

Question 1:

Given a parse tree in which each non-terminal node contains a syntactic category, and each leaf contains a word. All the words in the leaves of the sub-tree of each non-terminal node (except for the root) are considered to be a phrase. Your goal is to print all the phrases contained in the tree, and for each phrase its syntactic category.

- Describe how the data structure to store the tree in the computer.
- Write pseudo code for an *efficient* algorithm to print all the phrases and their categories.
- What is the computation time of your algorithm?

Example:

For the tree below your program should print:

NP – I

NP – him

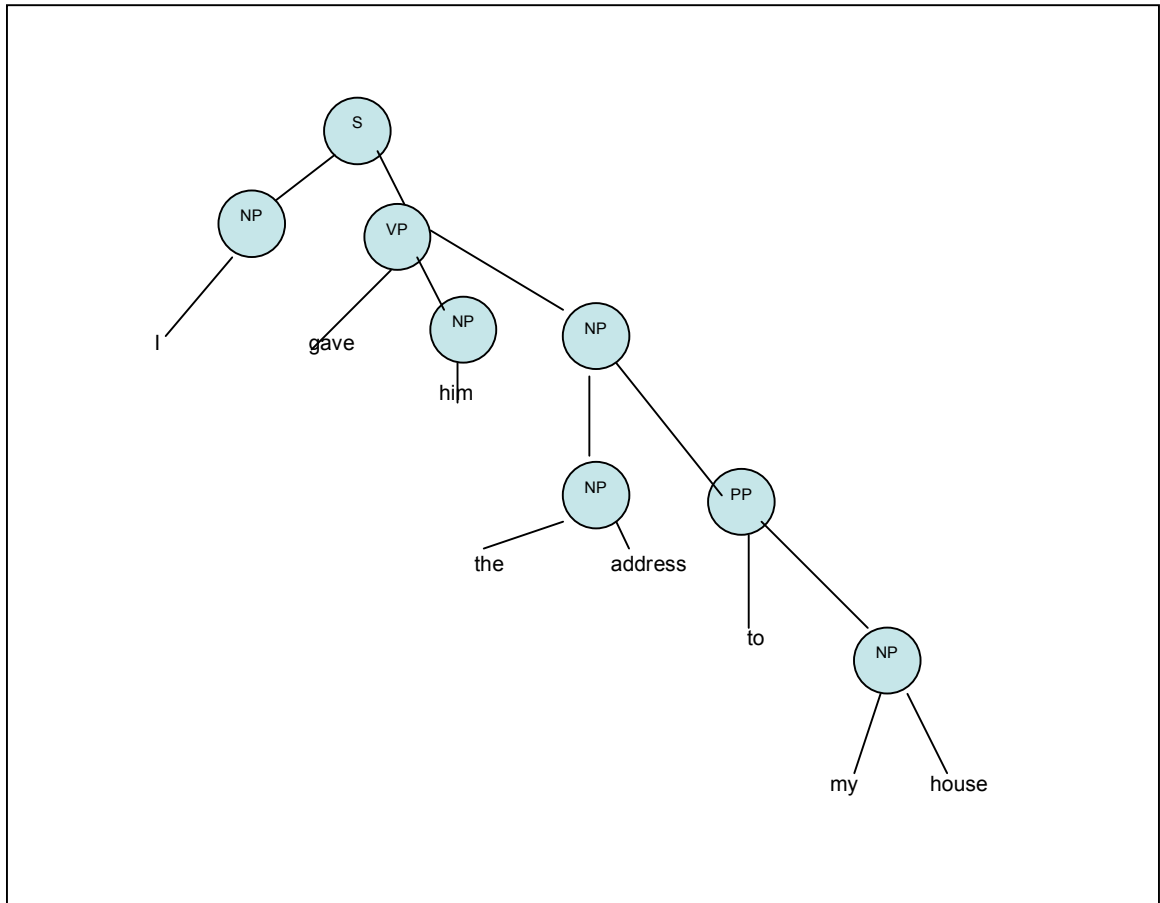
NP – the address

NP – my house

PP – to my house

NP the address to my house

VP gave him the address to my house



### Question 2

A list with  $n$  elements has a majority element if more than half of its elements are the same. Your goal is to find the majority element, or determine that the list does not contain a majority element. The elements in the array are not necessarily from some ordered domain like the integers, and so there can be no comparison of the form “is element  $i >$  element  $j$ ” (The elements may be GIF files). However, you can answer the question “is element  $i =$  element  $j$ ” in constant time.

Describe an algorithm to solve the problem and analyze the computation time. Your algorithm should have a faster run time than  $O(n^2)$  (There is an  $O(n)$  algorithm).

### Question 3

You are given a string of  $n$  characters in which all white space has vanished (for example “itwasthebestoftimes...”). You wish to reconstruct the document using a dictionary.

- a. Describe a dynamic programming algorithm that determines whether the string can be reconstituted as a sequence of valid words. The running time should be at most  $O(d*n^2)$  assuming that a call to the dictionary takes time  $O(d)$ .
- b. Give pseudo code for a function that outputs the valid sequence of words.

### Question 4

Suppose you are performing a chemistry experiment; you have billions of molecules formed from a chain of 5 smaller parts (labeled A, B, C, D, and E). Each part has a unique weight. While you know what the parts are, you don't know their order: it might be ABCDE, ACDEB, EABDC, or any other configuration.

You have access to equipment that will break each large molecule into two pieces, and measure the weight of the individual pieces accurately. To be clear: if the order of the parts is ABCDE, then the equipment will generate

A BCDE  
AB CDE  
ABC DE  
ABCD E

There are billions of these molecules, and each will be broken into only two subcomponents. Your equipment (known as a mass spectrometer) will return a histogram that marks the weights of the subcomponents observed.

Explain a strategy to determine the order of the parts in the molecule. What role do the weights of the individual parts play? Can your strategy scale to handle larger molecules (with hundreds of parts, rather than just 5)?