Birla Institute of Technology & Science, Pilani

Work-Integrated Learning Programmes Division
MTech in Data Science & Engineering
S1_2024-2025, DSECLZG519- Data Structures & Algorithms Design

Assignment 2 – PS03 - [URL Caching] - [Weightage 13%]

Read through this entire document very carefully before you start!

Problem Statement

You are working on a web crawler application that needs to efficiently check whether a URL has been previously processed to avoid redundant processing. To optimize this process, you decide to implement a simple Bloom filter for URL caching. Your task is to create a program that adheres to the following specifications:

●Implement a Bloom filter with a configurable size and number of hash functions for URL caching.

●Develop functions to add URLs to the Bloom filter and check whether a URL is likely to be in the filter.

●Ensure that the implementation provides a consistent output for the same set of operations across different environments.

●Create your own hash functions. You can create up to 3-4 hash functions.

●Test the code for large inputs, i.e. at least 1,00,000 nodes.

Your program should accept input in the following format:

1. For adding a URL: **ADD url**
2. For checking URL existence: **CONTAINS url**

Your program should produce output in the following format:

1. For adding a URL: Added: url
2. For checking URL existence: URL Existence Check for url: True/False

There is a correlation to the size of Bloom filter and number of hash functions used. Use different data and plot some graphs and include this in the document.

Requirements:

1. Implement the above problem statement using **Python 3.7**
2. Read the input from a file(**inputPS03.txt**), which contains the list of actions to be performed.
3. You will output your answers to a file (**outputPS03.txt**) for each line.
4. Perform an analysis for the features above and give the running time in terms of input size: n.

Sample Input

Input will be taken from the file(inputPS03.txt).

ADD https://example.com/page1

CONTAINS https://example.com/page1

ADD https://example.com/page2

CONTAINS https://example.com/page2

CONTAINS https://example.com/page3

Note that the input/output data shown here is only for understanding and testing, the actual file used for evaluation will be different.

Sample Output

Display the output in outputPS03.txt.

Added: https://example.com/page1

URL Existence Check for https://example.com/page1: True

Added: https://example.com/page2

URL Existence Check for https://example.com/page2: True

URL Existence Check for https://example.com/page3: False

Note that the input/output data shown here is only for understanding and testing, the actual file used for evaluation will be different.

1. Deliverables

1. PDF document **designPS03_<group id>.docx** detailing your solution.
2. **[Group id] _Contribution.xlsx** mentioning the contribution of each student in terms of percentage of work done. Columns must be "Student Registration Number", "Name", "Percentage of contribution out of 100%". If a student did not contribute at all, it will be 0%, if all contributed then 100% for all.
3. **inputPS03.txt** file used for testing
4. **outputPS03.txt** file generated while testing
5. .py file containing the python code. Create a single notebook. Do not fragment your code into multiple files
6. Zip all of the above files including the design document in a folder with the name:
   a. **[Group id] _A2_PS03_URL_Caching.zip** and submit the zipped file.
   b. Group Id should be given as **Gxxx** where **xxx** is your group number. For example, if your group is 26, then you will enter **G026** as your group id.

2. Instructions

1. It is compulsory to make use of the data structure(s) / algorithms mentioned in the problem statement.
2. Ensure that all data structure insert and delete operations throw appropriate messages when their capacity is empty or full. Also ensure basic error handling is implemented.
3. For the purposes of testing, you may implement some functions to print the data structures or other test data. But all such functions must be commented before submission.
4. Make sure that your read, understand, and follow all the instructions
5. Ensure that the input, prompt and output file guidelines are adhered to. Deviations from the mentioned formats will not be entertained.
6. The input, prompt and output samples shown here are only a representation of the syntax to be

used. Actual files used to evaluate the submissions will be different. Hence, do not hard code any values into the code.

7. Run time analysis is to be provided in asymptotic notations and not timestamp based runtimes in sec or milliseconds.

8. Please note that the design document must include:

   a. The data structure model you chose with justifications

   b. Details of each operations with the time complexity and reasons why the chosen operations are efficient for the given representation

   c. One alternate way of modeling the problem with the cost implications.

9. Writing a good technical report and well documented code is an art. Your report cannot exceed 4 pages. Your code must be modular and quite well documented.

10. You may ask queries in the dedicated discussion section. Beware that only hints will be provided and queries asked **in other channels will not be responded to**.

Instructions for use of Python:

1. Implement the above problem statement using Python 3.7+.

2. Use only native data types like lists and tuples in Python, do not use dictionaries provided in Python. Use of external libraries like graph, numpy, pandas library etc. is not allowed. The purpose of the assignment is for you to learn how these data structures are constructed and how they work internally.

3. Create a single *.py file for code. Do not fragment your code into multiple files.

4. Do not submit a Jupyter Notebook (no *.ipynb). These submissions will not be evaluated. You can create in Notebook and download as .py if needed.

5. Read the input file and create the output file in the root folder itself along with your .py file. Do not create separate folders for input and output files.

3. Deadline

1. The strict deadline for submission of the assignment is **Sunday, March 09, 2025 11:55PM.**

2. The deadline has been set considering extra days from the regular duration in order to accommodate any challenges you might face. No further extensions will be entertained.

3. Late submissions will not be evaluated.

4. How to submit

1. This is a group assignment.
2. Each group has to make one submission (only one, no resubmission) of solutions.
3. Each group should zip all the deliverables in one zip file and name the zipped file as mentioned above.
4. Assignments should be submitted via Canvas > Assignment section. Assignments submitted via other means like email etc. will not be graded.

5. Evaluation

1. The assignment carries **13 Marks**.
2. Grading will depend on:
   a. Fully executable code with all functionality working as expected
   b. Well-structured and commented code
   c. Accuracy of the run time analysis and design document.
   d. Every bug in the functionality will have negative marking.
   e. Marks will be deducted if your program fails to read the input file used for evaluation due to change / deviation from the required syntax.
   f. Use of only native data types and avoiding libraries like numpy, graph and pandas will get additional marks.
3. We encourage students to take the upcoming assignments and examinations seriously and submit only original work. Please note that plagiarism in assignments will be taken seriously. All groups that are booked under plagiarism will be given 0 marks and no further discussion will be entertained. Please refer to the detailed policy in ELearn Files.
4. Source code files which contain compilation errors will get at most 25% of the value of that question.

6. Readings

Text book: Algorithms Design: Foundations, Analysis and Internet Examples Michael T. Goodrich, Roberto Tamassia, 2006, Wiley (Students Edition). Chapters: **XX**