CS251 Homework 5: Strings (II)

Out: April 13, 2018 @ 9:00 PM
Due: April 27, 2018 @ 11:59 PM

Submission Instructions: Please submit a typeset PDF on blackboard. For multiple choice questions, you must provide an explanation along with your answer. Answers without explanations will receive 0 points, even if correct.

1. Finding patterns in DNA sequences is a common task in bioinformatics. A DNA sequence is composed of characters A, C, G and T representing adenine, cytosine, guanine and thymine, respectively. The KMP algorithm is used when the text and the pattern are not too long. Before running KMP, we must calculate the failure function of the pattern. We need to find the pattern "GACAGATGA" in a DNA sequence. Calculate the failure function for the given pattern. [2 points]

   F(G) = 0
   F(GA) = 0
   F(GAC) = 0
   F(GACA) = 0
   F(GACAG) = 1
   F(GACAGA) = 2
   F(GACAGAT) = 0
   F(GACAGATG) = 1
   F(GACAGATGA) = 2

   000012012

   GACAGATGA

2. Following the KMP algorithm, what is the number of comparisons required for finding the pattern "GACAGATGA" in "GGTACCCGACAGATGACAGA"? Help yourself with a drawing similar to the examples showed in class and paste it with your answer. [3 points]

GGTACCCGACAGATGACAGA
12
GACAGATGA
3
GACAGATGA
4
GACAGATGA
5
GACAGATGA
6
GACAGATGA
7
GACAGATGA
8

GACAGATGA
9-17
GACAGATGA

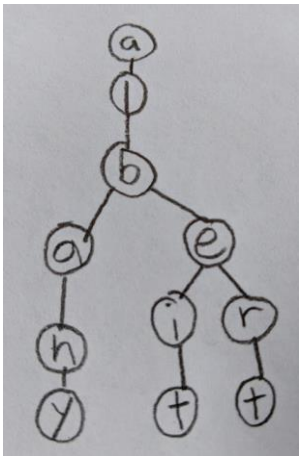
There are 17 comparisons required to find a match.

3. Consider a standard trie constructed from a chemistry book of 500 pages, each page having 2000 words (on average). Word lengths range from 1 letter to 1909 (There are long chemical names, and sometimes the author forgets to put hyphen to separate the groups. If you don't believe me, check this: https://en.wikipedia.org/wiki/Longest_word_in_English). We are interested in searching English words (so, our alphabet includes standard English alphabet, ignoring case, and numbers 0-9, totaling 36).
   We want to search the word "methylhydroxide". What will be the most number of comparisons to search the word? Show your work. [2 points]
   a. 68724
   b. 15
   c. 36
   d. 540

   The answer is B. This is because to search a string of length n in a trie, the maximum depth for the comparison to be made is n. Therefore in the search for methylhydroxide, the maximum comparisons would be equal to its length which is 15.

4. How many tree nodes are there in a standard trie constructed from the following three words: albert, albany, albeit? Show your work by drawing the tree. [2 points]
   a. 7
   b. 9
   c. 11
   d. 12

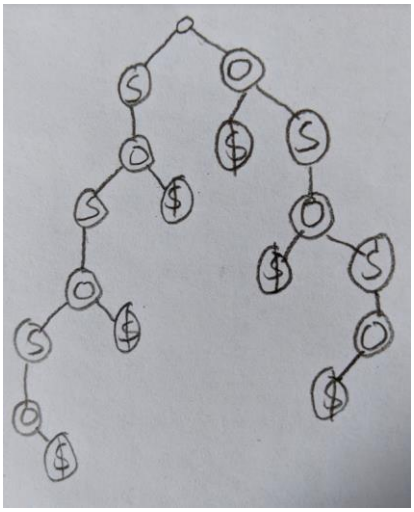
   The answer is C.



