CPSC 335
Spring 2014
Project #1 — empirical analysis

Introduction

In this project you will design, implement, and analyze straightforward algorithms for three problems. For each problem, you will design an algorithm, describe your algorithm using clear pseudocode, analyze it mathematically, implement your algorithm in Python, measure its performance, compare your experimental results with the efficiency class of your algorithm, and draw conclusions.

The hypothesis

This experiment will test the hypothesis that *for large values of n, the mathematically-derived efficiency class of an algorithm accurately predicts the observed running time of an implementation of that algorithm.*

The problems

All three problems involve string processing.

1. The *greatest string character* problem is:
   input: a string s of length $n > 0$
   output: the character c in s with the greatest integer value
   size: n

   This problem is related to the problem of determining which text encoding a file has used (ASCII, Latin-1, UTF-8, etc.).

There is a straighforward decrease-by-one algorithm that solves this problem in $O(n)$ time.

2. The next problem deals with finding strings that start and end with the same character, which we will call "oreos."

   The *longest oreo problem* problem is:

   input: a string s of length $n > 0$
   output: the longest nonempty substring u of s such that the first and last characters of u are identical
   size: n

   There is a straighforward decrease-by-one algorithm that solves this problem in $O(n^2)$ time.

3. The *longest repeated substring* problem is:

   input: a string s of length $n > 0$
   output: the longest nonempty substring u of s such that u appears more than once in s
   size: n

   Long repeated substrings often show up as a result of mistakenly copy-pasting something twice, so it can be helpful to detect them.

   There is a straighforward decrease-by-one algorithm that solves this problem in $O(n^3)$ time.

Algorithm design, pseudocode, and mathematical analysis

First, design an algorithm for each of the problems. This is not intended to be particularly difficult; this project focuses on empirical analysis, not algorithm design. So do not be surprised if your algorithms are simple. As stated above, there are straightforward algorithms that solve the problems in $O(n), O(n^2)$, and $O(n^3)$ time respectively.

Once you have worked out your algorithms, write clear pseudocode for each. As discussed in lecture, we consider pseudocode to be clear when a

typical student in this class could implement it without any further explanation.

Then, analyze each algorithm mathematically. The goal is to prove that each algorithm's worst case running time should be $O(f(n))$, for some specific $f(n)$. I expect that your algorithms' efficiency classes will probably be $O(n), O(n^2)$, and $O(n^3)$, but if not, they will almost certainly be one of the common efficiency classes listed in section 2.3 of the lecture notes.

Implementation

Implement each of your algorithms in Python 3. You must write your own code for the algorithm implementations, which should correspond directly to your pseudocode. The Python library may have built in functions to solve parts of these problems, but you can only use them if they really do correspond to the algorithm described by your pseudocode.

Each algorithm should be encapsulated as a single clearly-named function (which may call helper functions if you wish).

I have provided a template Python source file to help get you started. You may freely use any part of the template. It shows how to encapsulate an algorithm in a function, access command line arguments, load a text file into a Python string, and how to use the time.perf_counter() function to measure the run time of Python code.

Empirical analysis

To analyze the algorithms empirically you will need to run them against representative inputs of various sizes n, measure the elapsed running time (in seconds) of each run, graph these results on a scatter plot, and try to infer which complexity class each plot corresponds to.

Section 2.6 of the Levitin textbook, and section 2.4 of the lecture notes, describe how to conduct an empirical analysis in general terms. As discussed there, you should create a *test harness* program that runs your code and measures the elapsed time of the code corresponding to the algorithm in question. Your test program should perform the following steps:

1. Read a text filename and n value from the command line, read the file into a string, form a string s from the first n characters of the file, and print the value of n. (The provided template already does this step.)
2. Use your algorithm to find the greatest character in the string, while measuring how long the process takes.
3. Print out the output of your algorithm.
4. Print out the elapsed time.
5. Repeat steps 2–4 for your other two algorithms.

You will need to assemble your own corpus of large text files to use as problem instances. In order to use truly large values of n, these files may need to be quite large (e.g. multiple megabytes). One source of free, large text files is Project Gutenberg (`http://gutenberg.org`) which provides works of literature as plain text files.

Sample output

The following shows the output of a solution program when given the first $n = 100$ characters of the Project Gutenberg edition of *The Adventures of Huckleberry Finn*.

```
Loaded "pg76.txt" of length 593144
n = 100
largest character = 65279
elapsed time = 1.8826998712029308e-05
longest oreo = [e Project Gutenberg EBook of Adventures of Huckleberry Finn, Complete
by Mark Twain (Samuel Cle]
elapsed time = 0.0019716549995791866
```

```
longest repeated substring = [ of ]
elapsed time = 0.004276501000276767
```

What to measure

The goal is to draw a scatter plot graph for each algorithm's running times (a total of three plots). The values of n should be on the horizontal axis (x-axis) and the time values should be on the vertical axis (y-axis). Each plot should have a title and axis labels and be legible.

Each plot also needs to have enough data points to interpolate a fitting curve. 5 is the smallest number that might be reasonable. So run each algorithm for at least 5 different values of n. If possible, include at least one problem instance that's large enough to make each of the three algorithms run for at least one minute.

Since your algorithms will probably have differing time complexities, some of your implementations may run significantly faster than others. You may need to interrupt your program with CTRL-C, or temporarily disable some of your algorithms with **if** statements, in order to gather all the time data.

You can generate your best fit lines in software such as Excel or Mathematica, or draw them by hand.

Deliverables

Produce a written project report. Your report should include the following:

1. Your name(s), CWID(s), and an indication that the submission is for project 1.
2. Three scatter plots meeting the requirements described above.
3. Your pseudocode for all three algorithms.

4. Output from your program, for one instance of size $n = 100$ and another of size $n = 2000$.
5. Your complete Python source code.
6. Answers to the following questions, using complete sentences.
   (a) What did you use as problem instances? How representative are your problem instances of real-world printable strings, and why?
   (b) What is the efficiency class of each of your algorithms, according to your own mathematical analysis?
   (c) Are the best fit lines on your scatter plots consistent with these efficiency classes? Justify your answer.
   (d) Is this evidence *consistent* or *inconsistent* with the hypothesis stated on the first page? Justify your answer.

Your document *must* be uploaded to Titanium as a single PDF file.

Due Date

The project deadline is Friday, 2/21, 11:55 pm. Late submissions will not be accepted.