Introduction to Programming Homework 6 Solutions

Exercise 1 (Generators)

Make a module called `generators.py`. You will implement the following three generators.

- **a.** (Binary Decimal Expansion) Given a ratio m/n of positive integers with $m \leq n$, we can consider the binary expansion

$$\frac{m}{n} = \sum_{i=1}^{\infty} \frac{a_i}{2^i}.$$

Write a generator function `binary_decimal(m,n)` which yields the sequence a_i in order. For example,

```
import generators as gen
one_third = gen.binary_decimal(1,3)
for _ in range(5) :
  print(next(one_third))
```

should print

```
0
1
0
1
0
```

Be sure to raise an error on bad input.

In [ ]:

```
import random

def binary_decimal(m,n) :
    """ Given integers 0 <= m <= n, n != 0, yields
    the (shortest) binary decimal expansion of m/n """
    assert ( isinstance(m, int) and isinstance(n, int) and
             0 <= m <= n and n != 0 )
    rest = m
    while rest != 0 :
        rest *= 2
        coeff = rest // n
        yield coeff
        rest -= coeff * n
```

- **b.** (Biased coin from a fair one) A common problem in computer science is to take one pseudo-random number generator and use it to construct another. We can model a fair coin flip by using `random.randint(0, 1)` form the `random` module. This returns a `0` or a `1` with equal probability. **Using this fair coin**, write a function `biased_coin(m,n)` that returns `1` with probability exactly `m/n`.
  - Hint : Make use of part **a**.

In [ ]:

```
def biased_coin(m,n) :
    """ Using random.randint(0, 1) as a fair coin, this
    returns 1 with probablity m/n, where m,n must be
    non-negative integers with m <= n. """
    # binary_decimal checks m,n for us
    bin_dec = binary_decimal(m,n)
    for x in bin_dec :
        if x != random.randint(0, 1) :
            return x
    return 0 # if binary_decimal is exhausted
```

- **c.** (Shuffle generator) Write a generator `shuffle_gen(start, stop)` that returns random numbers between `start` and `stop` without repetition. Make sure this works for very large integers.
  - Hint: See the modern version of the Fisher-Yates shuffle at https://en.wikipedia.org/wiki/Fisher–Yates_shuffle#Modern_method but instead of *swapping* numbers into a list you should use a better data structure.

