Assignment #3

Assignment #3

Due: April 07, 2025

1. (20) Produce a stylesheet to transform the document shown in Fig. 1(a) to a document shown in Fig. 1(b).

```
<? Xml version = “1.0” encoding = “utf-8” standalone = “yes” ?>
<Movies>
        <Movie title = “King Kong”>
                <Version year = “1993”>
                        <Star>Fay Wray</Star>
                </Version>
                <Version year = “1976”>
                        <Star>Jeff Brideges</Star>
                        <Star>Jessica Lange</Star>
                <version year = “2005” />
        </Movie>
        <Movie title = “Footloose”>
                <Version year = “1984”>
                        <Star>Kevin Bacon</Star>
                        <Star>John Lithgow</Star>
                        <Star>Sarah Jessica Parkr</Star>
                </Version>
        </Movie>
</Movies>
```

```
<? Xml version = “1.0” encoding = “utf-8” standalone = “yes” ?>
<Movies>
        <Movie title = “King Kong”>
                <Star>Fay Wray</Star>
                <Star>Jeff Brideges</Star>
                <Star>Jessica Lange</Star>
        </Movie>
        <Movie title = “Footloose”>
                <Star>Kevin Bacon</Star>
                <Star>John Lithgow</Star>
                <Star>Sarah Jessica Parkr</Star>
        </Movie>
</Movies>
```

(b)

Fig. 1

(a)

Assignment #3

2. (15) In Fig. 2, we show a network, in which each node stands for a page and each arc for a link from a page to another. Please give the transition matrix for the network. Also, explain why the solution to the equation:

$A = \boldsymbol{M}A$ can be used as the estimation of page importance, where A is a vector of n variables and $\boldsymbol{M}$ is an $n \times n$ transition matrix.

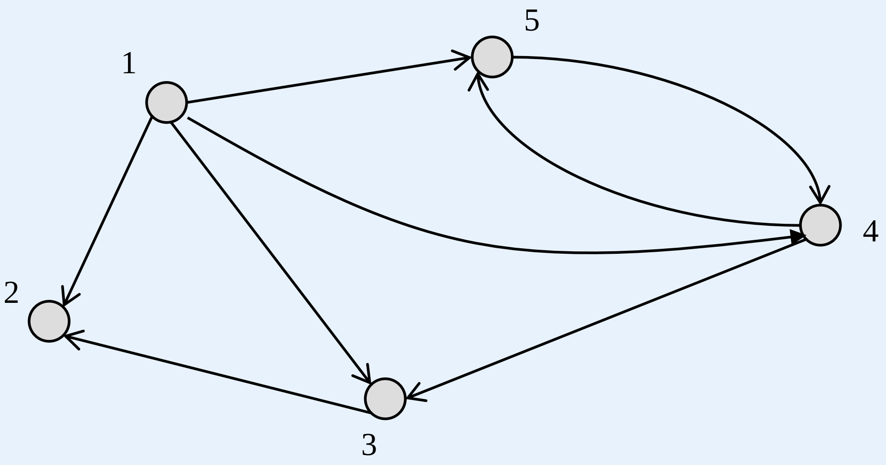

Fig. 2

Assignment #3

3. (10) Explain why the following equation (for estimate the importance of pages) works in the presence of spider traps and dead ends.

$$\mathbf{P}_{new} = \beta\mathbf{M}\mathbf{P}_{old} + (1 - \beta)\mathbf{T}$$

4. (20) Fig. 3 shows a tree encoding. The quadruples can be stored as a sequence sorted by LeftPos values by using the depth-first search. Design an algorithm to transform it into another sequence sorted by RightPos values.

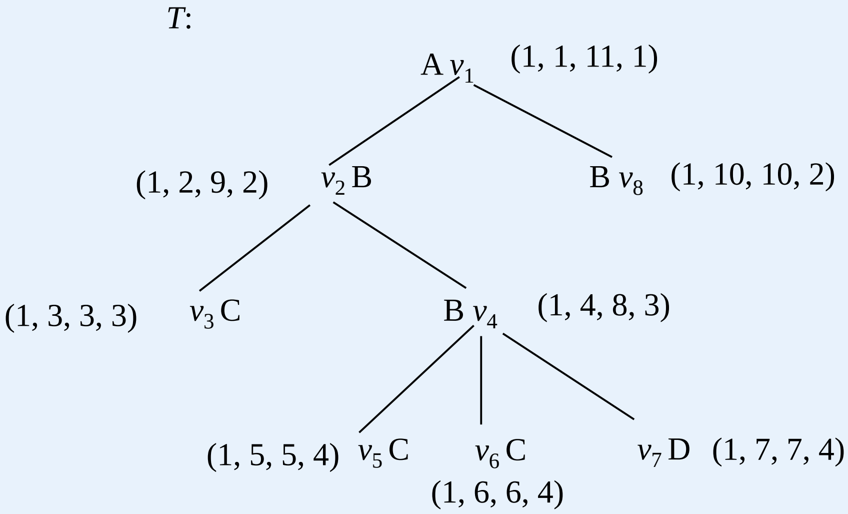

Fig. 3

Assignment #3

5. (15) In the following table, we show the key words of five documents, as well as the key word sequences sorted by frequencies. Please construct a trie for the sorted sequences and a header table for all the key words to speed up the evaluation of conjunctive queries of form word1 $\wedge$ word2 $\wedge$ … $\wedge$ word*i*. Also, show how a conjunctive query is evaluated by using the trie.

DocID		Items	Sorted item sequence
1		*f, a, c, i, j, m, p*	*c, f, i, a, m, p, j*
2		*a, b, c, h, f*	*c, f, a, b, h*
3		*b, i, f*	*f, b, i*
4		*b, c, i*	*c, b, i*
5		*a, f, c, m, p*	*c, f, a, m, p*

Fig. 4

Assignment #3

6. (20) Consider a bit data stream. It will be stored as a series of pairs of the form (*m*, *t*), where *m* is the number of bits in a bucket *B*, and *t* is the time when the most recent bit in *B* is received. A bucket is a segment in a bit data stream satisfy following conditions:

1. The size of every bucket is a power of 2 (2^i for some *i*).
2. As we look back in time, the sizes of the buckets never decrease.
3. For *m* = 1, 2, 4, 8, … up to some largest-size bucket, there are one or two buckets of each size, never zero and never more than two.
4. Each bucket begins somewhere within the current window, although (largest) bucket may be outside of the window.

Please briefly describe the method to form the buckets when receiving bits one by one.