Anh Tran
CS 351
Homework 4

Write up and Comparison

1. Description of the "compress_bitmap.py" file:
   a. There are 2 functions that are included:

i. *create_index(input_file, output_path,sorted)*: takes in the input file name, the output path to create output file and a Boolean value for 'sorted' parameter:

ii. *compress_index(bitmap_index, output_path,compression_method, word_size)*: reads the bitmap_index file, create a new output file with output_path and compress the bimap using compression_method with word_size

2. Compare the size of the bitmap indexes and compressed versions on the large test animals.txt file. Write an analysis on why you think they are different size. Did sorting help with the compression and by how much?

Table 1. Size of bitmap indexes of animal.txt unsorted vs sorted

Bitmap Index	
Sorted/ Unsorted	**File Size**
animals.txt_unsorted	1.7 MB
animals.txt_sorted	1.7 MB

Table 2. Compressed versions on large test file

WAH Compression				
Sorted/ Unsorted	**Word Size**	**File Size**	**Number of runs**	**Number of literals**
animals.txt_unsorted	8	1,559,576 bytes	76429	152147
	16	1,661,744 bytes	14025	92647
	32	1,650,000 bytes	1271	50345
	64	1,626,128 bytes	26	25382
animals.txt_sorted	8	279,496 bytes	226996	1580
	16	154,256 bytes	104962	1710
	32	159,312 bytes	49838	1778
	64	244,688 bytes	23604	1804

There is not much difference in size of bitmap indexes of unsorted vs sorted uncompressed file (~ 1.7MB). Since we only re-order (sort) the data but didn't compress or reduce the size of the file.

As for compressed files, the unsorted files require larger space, around 6 - 10 times larger for word size of 8,16,32,64 according to Table 2.

The reduction happens because we compress over the columns of the data and by sorting the data, more run chunks can be compressed consecutively, and we can keep adding the number of runs to the string that is used to store the number of fills without having to create a new compressed chunk unless run is full. That explains the number of runs for sorted files are significantly more than unsorted ones.

3. Did different word sizes have different compression ratios and why do you think that is?

Yes, different word sizes have different compression ratios. As the word size doubles, the number of runs significantly decrease, around ½ ratio since we grab larger word, there are more likelihood to be a literal, instead of fill. Regarding compressed file size, the unsorted files does not have much difference compared to bitmap index file size. As for sorted files, word size of 8 and 64 have larger size versus word size of 16 and 32.