CS 351 Assignment Four

Andrew Oakes

March 2019

1 Summary

The one and only file, wah.py, contains the code for creating a bitmap from a file named 'animals.txt' and compressing a sorted and unsorted version of the bitmap using the WAH algorithm with 32 and 64 bit words. It creates six files in total:

- unsorted_bitmap_animals.txt
- sorted_bitmap_animals.txt
- sorted_bitmap_compressed32_animals.txt
- unsorted_bitmap_compressed32_animals.txt
- sorted_bitmap_compressed64_animals.txt
- unsorted_bitmap_compressed64_animals.txt

2 Results

Table 1: Compression Ratio $\frac{newSize}{OldSize}$

	32-bit	**64-bit**
Sorted	.06791	.13204
Unsorted	.97068	.95613

Table 2: Runs of Zeroes

	32-bit	**64-bit**
Sorted	41045	19740
Unsorted	1260	18

Table 3: Runs of Ones

	32-bit	**64-bit**
Sorted	8795	3870
Unsorted	0	0

Table 4: Literals

	32-bit	**64-bit**
Sorted	1776	1798
Unsorted	50356	25390

3 Analysis

3.1 Sorting

From the results, we can tell the file with the best compression ratio was the 32-bit sorted compression, and the runner up was the 64-bit sorted compression. It is apparent that sorting the file makes a huge difference in the effectiveness of run length compression. The difference in the rate of compression between sorted and unsorted is roughly 90% in the case of the 32-bit compression, and 82% in the case of the 64-bit compression. The drastic difference in compression rate is due to the increase in the number of similar rows right next to each other in the sorted version, thus causing more runs and less literals. This is supported by the run counts in tables two and three, and the literal counts in table four.

3.2 Word Size

While word size seems to have less effect on the overall compression rate, it still is important that the word size is not too small or too large. From the data in the tables, we can see that the 64-bit compression was about half as effective as the 32-bit compression, likely due to the fact that 64 bits is larger than the average run size. This won't always be the case, but for the animals.txt file the runs are not large enough for that to be efficient.