5. A suffix trie is constructed from the word "sososo". What will be the number of nodes in the trie (in uncompressed form)? Show your work by drawing the tree. [3 points]
   a. 10
   b. 14

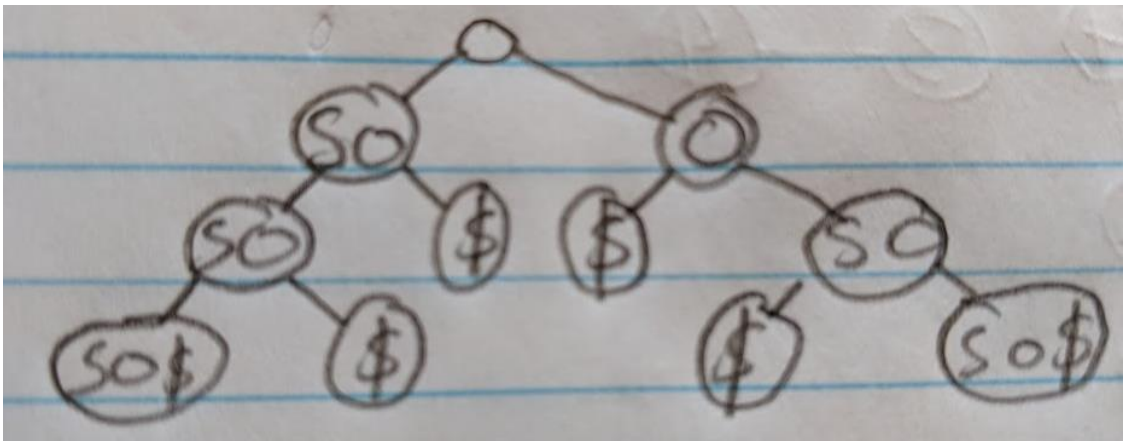
c. 18
d. 22

The answer is C.



6. Consider the same suffix trie from the previous problem, but in compressed form. What will be the number of nodes in the compressed suffix trie? Show your work by drawing the tree. [3 points]
   a. 9
   b. 10
   c. 11
   d. 12

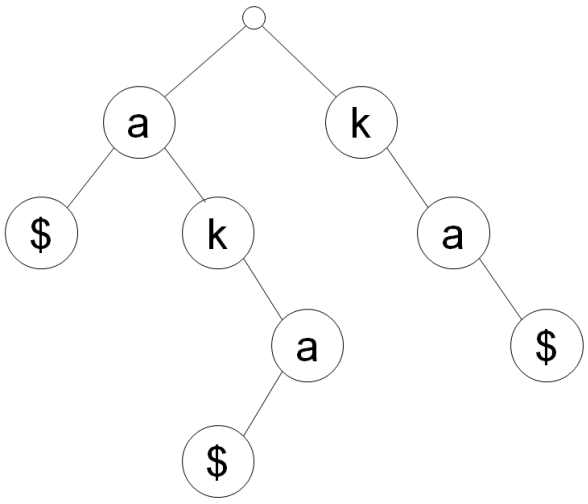
   The answer is C.



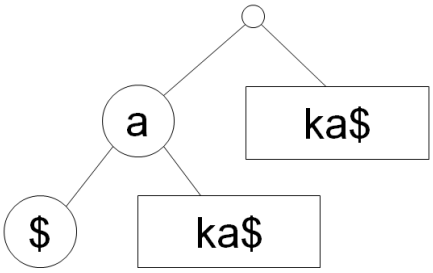
Hint for 5 and 6: For the example string “aka”, the suffix strings constructed will be:

1. aka$
2. ka$
3. a$

where $ is a special character used to denote the end of the string.
Then the uncompressed and compressed suffix trie will be the following:



[uncompressed suffix trie]



[compressed suffix trie]

7. Which uncompressed string is *the most inefficient* (in terms of the compression ratio) to be encoded using Run-Length Encoding? Which one is the *most efficient*? Show your work by providing Run-Length Encoding of each string. [2 points]
   a. AAAAAAAbbbXXXXXt
   b. AAAABBBAAACCC
   c. CGTACGTA
   d. CCGGTTAA

