


National University of Computer and Emerging Sciences, Lahore Campus

| | | | | |
|---|-------------|-----------------------|--------------|-------------|
|  | Course: | Information Retrieval | Course Code: | CS317 |
| | Program: | BS(Computer Science) | Semester: | Spring 2019 |
| | Duration: | 60 min | Total Marks: | 30 |
| | Paper Date: | Feb 26, 2019 | Weight: | 15% |
| | Section: | CS | Page(s): | 2 |
| | Exam: | Mid1 | Roll No. | |

Instructions/Notes: Please show all the working/reasoning, no marks will be given without it.
Write your answers neatly and in sequence on answer sheet.

Question 1 (12 points): Dictionary Compression (5+5+2) points

In Python3 each string takes **49 + (Length of String) bytes**. 49 bytes are overhead for each string, and **Length of String** is number of characters in the string.

For example:

- empty string takes 49 bytes
- "fish" takes 49+4=53 bytes
- "tropical" takes 49+8= 57 bytes
- "tropicalfish" takes 49+12=61 bytes

- Just consider the dictionary/vocabulary part of inverted index. The dictionary contains 8000 word and average length of word is 5. Currently each word is stored as a separate string (as shown in figure 1). On average, how much memory is consumed to store these words in Python3?
- On average, how much memory you will save if you store the **dictionary as String** (i.e. all the words as in one string) and keep the position of word in String, in index, as an integer. Example is shown in figure 2.
Note that it takes 28 bytes in python to store an integer in range $[0-10^7]$ and 32 bytes for integer in range $(10^7 - 10^{18})$
- Memory required to store a float number in python is 24 bytes. If you use float instead integer how much memory will you save?

| Words | |
|-------|------|
| a | |
| about | |
| an | |
| anna | |
| apple | |
| . | |
| . | |
| zulu | |

Figure 1: Index, showing only the dictionary

String: aaboutanannaapple....

| Words Position in String | |
|--------------------------|------|
| 0 | |
| 1 | |
| 6 | |
| 8 | |
| 12 | |
| . | |
| . | |

Figure 2: Keeping Dictionary(words) as a **String** and storing the starting position of word in index. For example **anna** starts at index 8 in **String**, so 8 is written in place of **anna**

Question 2 (5 points):

Decode the following sequence, encoded using gamma encoding.

1101001110111

Question 3 (3 points):

Consider the technique given in Manku[2007] of keeping sorted permuted tables of simhash of documents and answer the following question.

For 64 bit simhash and $k=3$. If simhash is split in 8 equal parts. How many permuted tables will be created?

Question 4 (5 points):

Merge the following two indexes into one. (Posting lists contain only document IDs in which the term occurs)

| Term | Postings |
|--------|------------|
| and | 2,3,5,6,20 |
| around | 6,7 |
| fish | 1,3,5,18 |

Table 1: Index 1

| Term | Postings |
|--------|--------------|
| apple | 8,9,10,11,30 |
| around | 9,12,15 |
| fish | 9,14,15,16 |
| world | 8,10,16 |

Table 2: Index 2

Question 5 (5 points):

Consider a part of robot.txt files of Wikipedia, and answer the following question

```
User-agent: Orthogaffe
Disallow:

User-agent: DOC
Disallow: /

User-agent: *
Allow: /w/api.php?action=mobileview
Allow: /w/load.php?
Allow: /api/rest/?doc
Disallow: /w/
Disallow: /api/
Disallow: /trap/
```

1. Which crawler is not allowed to crawl any page?
2. Which crawler is allowed to crawl all pages?
3. If you are trying to crawl Wikipedia, which of the pages you are not allowed to crawl?
https://foundation.wikimedia.org/wiki/Developer_app_guidelines
<https://en.wikipedia.org/w/api.php>
<https://en.wikipedia.org/api/>