جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

INFORMATION RETRIEVAL WITH APPLICATIONS OF NLP
ENCS4130

Assignment #3

Prepared by: Mazen Batrawi - 1190102

Instructor: Dr. Adnan Yahya

Section: 1

BIRZEIT

May – 2023

1.

a) Zipf's law states that the frequency of a word, ranked as the nth most frequent, can be expressed as the maximum frequency (fmax) divided by the value of n, if we adhere strictly to this law, we get the following:

f(a) + f(b) + f(c) + f(d) + f(e) = 25000
`$\frac{f}{1}+\frac{f}{2}+\frac{f}{3}+\frac{f}{4}+\frac{f}{5}=25000$
Multiplying both sides with 120
120f + 60f + 40f + 30f + 24f = 3000000
274f = 3000000
f = 10949
f(a) = 10949, f(b) = 5475, f(c) = 3650, f(d) = 2737, f(e) = 2190
Adding up these values, we get 25000.

b) For the non-positional postings list, we can consider a worst-case scenario where each word appears once in a document. Let's take the example of the word "a" and assume there are 10,949 documents in which "a" appears, with each document containing the word "a" exactly once. If we assume that the size of each posting entry requires 4 bytes, we can estimate the size of the postings list as follows:

a = 10949 * 4 = 37.8 Kbytes
b = 5475 * 4 = 21.9 Kbytes
c = 3650 * 4 = 14.6 Kbytes
d = 2737 * 4 = 11 Kbytes
e = 2190 * 4 = 8.8 Kbytes
The total is 94 Kbytes

Since more details are supplied regarding the precise placement of the word in the page, the positional postings list would be larger. Other details, such as the average number of occurrences per word and the average number of spots saved each occurrence, are required to estimate the size of the postings list.

2.

a) Number of bits = log2(600,000,000) = 30 bits.

number of full bytes = 30 / 8 = 4 bytes.

b) string length = voc. size * avg. word length = 500,000 * 8 = 4,000,000 characters.

The number of bits = log2(string length) = log2(4,000,000) = 22 bits.

The number of bytes = 22 / 8 = 3 bytes.

Pointer size (22 bits) = 22 * 500,000 = 11,000,000 bits = 1375 Kbytes

Pointer size (3 bytes) = 3 * 500,000 = 1,500,000 bytes = 1500 Kbytes.

c) size = #terms * #doc = 500,000 * 600,000,000 = 3 * 10^{14}

The estimation of the number of nonzero elements in the incidence matrix is based on the worst-case scenario. Assuming an average of 200 words per document, in the worst-case scenario, all these words would be unique and therefore contribute a nonzero value in the incidence matrix.

The number of nonzero elements can be calculated as follows: 200 words per document multiplied by 600,000,000 documents, resulting in a total of 1.2×10^{11} nonzero elements.

To estimate the size of the non-positional postings list, we once again consider the worst-case scenario. In this case, we assume that all words are unique across all documents.

If we assume that each posting entry requires 4 bytes to be represented, the total size of the postings list can be calculated as follows: 200 words per document multiplied by 600,000,000 documents, and then multiplied by 4 bytes. This results in a total postings list size of 4.8×10^{11} bytes.

3. a)

1	0
3	101
4	11000
39	11111000111
63	11111011111
127	1111110111111
1023	1111111110111111111
4095	11111111111011111111111

The sequence will be:

0101110001111100011111111011111111111101111111111111110111111111111111111110111 1

1111111

b) 1190102 = 100100010100011010110

removing the highest bit → 00100010100011010110

The length = 20

Gamma code for my id: 1111111111111111111100010001010001101011 0

c)

1110111111101010111110101011101011111111011101010

1110111 = 15

111101010 = 26

11111010101 = 53

11010 = 6

11111111011101010 = 490

The sequence is: 15, 26, 53, 6, 490.