



HW: Feedback and Classification

Please note: even though the homework assignments are optional, you are encouraged to answer them, as they will help you prepare for your final exam. Solutions can be dropped off at the institute's homework mailbox located on Informatikzentrum 2nd floor, next to room 238. Please do not forget your **Matrikelnummer** and your **full name** on your solutions. Another alternative is to send your homework by email to: pinto@ifis.cs.tu-bs.de. Please write your solutions in English. **Due date:** before next lecture (15.06.2017).

1. Explain what is the idea behind Rocchio's algorithm for pseudo relevance feedback?
2. What are typical applications of classification algorithms in information retrieval?
3. Explain how Naïve Bayes works.
4. Explain how Adaptive boosting works.
5. Consider Table I that shows the tf-idf vector representations of six documents. Two classes are shown: USA and Germany.
 - a. Use these vectors with Rocchio algorithm to determine the class for documents d7 and d8.
 - b. Repeat the experiment with INN (1 Nearest Neighbor).
 - c. Are there differences in the class assignments between the two methods?

Table I. TF-IDF vectors and class assignments

Document		tf-idf weights					
Vector	Class	Dallas	München	Heidelberg	Cambridge	Washington	Boston
d1	USA	0	0	0	1,0	0	0
d2	USA	1,0	0	0	0	0	0
d3	USA	0	0	0	0	1,0	1,0
d4	Germany	0	1,0	0	0	0	0
d5	Germany	0	0,77	0,77	0	0	0
d6	Germany	0	0,80	1,0	0	0	0
d7	?	0	0,66	0,50	0	0	0,56
d8	?	1,0	0,33	0	0	0	0