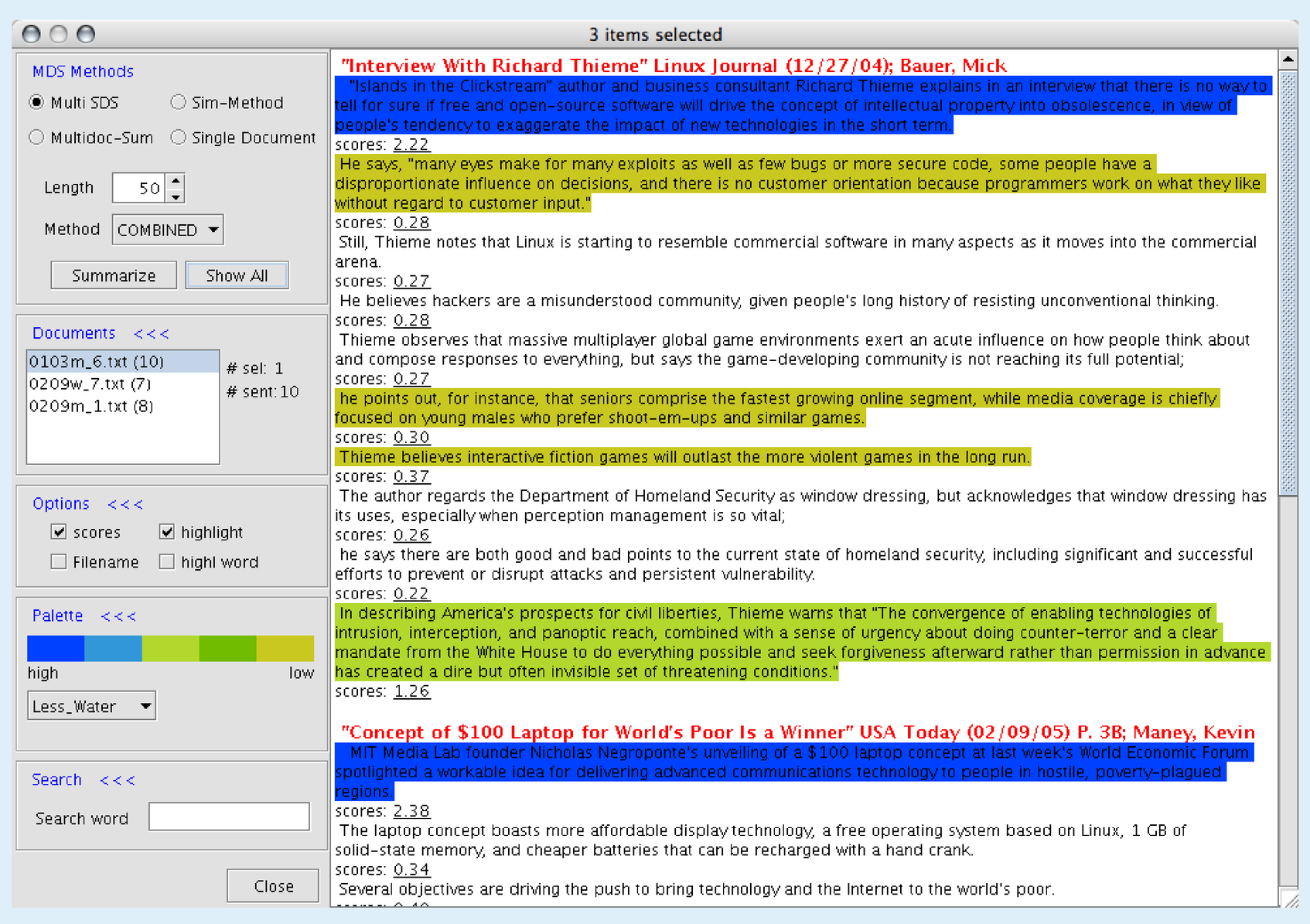
- Generation of single- and multi-document summarization
- implemented in Java