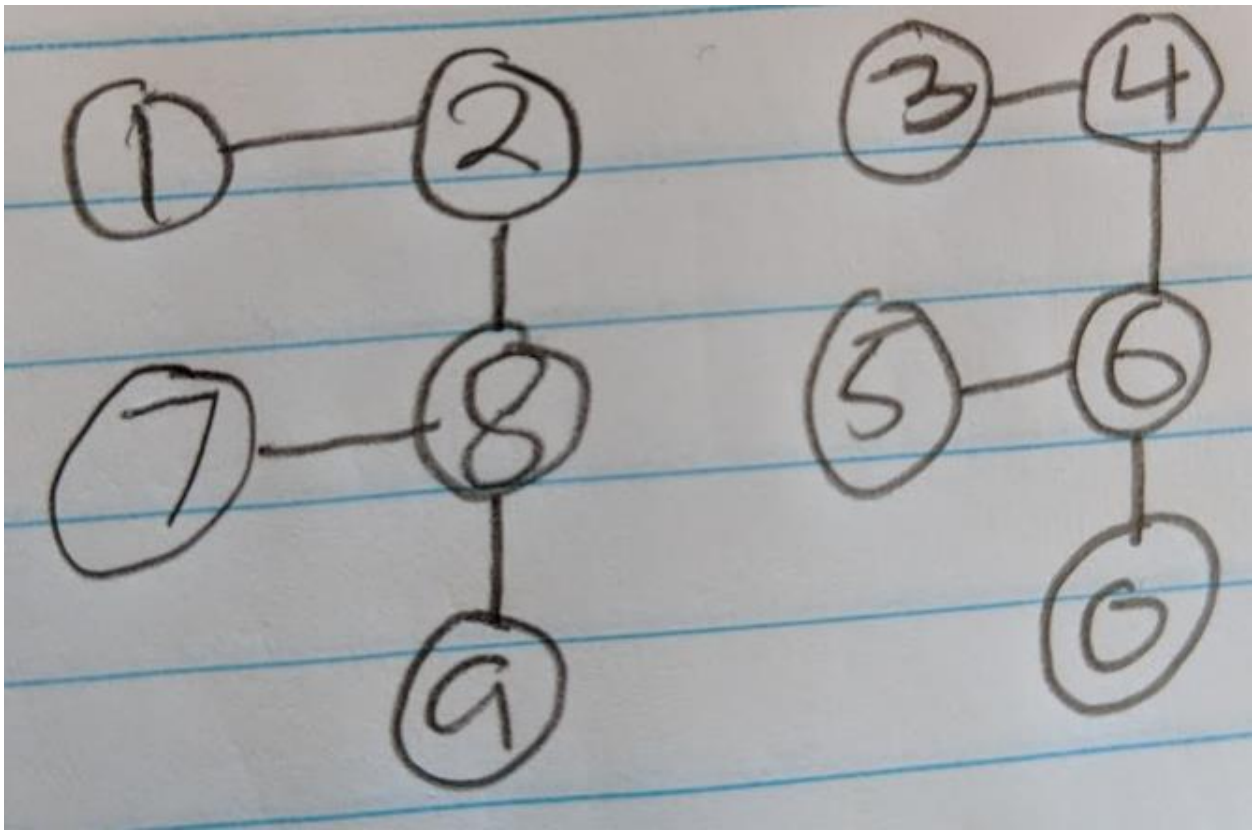
   A is 5A3b5X1t. B is 4A3B3A3C. C is 2 CGTA. D is 2C2G2T2A. D is the most inefficient because before encoding it has a length of 8 and after encoding it still has a length of 8. A is the most efficient because of how much it shortens it after encoding.

8. How many connected components do we have after performing the following sequence of union operations on a set of 10 items? Show your work by drawing the tree. [2 points]

```
union(1,2), union(3,4), union(5,6), union(7,8), union(8,9), union(2,8),
union(0,6), union(4,6)
```

