Crossword Compiler: A Data Structure, Algorithms, and Entropy

Matthew Fahrbach

Outline

Crossword Compiler

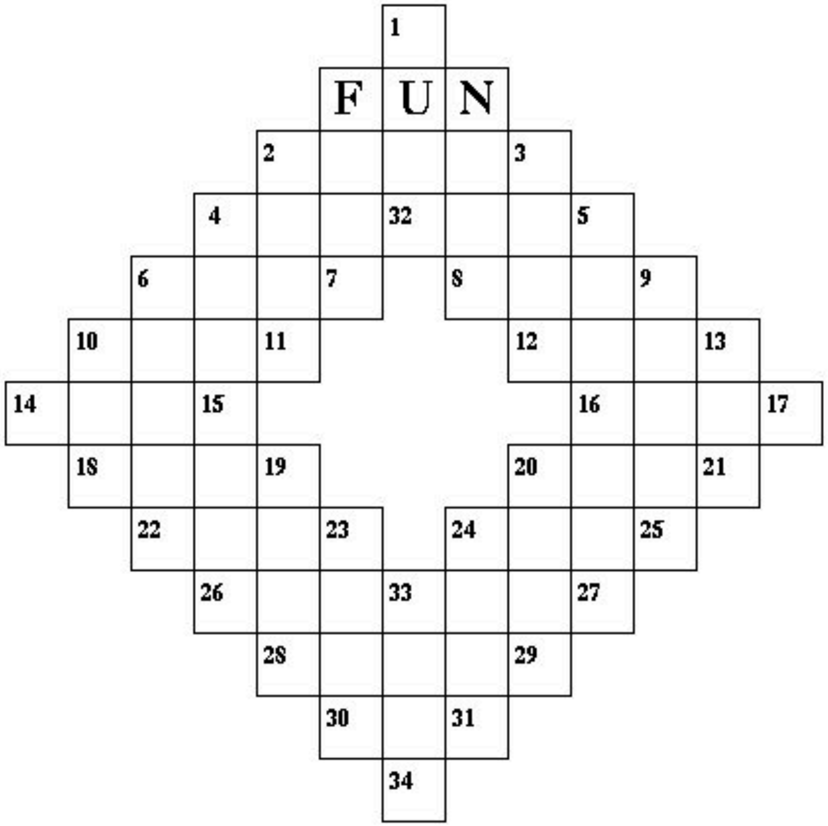
- Crossword Puzzles
- Filler Word Tree
- Word Ranking Algorithm
- Crossword Fill Algorithm

