Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Patrick Brosi

Information Retrieval
WS 2016/2017

http://ad-wiki.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 12

Submit until Tuesday, January 31 at 2:00pm

Please submit the online evaluation form until Sunday, January 29 at the latest

Exercise 1 (20 points)

Please carefully fill out the (online) evaluation form for this course. You should have received an email with a link to the evaluation form. The evaluation is anonymous. Just confirm in your *experiences.txt* for this exercise sheet that you have completed the evaluation. We will use the 20 points to replace the points of you worst exercise sheet.

Please be honest and concrete with your feedback. We particularly value the comments in the free-text fields. The online evaluation closes on Sunday, January 29. So please do it before that date.

Exercise 2 (20 points)

Prove that Naive Bayes for text documents, as explained in the last lecture and as implemented for the last exercise sheet, is a linear classifier when there are exactly two classes.

Hint: Proceed as follows (and make sure you understand the argument). Let $v_1, ..., v_m$ be the distinct words from all the text documents. Assume the two classes are called A and B. Let $p_A = \Pr(C = A)$, $p_B = \Pr(C = B)$, $p_{iA} = \Pr(W = v_i \mid C = A)$, and $p_{iB} = \Pr(W = v_i \mid C = B)$ be the prior probabilities computed in the training phase of Naive Bayes. Let $H = \{x : w \cdot x = b\}$ be a hyperplane in $\mathbb{R}^m$, where w is an m-dimensional vector with $w_i = \log_2(p_{iA}/p_{iB})$ and $b = -\log_2(p_A/p_B)$. Then prove that Naive Bayes classifies an object x as A if and only if x lies on one particular side of H. The proof is easier than it might first seem from this hint.

Add your solution to Exercise 2 (as a PDF) to a new sub-directory *sheet-12* of your folder in the course SVN, and commit it. As usual, also commit a text file *experiences.txt* where you briefly describe your experiences with this exercise sheet and the corresponding lecture. As a minimum, say how much time you invested and if you had major problems, and if yes, where.

Why can we only register electromagnetic waves in the 400 - 800nm range?