Assignment 3 [I got 36 / 40]

Joey Tawadrous - 109528652

Problem 1 [5 points]
Write down VB code and Gamma code for number 777.

777 in binary code:

777 / 2
388 / 2 1
194 / 2 0
97 / 2 0
48 / 2 1
24 / 2 0
12 / 2 0
6 / 2 0
3 / 2 0
1 / 2 1
0 1
= 1100001001

VB code:

- 00000110 10001001
- (start to fill 7 bytes if you did not finish put 1 on the 8th bit else 0)
- if less than 8 bytes, just put a 1 at the start i.e.
    - VB of 12
    - binary = 1100
    - VB = 10001100

Gamma:

Firstly

- offset is the binary, with the leading bit chopped off = 100001001

Secondly

We need unary code to get gamma code length.

- Next get length of offset (100001001) = 9
- Represent length as n 1s with a final 0.
- Unary code for 9 is 1111111110

Finally

- Gamma code of 777 is the concatenation of length and offset = 1111111110100001001

Problem 2 [7 points] (Example on slides T3 slide 55)

• Looking at a collection of web pages, you find that there are 7000 different terms in the first 20,000 tokens and 36,000 different terms in the first 1,500,000 tokens.

• Assume a search engine indexes a total of 30,000,000,000 (3 × 1010) pages, containing 200 tokens on average

• What is the size of the vocabulary of the indexed collection as predicted by Heaps' law (log based 10 and round result to integer)?

log(M1) = logk + blog(T1)
log(7000) = logk + blog(20,000)

log(M2) = logk + blog(T2)
log(36,000) = logk + blog(1,500,000)
(above is b(0.5) * log(1,500,000))

thus logk = log(7000) - 2.15 = 1.6945, k = 29.99 and b = 0.5
(above 2.15 is log(20,000) * 0.5)

log(M) = logk + ½ log(30,000,000,000 * 200) = 7.866
(above logk(29.99) +)

thus M = 10^7866 = 7*10^7
(above ^ = to the power of)

Problem 3 [7 points]
What is the Levenshtein distance between the following pairs of strings? "thorough and "throughout, write down your calculation steps.

Answer attached on separate page.

Problem 4 [5 points]
Define the terms recall and precision.

Precision is the fraction of retrieved documents that are relevant to the query.

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

Problem 5 [5 points]
The F-measure is defined as the harmonic mean of recall and precision. What is the advantage of using the harmonic mean when compared to the arithmetic mean?

Harmonic mean closer to average.

- If a search engine returns all documents in the collection, recall will be 1 and precision will be close to 0. This gives us an arithmetic mean evaluation of 0.5, which is too high. This strongly suggests that the arithmetic mean is an unsuitable measure to use.
- The harmonic mean on the other hand, will punish poor performance in either precision or recall, which gives more accurate evaluations. In the 'return everything' case, an evaluation of 0 will be returned for the harmonic mean.
- Extra: The harmonic mean is always less than or equal to the arithmetic mean. When the values of two numbers differ greatly, the harmonic mean is closer to their minimum than the arithmetic mean.

Problem 6 [6 points]
Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 → 2, 3 → 2, 2 → 1, 2 → 3. Write down the transition probability matrices for the surfers walk with teleporting, for the following three values of the teleport probability: (a) α = 0; (b) α = 0.5 and (c) α = 1.

Link Matrix

	1	2	3
1	0	1	0
2	1	0	1
3	0	1	0

a) a = 0

	1	2	3
1	0	1	0
2	0.5	0	0.5
3	0	1	0

b) a = 0.5
0.5 / 3 = 1/6

	1	2	3

1	1/6	2/3	1/6
2	5/12	1/6	5/12
3	1/6	2/3	1/6

c) a = 1
1 / 3 = 1/3

	1	2	3
1	1/3	1/3	1/3
2	1/3	1/3	1/3
3	1/3	1/3	1/3

Problem 7 [5 points]
Show that the PageRank of every page is at least α/N . What does this imply about the difference in PageRank values (over the various pages) as α (teleporting probability) becomes close to 1?

Quick Answer

- p(i -> j) >= (a / n) > 0
- This is for all 1< i, j <= n where the value of n is the number of nodes / webpages.

Longer Answer

$$\vec{x_i} = \sum_{j=1}^{N} (\vec{x_j} P_{ji}) \geq \sum_{j=1}^{N} (\vec{x_j} \alpha/\mathrm{N}) = \left(\frac{\alpha}{\mathrm{N}}\right) \sum_{j=1}^{N} \vec{x_j} = \alpha/\mathrm{N}$$

Final Part Answer

- As α becomes closer to 1, the impact of the web graph link structure gets smaller. As a result, the difference in PageRank values over various pages will also get smaller. Thus the pagerank will begin to lose its ability to differentiate between pages.