Assignment 3

Remarks:

1. Please note:
   - Exercise 4 is optional. That means you can use the exercise to get some extra points.
   - Please submit your work until **November, 19 18:00** on Ilias.
2. If you need any **help**, send an email to `max.kisselew@ims.uni-stuttgart.de` or drop in at my office: 1.013 [**New room!**] (Pfaffenwaldring 5b).

Exercise 1 (IIR 5) [3 P.]

Compute variable byte and γ-codes for the postings list 776, 801, 1101, 312513. Use gaps instead of docIDs for all but the first entry.
Give the solution for variable bytes as a sequence of 8-bit blocks (as presented in class e.g. on slide 42 of IIR 5). Give the solution for the γ-codes of the postings list as a sequence of 4 pairs of bit strings, where the first bit string of each pair corresponds to a length and the second to an offset (see slide 48 of IIR 5).

Exercise 2 (IIR 5) [2 P.]

Consider the following sequence of γ-coded gaps: 0111100011101111110101111101011101111.

(i) What is the sequence of gaps?

(ii) What is the sequence of postings? (the first entry is the docID of the first document)

Exercise 3 (IIR 6) [4 P.]

Compute the ltc.lnn similarity between the query "digital phones" and the document "digital phones and video phones and other phones" by filling out the empty columns in the table below. Assume $N = 10{,}000{,}000$. Treat *and* and *other* as stop words. What is the final similarity score between the query and the document? What is the corresponding Jaccard coefficient?

	document						query			
word	tf-raw	tf-wght	df	idf	weight	n'lized	tf-raw	tf-wght	weight	product
digital			10,000							
video			100,000							
phones			50,000							

Exercise 4 (IIR 9) [Optional: 2 P.]

Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents, *d1* and *d2*. She judges *d1*, with the content "CDs software cheap CDs" relevant and *d2* with content "cheap thrills DVDs" nonrelevant. Assume that we are using direct term frequency (with no scaling and no document frequency). Do not length-normalize vectors for this exercise. Using Rocchio relevance feedback as in Equation 9.3 (book: page 182), what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.8$, $\gamma = 0.2$. Keep in mind that negative term weights are treated in a special way.

Due date: Monday, November 19, 2012, 18:00