CPSC 331 Empirical Exercise II

Ali Akbari

TOTAL POINTS

3 / 3

QUESTION 1

1 Report 3 / 3

✓ **+ 3 pts** **Good:**

Clear and consist data collection.

Observations properly explained and connected to theory.

Results show expected behaviour, clear explanations included.

Report generally well-written.

- 0.5 pts Minor error

+ 2 pts Fair:

Some evidence of data, incorrect data etc.

Non-technical explanations, incorrect analysis/interpretation of data etc.

Data included but not explained adequately.

Report written well for the most part but has grammar / style issues.

+ 1 pts Deficient:

No evidence of data, little to no effort made to collect data.

Data not explained/connected to theory.

Poorly written report.

+ 0 pts Not done:

Nothing submitted.

CPSC 331 Emerpical Exercise #2 **Ali Akbari**

Problem:
Investigating the goodness of two hash functions on String objects.

Chose a prime close to N/0.65 = 11344/0.65 = 17452 $\approx$ 17467
Table Size = 17467

First String Hash Function Table results:
Percentage of empty slots: 52.115%
Maximum number of collisions per slot: 5
Average number of collisions per slot (only for the non-empty slots): 0.645
MyString Hash Function Table results:
Percentage of empty slots: 92.884%
Maximum number of collisions per slots: 47
Average number of collisions per slot (only for the non-empty slots): 8.93

Methodology:
The main function controlled reading the text file and calling other functions. The main function buffer reader was used to traverse the text file line by line and send each word to both hash functions to build the hash tables. A prime number was chosen for the table size that was close to N/0.65, N is the number of words in the text file. The table recorded the number of elements that were hashed to a slot. For the second hash function, each word was sent to java's hashCode and the result was also modulo by the table size, and the slot was incremented by one. To calculate the percent of empty slots I traversed through the hash tables and looked for empty slots, if I found one I increment a counter, and returned the percentage of empty slots which is equal to total empty slots/ table size.To calculate the maximum number of collisions in the table I looped/traversed the hash table to find the largest number in the slot which represents collision. To calculate the average of collisions per slot I added all the collisions, i.e traversing the hash table and adding its values in the slots, and then I returned the total collision divided by the number of slots/ table size.
Analysis of Result:
We can see from the statistics that the first hash function has more slots filled up (47.885 % filled up) and the myString hash function has only filled up 8% of the whole hash table. Since the input sizes(number of words in the text file) are the same and the table size are the same for both functions we know that there were more collisions in the myString hash function as it has only used 8% of the table. The statistics are in favor of the First String Hash function being better than the myString hash function which is expected. The goodness of the first hash function is due to the extra step it takes to help against collisions. The modulo operation of the prime number helps spread the inputs to a wider range of slots thus lowering the chances of collisions. Our statistics show that the first hash function not only has a lower average number of collisions per slot, but it also has a lower number for the maximum collisions in a single slot.

1 Report 3 / 3

✓ **+ 3 pts** **Good:**

Clear and consist data collection.

Observations properly explained and connected to theory.

Results show expected behaviour, clear explanations included.

Report generally well-written.

- 0.5 pts Minor error

+ 2 pts Fair:

Some evidence of data, incorrect data etc.

Non-technical explanations, incorrect analysis/interpretation of data etc.

Data included but not explained adequately.

Report written well for the most part but has grammar / style issues.

+ 1 pts Deficient:

No evidence of data, little to no effort made to collect data.

Data not explained/connected to theory.

Poorly written report.

+ 0 pts Not done:

Nothing submitted.