CMI/Aug-Nov 2020/Assignment 3 Sep 2020

Chennai Mathematical Institute

INFORMATION RETRIEVAL DEADLINE: OCT 12, 2020. MAX MARKS: 10.

Ramesh, an enthusiastic student from the Information Retrieval class, was excited when he heard about the “Dictionary as a String” idea to conserve space. So, after the class was over, he took some text from the CMI website[1]. He removed comma and periods. He replaced hyphens with a space. He converted the text to lower-case. Ramesh did not remove stop-words. Ramesh did not apply stemming or lemmatization. The modified text as Ramesh obtained after applying all these pre-processing steps is given to you.

Assume that each character consumes a byte of memory. As Ramesh is not good with programming, he seeks your help for the following questions.

(1) How many bytes (B1) would this text consume if saved as fixed length array entries where each array entry uses 20 bytes?
(2) How many bytes (B2) would this text consume if saved as “Dictionary as a String” with a blocking factor of (k=4) four?
(3) Print this text in the format of “Dictionary as a String” with a blocking factor of (k=4) four.
(4) Print this text front-coded with blocking factor k=4.

Note that we ignore the document frequency, postings-pointer and term-pointer storage in this assignment.

Submit one text/pdf file in the following format:

```
Roll-No#1, Roll-No#2
B1, B2
String that represents the ‘‘dictionary as a string’’ with k=4:
---
Front-Coded string with k=4:
---
Your comments/remarks/assumptions if any:
---
```

You may use any programming language or tools for this purpose. You do not need to submit your code. If you still wish to submit your code, push your code to github as a private project, include me as a collaborator, and just include the link in your assignment.

Mini-Project Idea: For your mini-project, you may extend this program with Variable-Byte encoding for postings, and build a tutorial site that I can use to teach these ideas in future years!

[1] https://www.cmi.ac.in//about/research.php

The text that Ramesh obtained after pre-processing:

the research groups in mathematics and computer science at cmi are among the best known in the country recently a research group has also been set up in physics the institute has nurtured an impressive collection of phd students the main areas of research in mathematics pursued at the institute are algebra analysis differential equations geometry and topology in computer science the main areas of research are formal methods in the specification and verification of software systems design and analysis of algorithms computational complexity theory and computer security in physics research is being carried out mainly in quantum field theory mathematical physics and string theory members of the faculty have strong academic ties with reputed institutions in india and abroad the institute has an active visitors programme and is open to offering flexible visiting positions at all levels the institute has exchange programmes with the ecole normale superieure paris and the ecole normale superieure de cachan each year the two top ranking senior bsc students from cmi spend the summer at ens paris working on research problems with faculty there in return four phd students from ens visit cmi each year to participate in research and teaching cmi is also involved in a number of collaborative research projects both with academic and industrial partners from 2005 2009 the institute was involved in a four year project on timed and distributed computing systems under the indo french networking programme jointly funded by the french ministry of science cnrs and the indian department of science and technology the project involved scientists from cmi imsc and iisc bangalore in india and the university of paris 7 ens de cachan and the university of bordeaux 1 in france a three year indo french research collaboration involving cmi and ens de cachan has been initiated in 2008 under the arcus programme funded by region ile de france cmi has had sponsored research projects with honeywell technology solutions laboratory bangalore siemens’ corporate technology research centre bangalore and tata research development and design centre pune the institute actively supports conferences and workshops and other activities that contribute to the growth of mathematics and computer science in the country