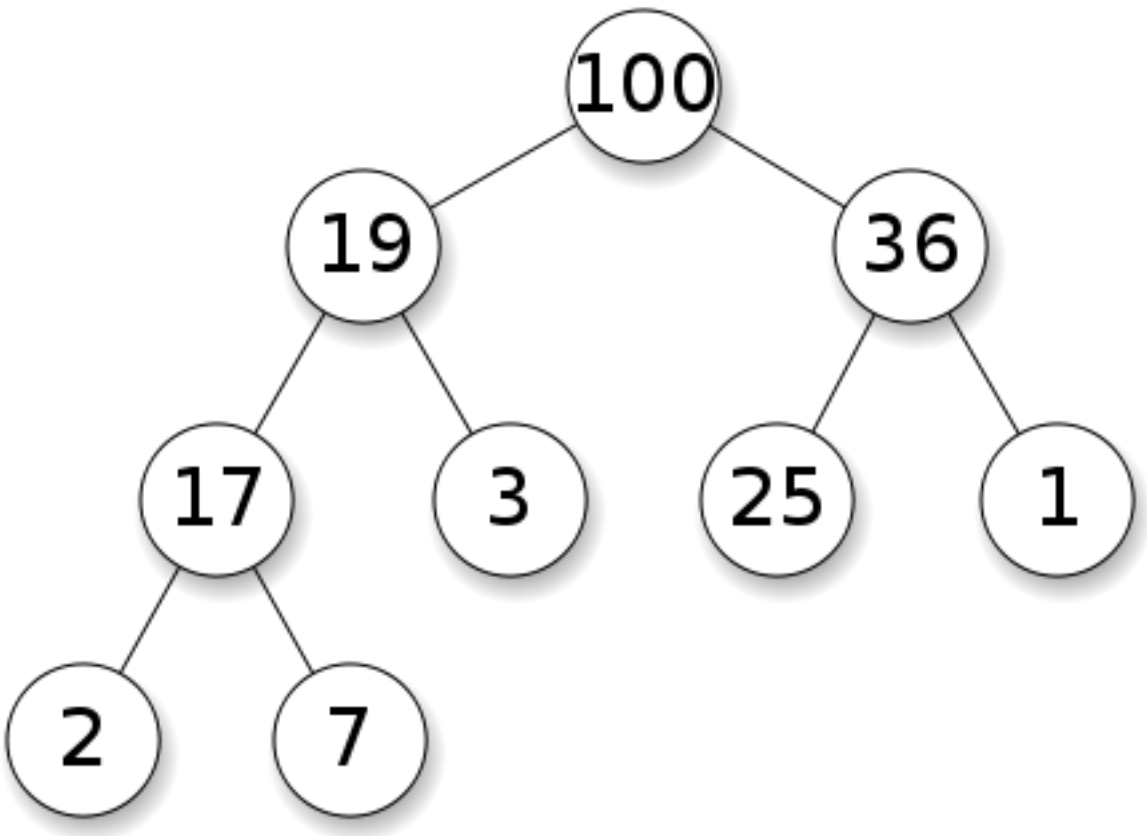
In [ ]:

```
def shuffle_gen(start, stop) :
    """ A generator that yields a random permutation
    of the numbers start, start + 1, ..., stop -1. """
    n = stop - start
    scratch = dict()
    for remaining in range(n, 0, -1):
        i = random.randrange(remaining)
        # recall that scratch.get(i,i) = scratch[i]
        # if it is defined and scratch.get(i,i) = i otherwise
        yield scratch.get(i,i) + start
        scratch[i] = scratch.get(remaining - 1, remaining - 1)
        # we don't need to keep track of this anymore
        # note : we use `.pop` instead of `del` because
        # scratch[remaining - 1] may not be set
        scratch.pop(remaining - 1, None)
```

Exercise 2 (Heap)

Create a module called `heap.py`.

A binary **heap** is a binary tree with the property that for every node `x` one has `x.data >= x.left.data` and `x.data >= x.right.data`. For example, this image represents a heap



In this exercise, we will create a `Heap` class that will manage a tree with the heap property. **Instead** of using the `Node` class from the previous homework, the `Heap` will **store all data in a `list` in level-order** (i.e. we read off each level of the tree from left to right). For example, the data in the tree above will stored as the list

```
data_list = [100, 19, 36, 17, 3, 25, 1, 2, 7]
```

You can see that the node at index `i` of `data_list` has the left child at index `2*i+1` and the right child at index `2*i+2`. Similarly, you can use the index in the list to find a node's parent.

You will notice that with this structure, the maximum of the data in the tree is always the first element of the list. Our goal will be to write methods that allow us to insert new data into the heap and to remove the largest value from the heap all while preserving the heap property.

- **a.** Create a class called `Heap`. Inside, you will store a private `self._data_list` attribute. The init method should simply have type `__init__(self)` where you initialize the data list. In part **b**, we will be inserting elements to the end of `self._data_list` and then using the following methods to force the heap property using the following methods :
  - `def _bubble_down(self, i)`:
    - start with the node at index `i`
    - if the data at index `i` is larger than that of its children, stop.
    - if not, switch the data at index `i` with either child that has greater value.
    - continue doing this process to the child until you stop.
  - `def _bubble_up(self, i)`:
    - start with the node at index `i`
    - if the data at index `i` is smaller than that of its parent, stop.
    - if not, switch the data at index `i` with that of its parent.
    - continue doing this process to the parent until you stop.

Do not use recursion for the above methods!

- **b.** We will now us the `_bubble_down` and `_bubble_up` methods to insert elements into the heap and delete the largest value. Implement the following methods :
  - `def insert(self, value)`:
    - add `value` at the end of `self._data_list`
    - run `self._bubble_up` to move the newly inserted element into a position that doesn't violate the heap property.
  - `def pop_max(self)`:
    - record the first value of `self._data_list`
    - replace the first value of `self._data_list` with the last value
    - use `self._bubble_down` to push this replaced value down until it doesn't violate the heap property.
    - return the recorded first value as the maximum in the heap.

