Compressie

Zonder dat je het merkt, heb je regelmatig te maken met gecomprimeerde computerbestanden. Daardoor passen er heel veel nummers op een iPod en past een hele film op een dvd. We bekijken in deze opgave een vereenvoudigde compressiemethode.

De rij 000000001111110000001 kan met deze methode gecomprimeerd worden tot 80616011. In de nieuwe rij wordt elk teken voorafgegaan door het aantal keren dat het voorkomt:
8 keer een 0, 6 keer een 1, 6 keer een 0 en nog één 1.
De oorspronkelijke rij van 21 tekens is zo teruggebracht tot 8 tekens. Als het aantal tekens na compressie kleiner is geworden, noemen we de compressie voordelig.
Voor deze rij is de compressieratio 0,62. De compressieratio kunnen we als volgt berekenen:

$$compressieratio = \frac{aantal\,tekens\,voor\,compressie - aantal\,tekens\,na\,compressie}{aantal\,tekens\,voor\,compressie}$$

Bekijk de rij 001100110011000111000. Op deze rij wordt bovenstaande compressiemethode toegepast.

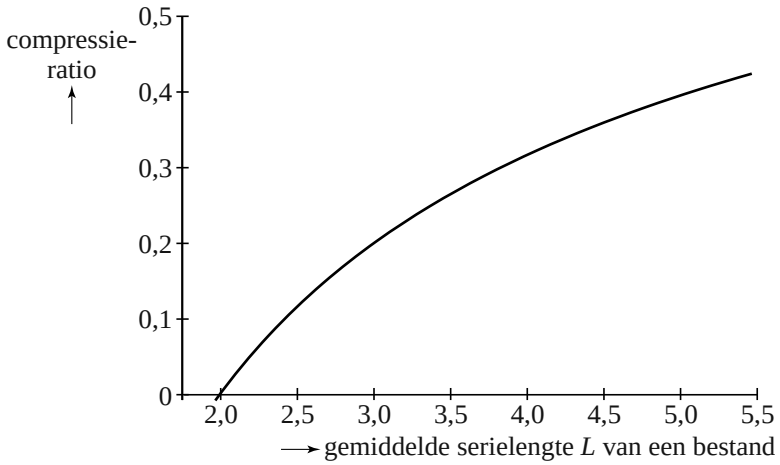
3p **19** Bereken de compressieratio van deze rij en geef aan of de compressie voordelig is.

Een rij met veel dezelfde tekens achter elkaar is natuurlijk beter te comprimeren dan een rij waarin de tekens vaak wisselen.
Bij een rij waarin de tekens vaak wisselen kan de compressieratio zelfs negatief worden. Het aantal tekens is dan na compressie groter dan ervoor.

3p **20** Onderzoek wat de kleinste waarde is die de compressieratio volgens de formule kan aannemen.

Omdat computerbestanden bestaan uit rijen tekens kunnen we hiervan ook de compressieratio berekenen.
Voor grote bestanden is er een verband tussen de compressieratio en het gemiddelde aantal gelijke tekens achter elkaar (de gemiddelde serielengte). Het verband hiertussen is in de figuur weergegeven.

figuur



Bij deze grafiek past de volgende formule:

$$compressieratio = 1 - \frac{1,4709}{L^{0,561}}$$

Hierin is L de gemiddelde serielengte.

De compressieratio van een bestand is 0,40.

3p **21** Bereken met de formule de gemiddelde serielengte van dit bestand. Geef je antwoord in twee decimalen nauwkeurig.

Informatici doen veel onderzoek naar de compressie van grote bestanden. Zij genereren met behulp van een toevalsgenerator rijen enen en nullen om te onderzoeken onder welke omstandigheden compressie voordelig is. Als de toevalsgenerator zo wordt ingesteld dat de kans op een 0 gelijk is aan 0,8 en de kans op een 1 gelijk is aan 0,2, dan zal er een behoorlijke kans zijn dat het begin van een rij uit 3 of meer dezelfde tekens bestaat.

4p **22** Bereken deze kans.