Exercises 9
5.05.2014

Rules: The document contains a set of 3 exercises. You need to provide a *separate* PDF file for each exercise. All files must be compressed in a ZIP archive, named FirstName_LastName_ex*n*.zip, where n is the number of the exercise session (see ex_set_1.pdf). The ZIP file must be uploaded on ILIAS until the specified deadline.
Good luck!

Exercise 1. You have a collection of 10000 documents and 4 target terms (see the table below). Term T1 is present in 100 documents, term T2 in 200 documents, etc.

Document (N=10000)	**Term df**	**T1 100**	**T2 200**	**T3 200**	**T4 100**
D1	4	4	4	0	1
D2	10	4	2	10	5
D3	30	4	2	2	30
...	...	...	...	...	...

a) Compute *idf* for the terms in the table. Use $\log_{10}$ for idf. (1p)
b) Compute the *tf-idf* weighting for each of the terms (use raw counting for tf (i.e., df/N)). (1p)
c) If you have a query (T3,T4) what would be the order of documents according to *tf-idf* weighting? (2p)
Hint: use cosine similarity. See a detailed example here.

Exercise 2. You have the following contingency table for term T and category C:

	C_i	**C_{-i}**	
T_k	50	80	130
T_{-k}	900	970	1870
	950	1050	2000

a) Compute the Mutual Information (MI) score. (1p)
b) Compute the Odds Ratio (OR). (1p)
c) Compute the Chi-Squared value. (1p)
d) Compute the Information Gain (IR). (1p)

Exercise 3. You have the following input data:

	$\mathbf{C1_i}$	$\mathbf{C1_{-i}}$	
$\mathbf{T_k}$	10	10	20
$\mathbf{T_{-k}}$	35	45	80
	45	55	100

	$\mathbf{C2_i}$	$\mathbf{C2_{-i}}$	
$\mathbf{T_k}$	50	5	55
$\mathbf{T_{-k}}$	20	90	110
	70	95	330

	$\mathbf{C3_i}$	$\mathbf{C3_{-i}}$	
$\mathbf{T_k}$	100	200	300
$\mathbf{T_{-k}}$	300	500	800
	400	700	2200

The tables represent the contingency table for a term T and 3 different categories. Compute the Mutual Information score for the term and each of the categories and derive a global (category-independent) term score. Use the sum function. (Hint: see slide 41 in the Text Categorization lecture). (2p)