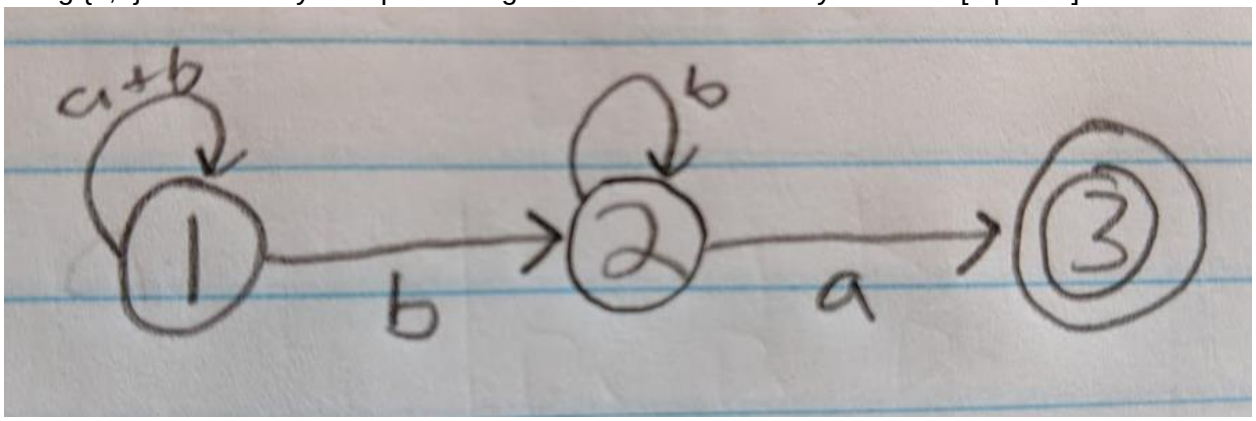
   a. 1
   b. 2
   c. 3
   d. 4

   The answer is B.



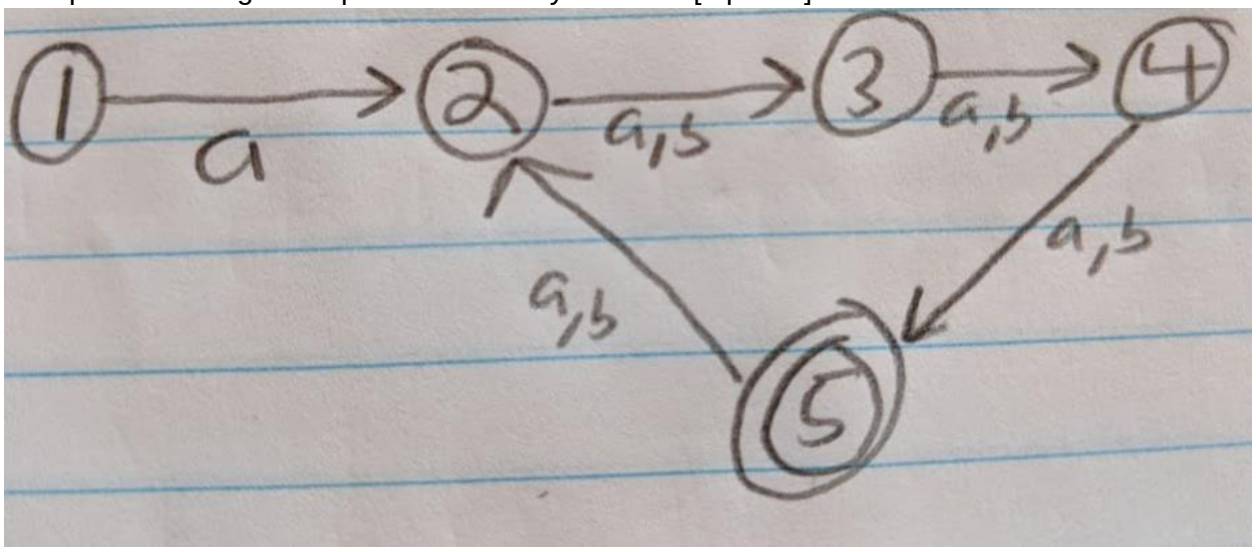
9. Suppose we are working with a Union-Find data structure as described in the lecture slides. Suppose we consider a set of 10 items that are eventually joined into a single set via 9 union operations (assume that no path compression is applied). Can you have a sequence of union operations that will result in the root having three children? Explain why or why not. [3 points]
   You can have a sequence where there will be three children. If there is a union-by-weight then you can merge tree with a smaller weight into a tree with a larger weight and can cause the root to have three children.

