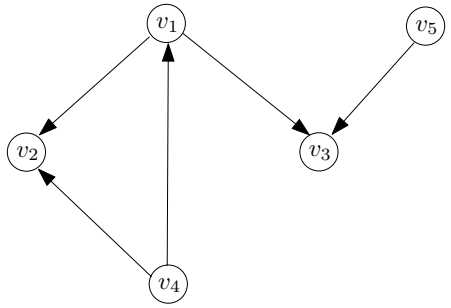
WST540: Exercise List 3

Problem 1. Consider the following graph:



Each node represents a website and each edge represents a hyperlink. Suppose that a web crawler initially knows only v_4. Using the BFS algorithm discussed in Lecture 4, which websites will be discovered by the crawler when it finishes?

Solution. The BFS algorithm visits v_4, v_2, v_1 and then v_3. It is not able to discover v_5.

Problem 2. Suppose that we have the following document collection:

document ID	content
1	the old night keeper keeps the keep in the town
2	in the big old gown in the big old house
3	the house in the town had the big old keep
4	where the old night keeper never did sleep
5	the night keeper keeps the keep in the night
6	and keeps in the dark and sleeps in the light

Also, consider that our dictionary has words {big, dark, gown, house, keep, light, night, old, sleep, town}. Assume that, after stemming, we have the following equivalence:

$$\begin{aligned} \text{keep} &= \text{keeper, keeps} \\ \text{sleeps} &= \text{sleep} \end{aligned}$$

Give the document-level inverted lists of all the words in the dictionary. Each entry of an inverted list should have the format (doc id, term frequency).

Solution.

term w	inverted list for w
big	(2, 2), (3, 1)
dark	(6, 1)
gown	(2, 1)
house	(2, 1), (3, 1)
keep	(1, 3), (3, 1), (4, 1), (5, 3), (6, 1)
light	(6, 1)
night	(1, 1), (4, 1), (5, 2)
old	(1, 1), (2, 2), (3, 1), (4, 1)
sleep	(4, 1), (6, 1)
town	(1, 1), (3, 1)

Problem 3. Give the Eilas' gamma and delta codes of 23.

Solution. Gamma code: 111100111. Delta code: 110010111.

Problem 4. Consider the following inverted list, where each entry is in the format of (doc id, term freq):

(1, 1), (4, 1), (5, 2)

Give the bit sequence that compresses the above list based on the following ideas:

- For the i-th ($i \geq 2$) pair, represent its id by storing in Elias' gamma code the difference from the id of the $(i-1)$-th pair.
- Store each term-frequency value in Elias' gamma code.

Solution. We will store the following sequence of values: 1, 1, 3, 1, 1, 2. The bit sequence is: 0010100100.

Problem 5. Give the word-level inverted lists for Problem 3. Each entry of an inverted list should have the format (doc id, term frequency, position 1, position 2, ...).

Solution.

term w	inverted list for w
big	(2, 2, 3, 8), (3, 1, 8)
dark	(6, 1, 5)
gown	(2, 1, 5)
house	(2, 1, 10), (3, 1, 2)
keep	(1, 3, 4, 5, 7), (3, 1, 10), (4, 1, 5), (5, 3, 3, 4, 6), (6, 1, 2)
light	(6, 1, 10)
night	(1, 1, 3), (4, 1, 4), (5, 2, 2, 9)
old	(1, 1, 2), (2, 2, 4, 8), (3, 1, 9), (4, 1, 3)
sleep	(4, 1, 8), (6, 1, 7)
town	(1, 1, 10), (3, 1, 5)