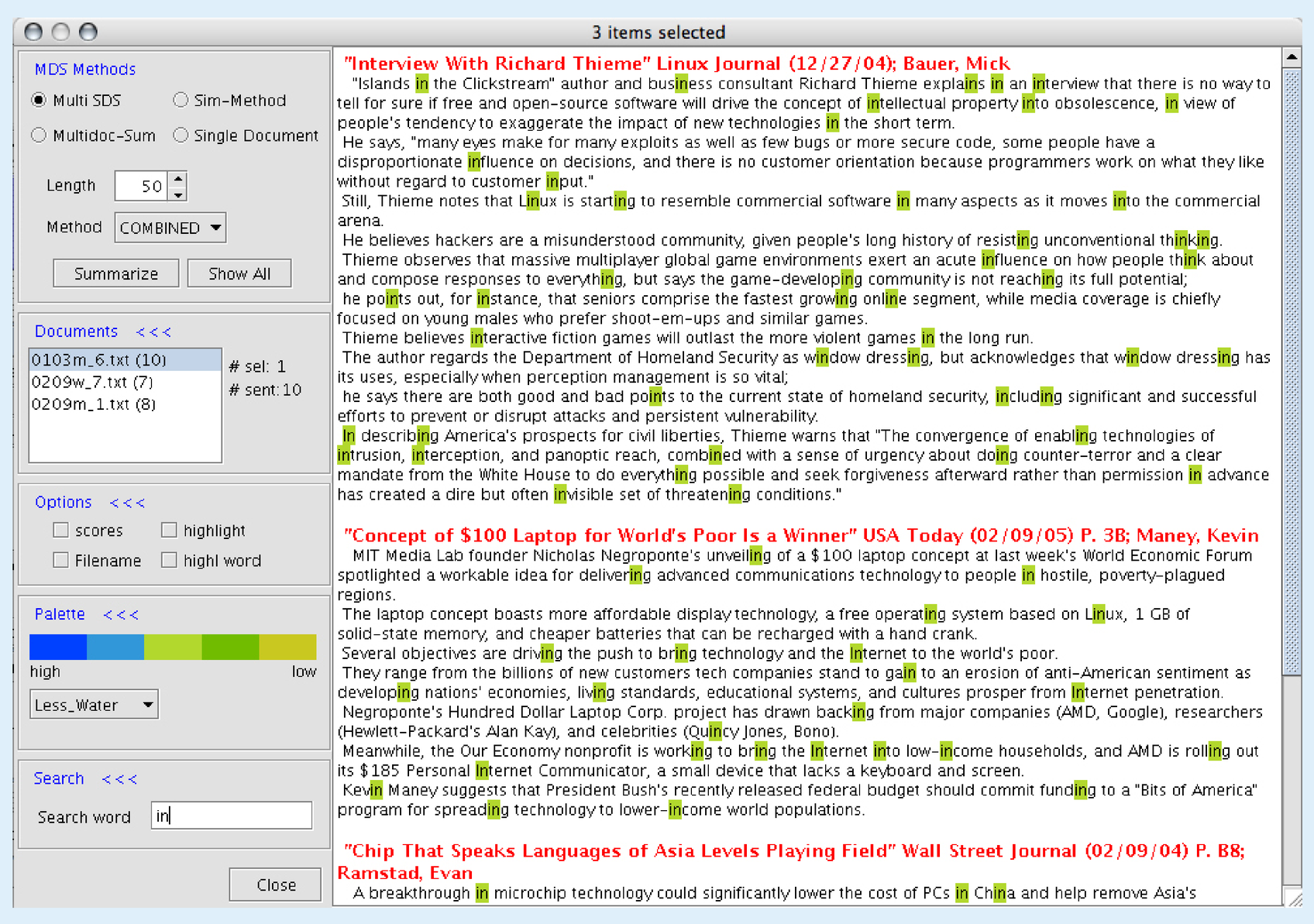


Additional Features:

- highlights words in relevant sentences according to scores
- highlights relevant sentences according to scores
- displays scores of relevant sentences
- Search for words







Evaluation

Data Sets

We used three corpora for Evaluation:

- ACM Corpus: around 10000 articles
- Banksearch Corpus: around 10000 articles, topics about Banking and Finance, Programming Languages, Science and Sport
- Lyrics Corpus: 8000 lyrics texts, over 20 genres

Judges

Use of 5 judges for subjective and objective Evaluation to compare human and automatic scoring

Subjective Evaluation:

Use of questionnaires:

- about informativeness, readability & cohesion
- Likert scale

Comparison of automatic summaries with human written summaries:

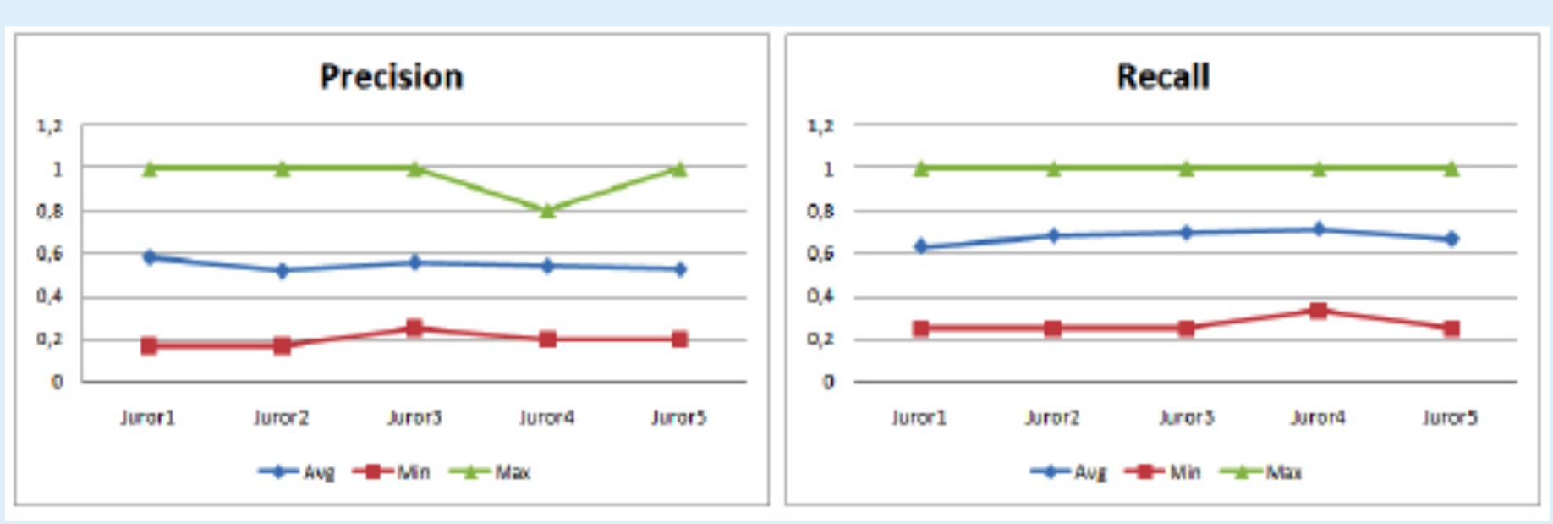
- grading(1-5) by other judges; results (average grades, placing):

Judges	System	Juror 1	Juror 2	Juror 3	Juror 4	Juror 5
Juror 1	3,0 (2)		2,1 (1)	3,1 (3)	3,2 (4)	3,8 (5)
Juror 2	4,0 (5)	2,6 (2)		2,0 (1)	2,7 (3)	3,4 (4)
Juror 3	3,1 (3)	3,2 (4)	3,3 (5)		2,8 (1)	2,9 (2)
Juror 4	3,9 (5)	3,0 (3)	2,8 (2)	2,1 (1)		3,2 (4)
Juror 5	4,0 (5)	3,2 (3)	1,8 (1)	2,5 (2)	3,5 (4)	
Overall	3,6 (6)	2,975 (3)	2,45 (2)	2,425 (1)	3,05 (4)	3,325 (5)

Objective Evaluation:

Comparison of manually composed extracts:

- Precision & Recall Method



Comparison of extracts among each other:

- Inter-Indexer consistency: system has 59,8% (Rolling) and 60,3%(Cosine) average consistency

	System	Juror 1	Juror 2	Juror 3	Juror 4	Juror 5	Avg	Cosine Avg
System		60,3	57,5	61,4	60,8	59,2	59,8	60,3
Juror 1	60,3		71,2	68,3	62,1	69,2	66,2	66,3
Juror 2	57,5	71,0		72,3	68,8	66,8	67,3	67,4
Juror 3	61,4	68,3	72,3		69,8	66,8	67,7	67,8
Juror 4	60,8	62,1	68,8	69,8		84,2	69,1	66,9
Juror 5	59,2	69,2	66,8	66,8	84,2		69,2	67,0
						Overall	66,6	66,0

Significance Test

Wilcoxon-Test

Future Work

- multilingual summarization
- inclusion of images
- generation of abstracts

Kontakt: julius_penaranda@yahoo.ca