Trasformata di Burrows-Wheeler (BWT)

Permuta una stringa s per ottenere sequenze lunghe di simboli uguali.

Codifica

1. si costruisce la matrice $|s| \times |s|$ di tutte le rotazioni cicliche di s;
2. si ordinano alfabeticamente le righe;
3. si restituisce l'ultima colonna L e l'indice i della parola originale nella matrice ordinata.

Decodifica

1. si ordina L per ottenere la prima colonna F della matrice ordinata;
2. la j-esima occorrenza di σ in F corrisponde alla j-esima occorrenza di σ in L;
3. partendo da $j = i$, si manda in output $F[j]$ e si aggiorna j all'indice della riga collegata, ripetendo finché non sono stati emessi $|L|$ simboli (ovvero la stringa s).

La trasformata non è surgettiva, quindi non tutte le stringhe si possono decodificare. In particolare la stringa codificata non può iniziare con il simbolo più piccolo, altrimenti la permutazione ha almeno due parti $(0)(\cdots)$; lo stesso vale per stringhe che terminano con il simbolo più grande.

Perché funziona?

Ordinando le righe si raggruppano quelle che iniziano con la stessa parola; nell'ultima colonna si trova la lettera che le precede, che spesso sarà la stessa — lettere uguali finiscono accanto in L perché provengono da contesti simili. Per esempio, su un testo inglese ci sarà un gruppo consistente di righe che hanno *he* nelle prime posizioni e t nell'ultima.

Questo significa che la trasformata crea maggiore località se il testo è *bilanciato*, ovvero se

$$\forall a \in \Sigma \,.\, \forall u, v \in s \text{ t.c. } |u| = |v| \;.\; \left||u|_a - |v|_a\right| \leq 1.$$

Una parola è detta circolarmente bilanciata se le sue rotazioni cicliche sono bilanciate.

L'approccio suggerito da Burrows e Wheeler è BWT $\to$ MTF $\to$ Huffman.