10. Construct a regular expression and a deterministic finite automaton that accepts all strings using {a,b} where every accepted string ends with "ba" . Show your work. [2 points]



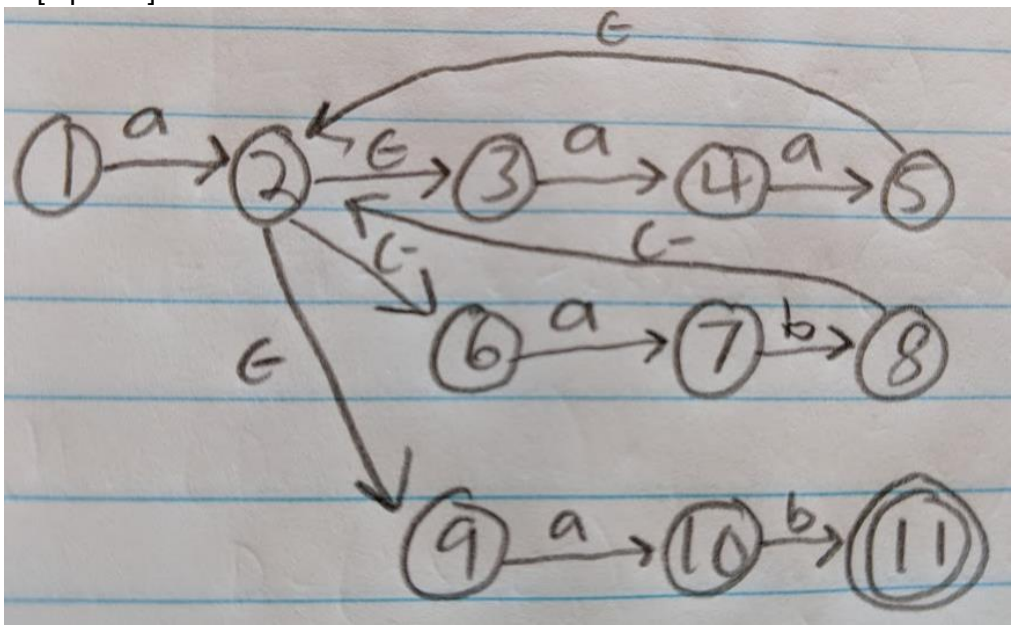
State 1 can accept any number of a's or b's or null. State 2 accepts at least one b and any number of b's. State 3 accepts only 1 a.

11. Construct a deterministic finite automaton that accepts all non empty strings using {a,b} where every accepted string starts with "a" and the length of the string is a multiple of 4. Also provide its regular expression. Show your work. [3 points]



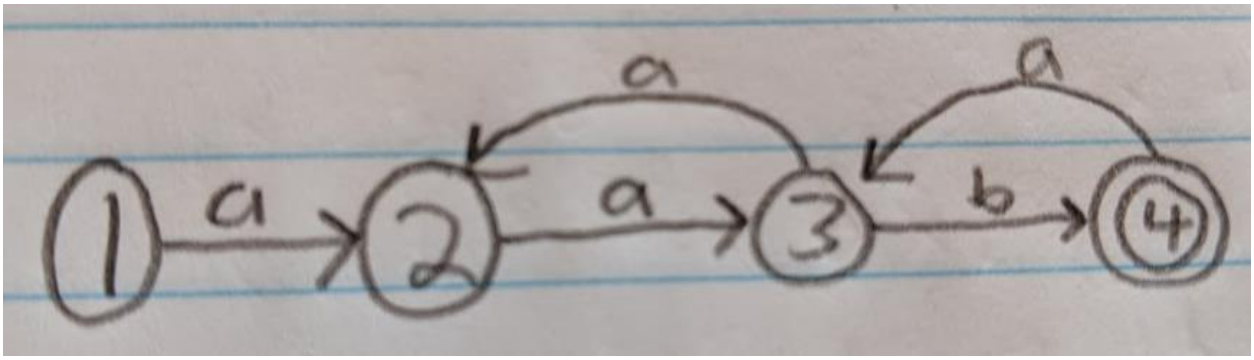
State 1-2 is for the start a. State 2 is where all strings have a remainder of 1. State 3 and 4 are the same as state 2. State 5 is where it will accept it if it is a multiple of 4.

12. a) Construct an Epsilon NFA for the regular expression `a(aa|ab)*ab`. Show your work. [2 points]



States 3-4-5 is for aa. States 6-7-8 are for ab. States 9-10-11 are for ab.

b) Construct a DFA for the NFA in (a). Show your work. [4 points]



When it reaches State 4 and an a comes it will go back to 3 then back to 2. If ab comes then State 4 will accept it.

Total: 33 points.