Information Theory

- Claude Elwood Shannon
- Entropy
- Redundancy

Existence of Infinitely Large 2-Dimensional Crosswords

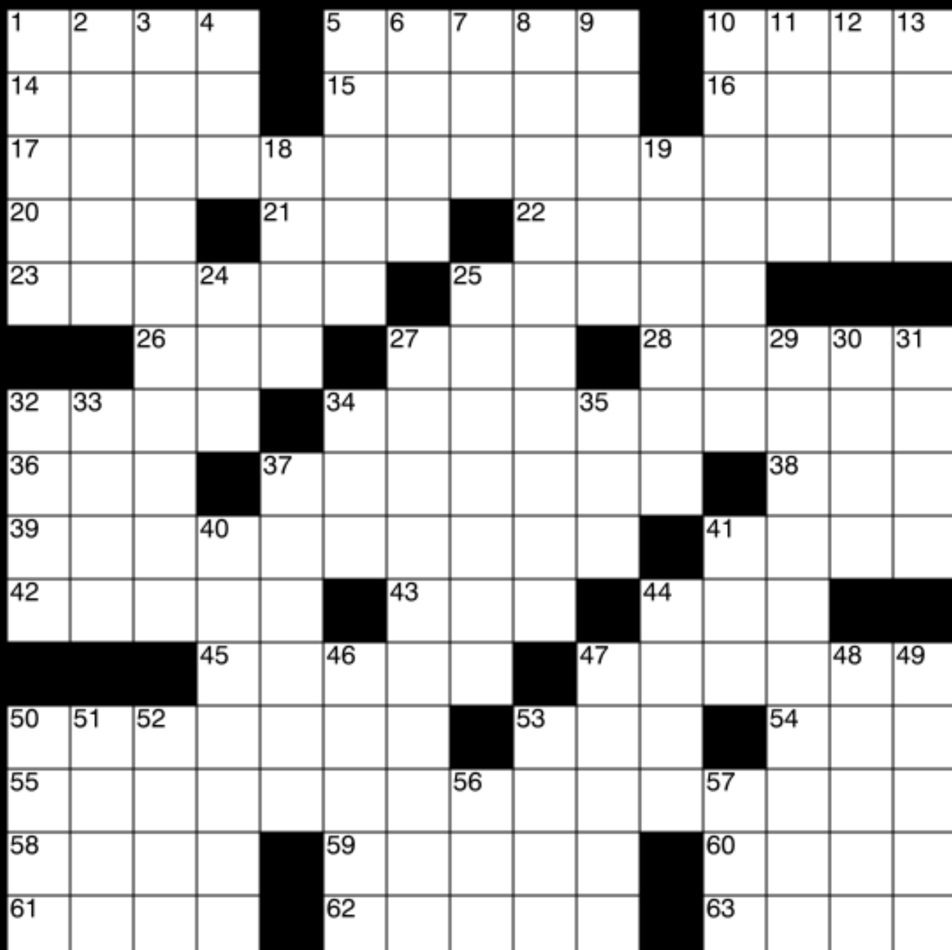
New York World Crossword



By Arthur Wynne, December 21st, 1913

- The first crosswords appeared in English children's puzzle books during the 19^{th} century
- Arthur Wynne was a Journalist from Liverpool, England
- By the 1920s crosswords appeared in almost all American newspapers

New York Times Crossword



- 3-letter word minimum
- A word may only be used once
- Puzzle must have rotational symmetry
- Theme should be interesting and narrowly-defined
- Difficulty increases throughout the week
- Daily Crossword – 15x15
  Sunday Crossword – 21x21 or 23x23

Setup

- Empty crossword template (.txt)
- Themed words (100)
  Percent themed words (10%)
- Filler words – Unix Dictionary (200K)

