LESSON 2: TEXT COMPRESSION

AP CSP – adapted from Code.org

GOALS: WE WILL BE ABLE TO…

- Collaborate with a peer to find a solution to a text compression problem using the Text Compression Widget (lossless compression scheme)
- Explain why the optimal amount of compression is impossible or “hard” to identify
- Explain some factors that make compression challenging
- Develop a strategy (heuristic algorithm) for compressing text.
- Describe the purpose and rationale for lossless compression

WARM-UP: ABBR IN UR TXT MSGS

- When you send text messages to a friend, do you spell every word correctly?
- Why do you use these abbreviations? What is the benefit?
- When you abbreviate or use coded language to shorten the original text, you are “compressing text.” Computers do this too, *in order to save time and space*

WHAT'S THIS ABOUT? – COMPRESSION: SAME DATA, FEWER BITS

- The art and science of compression is about figuring out how to represent the SAME DATA with FEWER BITS
- Why is this important? One reason is that storage space is limited and you'd always prefer to use fewer bits if possible… a much more compelling reason is that there is an upper limit to how quickly bits can be transmitted over the Internet
- What if we need to send a large amount of text faster over the Internet, but we've reached the physical limit of how fast we can send bits?
- Our only choice is to somehow capture the same information with fewer bits; we call this **compression**

ACTIVITY: DECODE THIS MYSTERY TEXT

- Activity Guide – “Decode this message – Activity Guide”
- What was the original text?

RECAP: HOW MUCH WAS IT COMPRESSED?

- To answer, we need to compare the number of characters in the original poem to the number of characters needed to represent the compressed version.

- Important notes:
  - The compressed poem is not just the compressed text, but also the key to solve it.
  - Thus, you must account for the total number of characters in the message plus the total number of characters in the key to see how much you've compressed the original

USE THE TEXT COMPRESSION WIDGET

- https://docs.google.com/document/d/1Z3kwOLtnzV-vFvNYVCI5CwJlqW7oMPHsmLEQNog-0jM/edit
- Video: Text Compression with Aloe Blacc
- Challenge: compress your assigned poem as much as possible.
  - Compare with other groups to see if you can do better
  - Try to develop a general strategy that will lead to a good compression

DISCUSSION: PROPERTIES AND CHALLENGES WITH COMPRESSION

- What makes doing this compression hard?
- There is a tipping point: when you have a large dictionary
- Do we think that these compression amounts that we've found are the best? Is there a way to know what the best compression is?
- But is there a process a person can follow to find the best (or pretty good) compression for a piece of text?

DEVELOP A HEURISTIC FOR DOING A COMPRESSION

- Heuristic: a problem solving approach(typically an algorithm) to find a satisfactory solution where finding an optimal or exact solution is impractical or impossible.

- Instructions:
  - Continue working on compressing your poem using the Text Compression Widget. As you do,develop a set of rules, or a "heuristic" that generally seems to provide good results.
  - Record your heuristic as a list of steps that someone else unfamiliar with the problem could follow and still end up with decent compression.

THE POINT:

- There is no real way to determine for sure that you've got the best compression besides trying everything possible by brute force.
- Heuristics are techniques for at least making progress toward a "good enough" solution.
- Following the same heuristic might lead to different results.

HEURISTIC DISCUSSION:

- Do you think it's possible to describe (or write) a specific set of instructions that a person could follow that would always result in better text compression than your heuristic? Why or why not?

- Is there a way to know that a compressed piece of text is compressed the most possible? If yes, describe how you could determine it. If no, why not?

WRAP-UP

- What did all groups' processes for compression have in common?
- Will following this process always lead to the same compression? (I.e. two people following the process for the same poem, will result in the same compression?)

EXIT TICKET

- Explain the difference between lossless compression and lossy compression
- What is a heuristic?

COMPRESSION IN THE REAL WORLD (.ZIP)

- There is a compression algorithm called LZW compression upon which the common "zip" utility is based.
- Zip compression does something very similar to what you did today with the text compression widget.
- See animation of lzw in action (link)
  - Doesn't compress it the most, but leads to better compression over time.
- Do you want to use zip compression for real?
  - Windows and Mac both have it
- Warning: results may vary:
  - Zip works really well for text, but only on large files
  - Zip is meant for text. It might not work well on non-text files

LOSSLESS COMPRESSION

- A method or protocol for using fewer bits to represent the original information
- The way we represented compressed data in this lesson, with a “dictionary” of repeated patterns is somewhat similar to the LZW compression scheme, which is used for zip files and GIF image file format
- **Lossless Compression:** a data compression algorithm that allows the original data to be perfectly constructed from the compressed data

HEURISTICS

- There is no single correct way to compress text using the method we used in this lesson because:
  - There is no known algorithm for finding an optimal solution
  - We don't even know a way to verify whether a given solution is optimal
- There is no way to prove it or derive it beyond trying all possibilities by brute force.
- This is an example of an algorithm that cannot run in a "reasonable amount of time"
- **Heuristic Approach:** a problem-solving approach (algorithm) to find a satisfactory solution where finding an optimal or exact solution is impractical or impossible.