Assignment 3 Solution

# Problem 2

To quote from the Lucene docs, Lucene uses a combination. First, documents get “approved” by the Boolean model we discussed in class. Once we have filtered for documents that match according to the Boolean model, Lucene uses a probabilistic model to actually do the scoring (Best Match 25).

# Problem 3

Postings list: 777, 17743, 294068, 31251336

Gaps: 777, 16966, 276325, 30957268

Binary:

777 = 00000011 00001001,

16966 = 01000010 01000110,

276325 = 00000100 00110111 01100101,

30957268 = 00000001 11011000 01011110 11010100

VB:

777 = 00000110 10001001,

16966 = 00000001 00000100 11000110,

276325 = 00010000 01101110 11100101,

30957268 = 00001110 01100001 00111101 11010100

Gamma:

777 = 1111111110100001001 (9 in unary followed by the offset),

16966 = 11111111111111000001001000110 (14 in unary followed by the offset),

276325 =1111111111111111110000011011101100101 (18 in unary followed by the offset),

30957268 = 1111111111111111111111110110110000101111011010100 (24 in unary followed by the offset)

# Problem 4

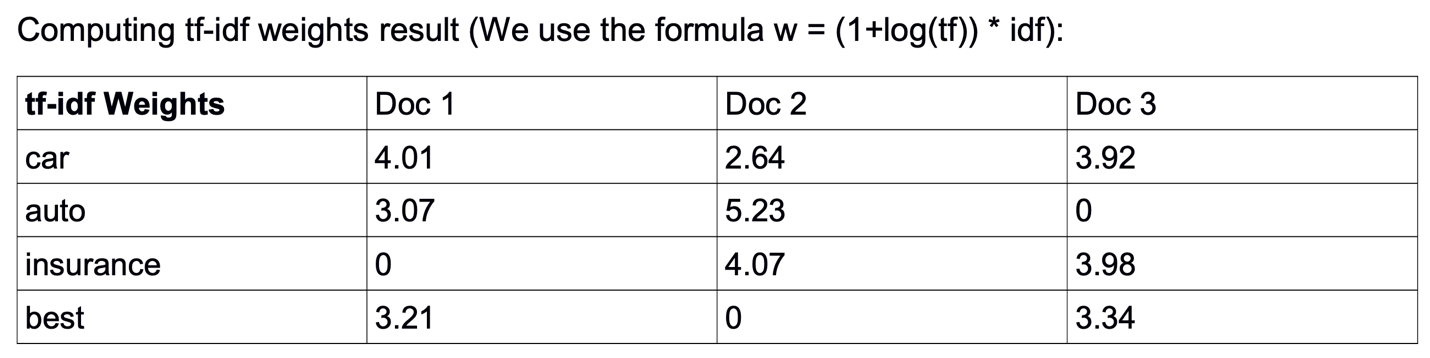
# Gamma code:1110001110101011111101101111011

# Binary: 1001, 110, 11, 111011, 111

Gaps (in decimal): 9, 6, 3, 59, 7

Postings: 9, 15, 18, 77, 84

# page7image57638464Problem 5



After normalization (dividing by L2 norm of the document):

|  |  |  |  |
| --- | --- | --- | --- |
| **Norm’d Weights** | Doc 1 | Doc 2 | Doc 3 |
| car | 0.67 | 0.37 | 0.60 |
| auto | 0.51 | 0.73 | 0 |
| insurance | 0 | 0.57 | 0.61 |
| best | 0.53 | 0 | 0.51 |

# Problem 6

1. The idf of a term that occurs in every document will be log (N/N) = 0. Thus, since stop words tend to occur in every document, they will be ignored automatically during ranking because their tf-idf weight will be 0.
2. Changing the base of the log means dividing the score of a document by logB(b), where “b” is the old base, and “B” is the new base. This value is a constant and, thus, it does not impact the ranking of the documents.