Practicum 1: Entropie

Het eerste practicum betreft Shannon entropie als maat voor de hoeveelheid informatie van een informatiebron. Het is de bedoeling dat je een computerprogramma schrijft dat de Shannon-entropie berekent van een tekstdocument (in UTF-8 formaat) dat je beschouwt als een informatiebron zonder geheugen of een informatiebron met geheugen *N*. De programmeeropdracht wordt uitgevoerd in Python vertrekkend van een template die beschikbaar wordt gesteld op Ufora. Daarna analyseer je drie tekstdocumenten en bespreek je de entropie van deze documenten. Meer bepaald wordt het volgende verwacht:

1. Beschouw drie tekstdocumenten naar keuze in UTF-8 formaat bestaande uit minstens 10.000 karakters. Selecteer verschillende types tekstdocumenten zoals b.v. een wetenschappelijk artikel, een stuk proza, een nieuwsbericht, een kinderverhaal, een gedicht, enz. Elk document is in eenzelfde taal geschreven, b.v. Nederlands, Engels, Frans.
2. Codeer een computerprogramma dat de entropie van een tekstdocument berekent. Vul hiervoor de functie

   ```
   get_entropy(data: list, memory: int = 0) -> float
   ```

   aan in de template (`template_lab1.py`) die beschikbaar wordt gesteld op Ufora. Hierbij wordt het document als een informatiebron zonder geheugen (d.i. H(X)) versus een informatiebron met geheugen *N* beschouwd (d.i. $H(X_N | X_0 , X_1 , \dots , X_{N-1})$. Gebruik de karakters uit het document als symbolen.
3. Stel een verslag op met de volgende elementen:
   a. Bespreek de drie tekstdocumenten qua type en aantal karakters, en visualiseer de distributie van de karakters in deze tekstbestanden.
   b. Vergelijk de distributie van karakters met de gemiddelde distributie van karakters van de gekozen taal. (Je kan gemiddelde distributies per taal terugvinden online.) Bespreek de verschillen voor elk van de drie documenten met de gemiddelde distributie. Citeer ook waar je de gemiddelde distributie gevonden hebt.
   c. Visualiseer de entropie van de drie tekstdocumenten als functie van het geheugen *N* waarbij je *N* varieert van 0 (geen geheugen) tot 10 (of tot wanneer de entropie convergeert of het computationeel onmogelijk wordt).
   d. Bespreek en verklaar het verloop van entropie als functie van het geheugen.
   e. Analyseer en bespreek de verschillen in entropie voor de verschillende documenten. Waarom heeft het ene document een hogere/lagere entropie i.v.m. de andere documenten? Waarom stijgt of daalt de entropie sneller voor het ene document i.v.m. de andere documenten?
   f. Comprimeer nu de tekstbestanden en bereken de entropie opnieuw en vergelijk met de entropie van de niet-gecomprimeerde tekstbestanden. (Je mag hiervoor het geheugen op 0 houden.) Gebruik 7zip om je bestanden comprimeren. Dit kan gebeuren met het volgende commando

      ```
      7z a {output}.7z {text_naam}.txt -mx=9
      ```

Je zal merken dat Python klaagt wanneer je het gecomprimeerde bestand probeert in te lezen aangezien het niet-bestaande utf-8 karakters bevat. Dit kan je eenvoudig oplossen door het bestand bytegewijs in te lezen m.b.v. `open("{output}.7z", "rb")`. Lees het originele tekstbestand dan ook bytegewijs in voor deze vraag, om een eerlijke vergelijking te kunnen maken van de entropie voor en na compressie. Merk op dat een symbool dan geen karakter meer is, maar een byte.

Je dient de code, de drie tekstbestanden, en het verslag in via Ufora. Zowel de code als het verslag wordt geëvalueerd. De code wordt gequoteerd op correctheid en volledigheid. Het verslag wordt gequoteerd op nauwkeurigheid en vorm (taalgebruik, layout, leesbare grafieken, enz.)

Indiendatum: vrijdag 21 oktober 2021, via Ufora.