Data Compression

“Most software today is very much like an Egyptian pyramid with millions of bricks piled on top of each other, with no structural integrity, but just done by brute force and thousands of slaves.”
–Alan Kay

LZ 78 (Lempel-Ziv)

The encoder and decoder each maintain a dictionary containing previously seen words

1. Initially the dictionary contains just the empty string

2. The encoder communicates a dictionary entry to the decoder by sending an index in the dictionary to all but the last character + the last character

   - The encoder looks at progressively longer prefixes of the remaining input until it finds the first prefix $\alpha x$ such that $\alpha$ is a word in the dictionary, $x$ is a symbol, and $\alpha x$ is not in the dictionary.

   - The encoder transmits $(i, x)$ to the decoder, where $i$ is the index of $\alpha$ in the dictionary

   - The decoder uses its dictionary to retrieve $\alpha$ at index $i$, and then adds $\alpha x$ to its dictionary
LZ 78: Implementation

Common to implement dictionary as a trie

- **trie**: from retrieval, rhymes with *try*, to distinguish it from *tree*
- **trie**: a tree that uses parts of the key to perform search - a trie associates keys with edges, not nodes
- Let $S$ be a set of strings from alphabet $A$ such that no string in $S$ is a prefix of another
- A trie for $S$ is an ordered tree $T$ such that:
  - each edge of $T$ is labeled with symbols from $A$
  - the ordering of edges of a node follows the canonical ordering of $A$
  - labels of edges on the path from the root to a node in $T$ forms a prefix of a string in $S$
- If the set $A$ of symbols is the set of 256 possible bytes, then each node of the trie might have an array of length 256 to store its children
- To save space, could store the children of a trie node in a linked list until the number of children is sufficiently large ($\geq 10$, for example), and then switch to an array
- Or the children of a trie node could be stored in a hash table

The integers used to represent dictionary entries are indices into an array of pointers into the trie