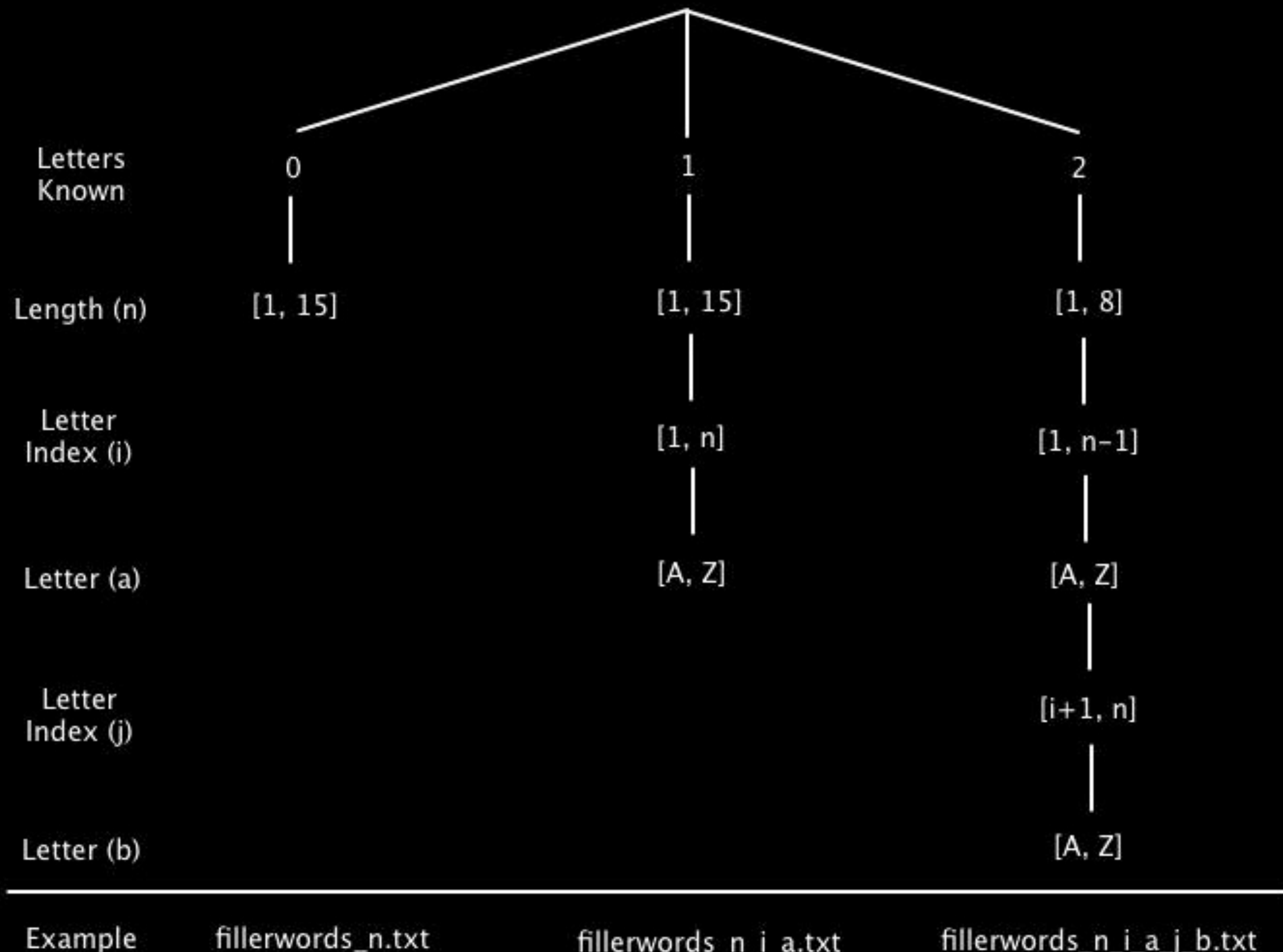
```
. . . . . # # . . . . # . . . . # . . . . # #
. . . . . . # . . . . # . . . . # . . . . . #
. . . . . . # . . . . # . . . . # . . . . . .
. . . . # . . . . # . . . # . . . . # . . . .
. . . # . . . . # . . . . . # . . . . # . . .
# # # . . . . # . . . . . . . # . . . . # # #
. . . . . . # . . . . # . . . . # . . . . . .
. . . . . # . . . . . . # . . . . # . . . . .
. . . . # . . . . . # . . . # . . . # . . . .
. . . # . . . . . # . . . # . . . . . # . . .
# # . . . . . # # . . . . . . . # . . . . . .
# . . . . . # . . . . # . . . . # . . . . . #
. . . . . . # . . . . . . . # # . . . . . # #
. . . # . . . . . # . . . # . . . . . # . . .
. . . . # . . . # . . . # . . . . . # . . . .
. . . . . # . . . . # . . . . . . # . . . . .
. . . . . . # . . . . # . . . . # . . . . . .
# # # . . . . # . . . . . . . # . . . . # # #
. . . # . . . . # . . . . . # . . . . # . . .
. . . . # . . . . # . . . # . . . . # . . . .
. . . . . . # . . . . # . . . . # . . . . . .
# . . . . . # . . . . # . . . . # . . . . . .
# # . . . . # . . . . # . . . . # # . . . . .
```

Filler Word Tree

- Composed of directories and text files
- Allows for quick lookup of partially filled words (regular expressions)
- A node (text file) in the tree contains a subset of the matching words

- *Example:* Query(Q--B) ⇒ fillerwords_4_1_Q_4_B.txt ⇒ [QUAB, QUIB]
  Count(Q--B) ⇒ 2

Filler Word Tree



Filler Word Tree

- Requires hours of preprocessing
- The structure is recyclable
- Words may be added or removed without regenerating the entire tree
- Unfortunately these 200K words require 200MB and are initially read from the disk (though memory is not a problem)

Word Rank Algorithm

- *Inputs:* Crossword, starting cell, and direction
- *Outputs:* Top ten matching words ranked by the sum of their stemming word count

Word Rank Algorithm

Q			B
			A
		E	Y

Example: RankWords(0, 0, Across)

Query(Q--B) $\Rightarrow$ [QUAB, QUIB]

$QUAB_{score}$ = Count(Q--) + Count(U--) + Count(A-E) + Count(BAY)
= 2 + 34 + 17 + 1
= 54

$QUIB_{score}$ = Count(Q--) + Count(U--) + Count(I-E) + Count(BAY)
= 2 + 34 + 5 + 1
= 42