Observe that you can now sort a list of data by inserting elements one by one into a heap and then using `.pop_max()` to read off the sorted list from largest to smallest.

- **c.** Create a global function called `sort_all_lines(path_to_file)` which will read a file at `path_to_file`. Each line of `path_to_file` will be a **comma separated list** of integers. **Use the heap class you built above** to sort each line from greatest to smallest. Write you data to a file named as follows : if `path_to_file = 'dir1/dir2/filename.extension` save your file to `'dir1/dir2/filename_sorted.extension'`. For example, if file `'test_data.csv'` contains

  ```
  4,6,1,47,241,352,7,0,-140,352
  35,2601,362,1350305,-9352,38351
  ```

  you code should write a file called `test_data_sorted.csv` as

  ```
  352, 352, 241, 47, 7, 6, 4, 1, 0, -140
  1350305, 38351, 2601, 362, 35, -9352
  ```

In [ ]:

```
def is_real_num(x) :
    """ Returns True if x is an int, bool, or float. """
    return isinstance(x, (int, float, bool))

class Heap :
    def __init__(self) :
        self._data_list = []

    @property
    def size(self) :
        return len(self._data_list)

    def _bubble_up(self, c_idx) :
        # since this is a private method, we assume
        # c_idx is a valid and non-negative index
        # we have c_idx == 2*p_idx + 1 or c_idx ==  2*p_idx + 2
        data = self._data_list
        while True :
            if c_idx == 0 : break
            p_idx = (c_idx + 1)//2 - 1
            parent, child = data[p_idx], data[c_idx]
            if parent < child :
                data[p_idx], data[c_idx] = child, parent
                c_idx = p_idx
            else :
                break

    def _bubble_down(self, p_idx) :
        data = self._data_list
        while True :
            # since this is a private method, we assume
            # p_idx is a valid and non-negative index
            l_idx, r_idx = 2*p_idx + 1, 2*p_idx + 2
            parent = data[p_idx]
            if l_idx < len(data) :
                left = data[l_idx]
            else : # if no left child, there is none at all
                break
            if r_idx < len(data) :
                right = data[r_idx]
            else : # only a left child, which is also terminal
                if left > parent :
                    data[p_idx], data[l_idx] = left, parent
                break # left can't have any children
            if left <= parent and right <= parent :
                break # parent dominates
            if left > right :
                data[p_idx], data[l_idx] = left, parent
                p_idx = l_idx
            else :
                data[p_idx], data[r_idx] = right, parent
                p_idx = r_idx

    def insert(self, value) :
        assert is_real_num(value)
        new_idx = len(self._data_list)
```

```
        self._data_list.append(value)

        self._bubble_up(new_idx)

    def pop_max(self) :
        data = self._data_list
        if self.size > 1 :
            top = data[0]
            data[0] = data.pop()
            self._bubble_down(0)
            return top
        elif self.size == 1 :
            return data.pop()
        else :
            raise IndexError("pop from empty Heap")

def new_path_with_suffix(path, suffix) :
    split = path.split('.')
    split[len(split) - 2] = split[len(split) - 2] + suffix
    return '.'.join(split)

def sort_all_lines(path) :
    assert isinstance(path, str) and len(path) > 0
    # we create out new file name
    sorted_path = new_path_with_suffix(path, '_sorted')
    # note : fp stands for file pointer
    with open(path, 'r') as read_fp :
        with open(sorted_path, 'w') as write_fp :
            heap = Heap()
            line_count = 0 # only used for error message
            for line in read_fp :
                line_count += 1
                sorted_data = []
                try :
                    # note : eval will turn ints into ints
                    # and floats into floats
                    data = map(eval, line.split(','))
                    for v in data :
                        # we reuse the same heap
                        heap.insert(v)
                    while heap.size > 0 :
                        sorted_data.append(heap.pop_max())
                    sorted_data_strs = map(str, sorted_data)
                    write_fp.write(','.join(sorted_data_strs) + '\n')
                except :
                    write_fp.write("Error : line {} is not a ".format(line_cou
nt)
                                   + "valid comma separated list of real numbers
.\n")
    # the with statements close everything for us
```