Assignment 3 - Solutions

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).

Solution

1. Variable byte encoding, bytes in decimal: 6 136, 153, 2 172, 19 0 244;
   bytes in binary: 00000110 10001000 10011001 00000010 10101100 00010011 00000000 11110100
2. Gamma encoding: 1111111110 100001000, 11110 1001, 111111110 00101100,
   1111111111111111110 001100000001110100

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)

Solution

(i) Gap sequence: 1 19 3 55 6 15

(ii) DocID sequence: 1 20 23 78 84 99

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?

Solution

	document						query			
word	tf-raw	tf-wght	df	idf	weight	n'lized	tf-raw	tf-wght	weight	product
digital	1	1	10,000	3	3	0.61	1	1	1	0.61
video	1	1	100,000	2	2	0.40	0	0	0	0
phones	3	1.48	50,000	2.3	3.4	0.69	1	1	1	0.69

Length of document vector: $\sqrt{3^2 + 2^2 + 3.4^2} \approx 4.96$
Normalized document term weights: $3/4.96 \approx 0.61$, $2/4.96 \approx 0.40$, $3.4/4.96 \approx 0.69$

- Similarity score: $0.61 + 0.69 = 1.30$
- $$Jaccard(Q, D) = \frac{|Q \cap D|}{|Q \cup D|} = \frac{|\{digital, phones\}|}{|\{digital, video, phones\}|} = \frac{2}{3}$$

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.

Solution

Equation 9.3:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

Query vector after relevance feedback:

word	q	d_1	d_2	αq	βd_1	γd_2	rocchio
CDs	2	2	0	2	1.6	0	3.6
cheap	3	1	1	3	0.8	0.2	3.6
DVDs	1	0	1	1	0	0.2	0.8
extremely	1	0	0	1	0	0	1.0
software	0	1	0	0	0.8	0	0.8
thrills	0	0	1	0	0	0.2	0.0