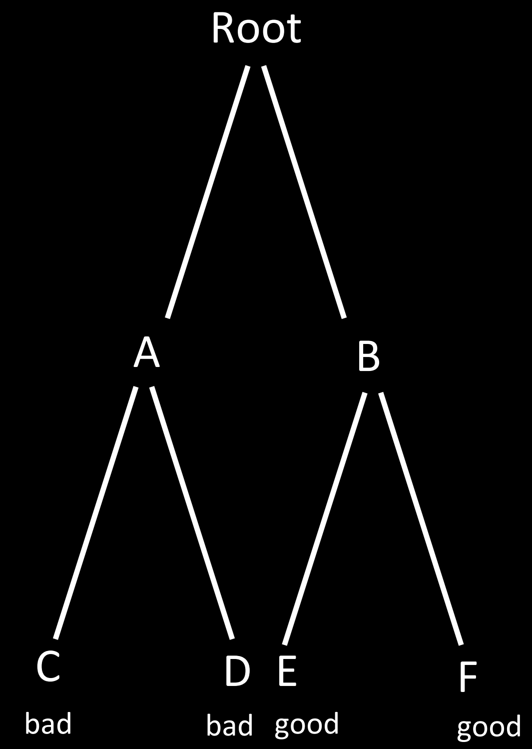
Return [QUAB, QUIB]

Word Rank Algorithm

- Implicitly processes languages for common word structures
- Higher ranked words are more likely to fill and complete the crossword
- If there is an index in the fitting word that has no perpendicular, stemming words it is not returned in the ranked list (pruning for backtracking)

Recursive Backtracking

1. Starting at Root, your options are A and B. You choose A.
2. At A, your options are C and D. You choose C.
3. C is bad. Go back to A
4. At A, you have already tried C, and it failed. Try D.
5. D is bad. Go back to A.
6. At A, you have no options left to try. Go back to Root.
7. At Root, you have already tried A. Try B.
8. At B, your options are E and F. Try E.
9. E is Good. You are finished.



Crossword Fill Algorithm

- *General Algorithm:* Heuristic Backtracking
- Fills an empty crossword with a percentage of themed words and then completes it using the filler word tree

Crossword Fill Algorithm

- Heuristic elements have been implemented experimentally to improve performance
- Select the ranked words at random instead of choosing the highest ranked word at each intersection
- Recursive limits prevent the program from exhaustively searching for a solution down a failing path

Crossword Compiler Demonstration

```
. . . . . # # . . . P # . . . . # . . . . # #
. . . . . . # . . . E # . . . . # . . . . . #
G A N D H I # . . . G # . . . . # . . . . . .
. . . . # . . . . # G . . # . . . . # . . . .
. . . # . . . . # B Y . . . # . . . . # . . .
# # # . . . . # . E L . . . . # D E A N # # #
. . J . . . # . . E E # . . . . # . . . . . L
. . O . . # . . . R E . # . . . . # . . . . Y
. . H . # . . . . Y # . . . # . . . # . . . O
. . N # . . . . . # . . . # . . . . . # . . N
# # D . . . . # # K I N N I C K # . . . . . S
# . G . . . # . . . . # . . . . # . . . . . #
D I R N D L # . . . . . . . # # . . . . . # #
. . A # . . . . . # . . . # S . . . . # . . .
. . H . # . . . # . . . # . W . . . # F I L M
. . A . . # . . . . # . . . I . . # . . . . .
. . M . . E # . . . . # . . N . # . . . . . .
# # # . . D . # . . . . . . G # . . . . # # #
. . . # . W . . # L O U I S # . . . . # . O .
. . . . # A . . . # . . . # . . . . # . . W .
. . . . . R # . . . . # . . . . # . . . . E .
# . . . . D # . . . . # . . . . # . . . . N .
# # . . . S # . . . . # . . . . # # . . . S .

Press <Enter> to continue
```

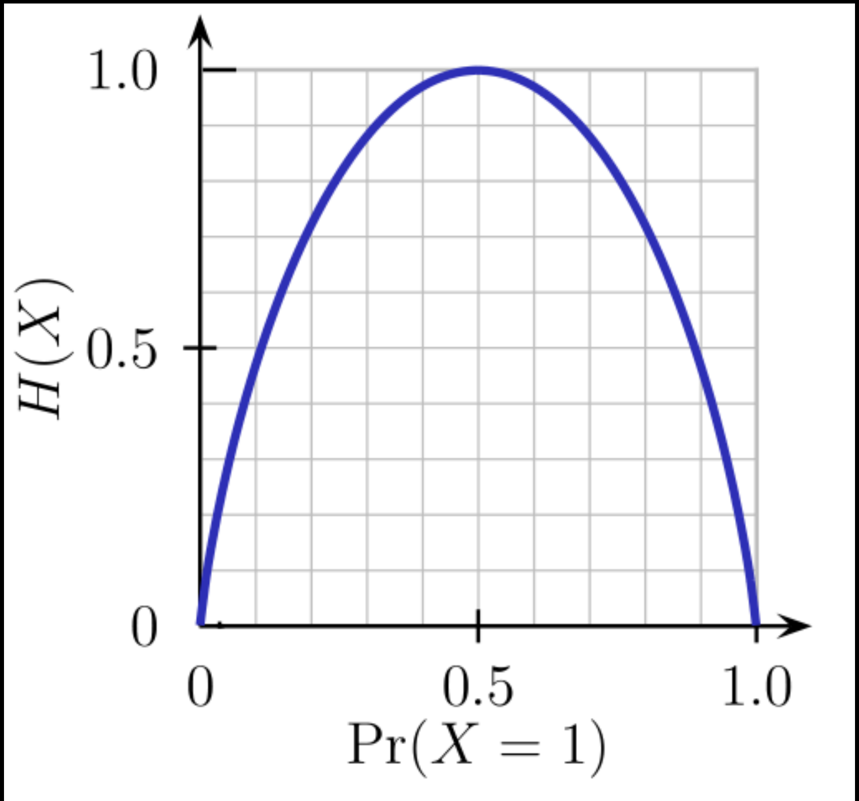
Claude Elwood Shannon

- A Mathematical Theory of Communication – 1948
- The Father of Information Theory

Entropy (Information Theory)

- *Definition*: Entropy is the measure of uncertainty in a random variable
- Measured in bits
- High entropy implies less predictability
- Provides a limit on the best possible lossless compression of any transmitted data

Entropy (Information Theory)



Let's start with a fair coin flip

- Heads and tails are equally likely
- Entropy of one flip is one bit
- Entropy of two flips is two bits

Now suppose the coin always lands on tails.

- How predictable is this?

Entropy (Information Theory)

Information Content

$$h(a_i) = \log_2 \frac{1}{w_i}$$

Shannon's Entropy

$$H(X) = \sum_{w_i > 0} w_i h(a_i) = \sum_{w_i > 0} w_i \log_2 \frac{1}{w_i} = -\sum_{w_i > 0} w_i \log_2 w_i$$

Example

Symbol (a_i)	A	B	C	D	Sum
Weights (w_i)	0.50	0.25	0.15	0.10	1
Information Content (in bits) ($-\log_2 w_i$)	1	2	2.737	3.322	
Entropy ($-w_i \log_2 w_i$)	0.5	0.5	0.411	0.332	$H(X) = 1.743$

Redundancy (Information Theory)

- *Definition*: Number of bits in the transmitted data minus its entropy
- Wasted “space” when transmitting data
- Compression reduces redundancy

The Existence of Large 2-Dimensional Crosswords

- The redundancy of a language is related to the existence of crossword puzzles
- Zero redundancy is trivial
- If the redundancy is too high the language imposes too many constraints for large crosswords to be possible

The Existence of Large 2-Dimensional Crosswords

- A more detailed analysis shows that large 2-dimensional crossword puzzles are only possible when the redundancy is less than 50%.
- If the redundancy is less than 33%, 3-dimensional crossword puzzles should be possible, etc.

The Existence of Large 2-Dimensional Crosswords

- Edgar Gilbert is an American coding theorist and longtime researcher at Bell Laboratories
- Motivated by Shannon's assertions he estimated the entropy of English text to be 41.5% when eliminating words of length 1 and 2
- Infinitely large 2-dimensional crosswords are possible to construct, but 3-dimensional crosswords are not

References

- Crossword History - *www.crosswordtournament.com/more/wynne.html*
- Recursive Backtracking - *www.cis.upenn.edu/~matuszek/cit594-2002/pages/backtracking.html*
- Claude Shannon and Information Theory - *Wikipedia*
- Crossword Puzzles and Shannon - *IEEE Information Theory Society Newsletter*, Vol. 51, No. 3, September 2001

This presentation discusses an algorithm developed for Bobo Strategy, Inc. as part of employment. The ideas within are presented with the permission of Bobo Strategy, Inc. in this academic setting.