Entropie in drie Engelse teksten van verschillende genres

Elias Nijs

Telecommunicatie, Informatica
Universiteit Gent

21 October 2022

Entropie in drie Engelse teksten van verschillende genres

Elias Nijs

Telecommunicatie, Informatica
Universiteit Gent

1. Introductie

In dit verslag worden de resultaten omtrent het onderzoek in verband met entropie voor het vak telecommunicatie aan de Universiteit Gent besproken.

In dit onderzoek werden drie teksten geanalyseerd met het oog op de hoeveel entropie deze bezitten. Hiervoor werd eerst een basis analyse uitgevoerd die de lengte en karater distributies bekijkt. Daarna werd de hoeveel entropie met verschillende hoeveelheden geheugen bekeken in elke tekst. Om dit te doen werd een entropie functie geschreven in python. Tot slot werd ook bekeken hoe de entropie zich gedraagt na compressie met 7zip.

De drie teksten zijn:

1. Genesis[1]
2. The C Programming Language[2]
3. Othello, the Moore of Venice[3]

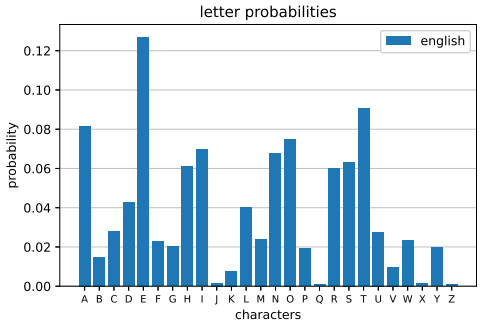
Voordat we beginnen, nog een praktische noot. Bij dit verslag hoort een jupyter notebook en een python bestand. Men kan deze samen met de teksten, afbeeldingen van de grafieken en enkele scripts terugvinden op de volgende link:

https://eliasnijs.xyz/unief/telepract1.zip

2. Basis Analyse van de teksten

We beginnen met een basis analyse van de teksten. Hier bekijken we de lengtes en karakterdistributies. Daarnaast zullen we ook de karakter distributie van de gemiddelde Engelse tekst vergelijken met degene van de onze teksten.

Helaas was voor de Engelse distributie enkel de distributie van de letters, zonder onderscheid tussen hoofd en kleine letters, te vinden.[4] Daarom, opdat we een eerlijke vergelijking zouden hebben, werden de teksten voor de vergelijking gefiltered op enkel letters.



2.1. Lengte

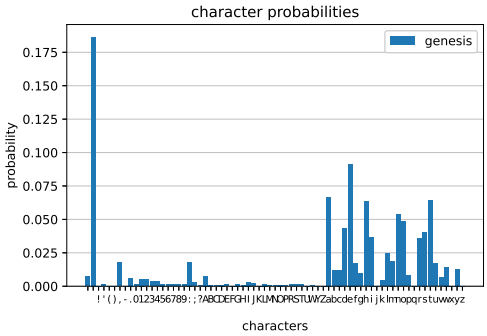
(vraag a)

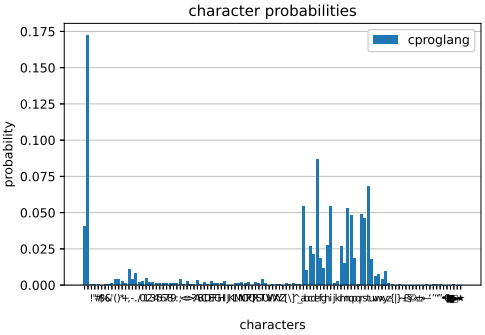
We kijken eerst naar de lengten, deze zien er als volt uit. Genesis bestaat uit *207327* karakters, The C Programming Language bestaat uit *432963* karakters en Othello, the Moore of Venice bestaat uit *153428* karakters.

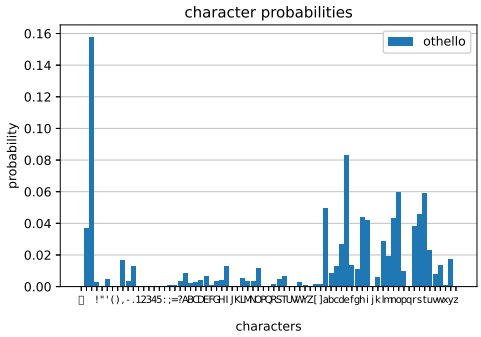
2.2. Karakter verdeling

(vraag a)

Als volgt kijken we naar de karakterverdelingen.







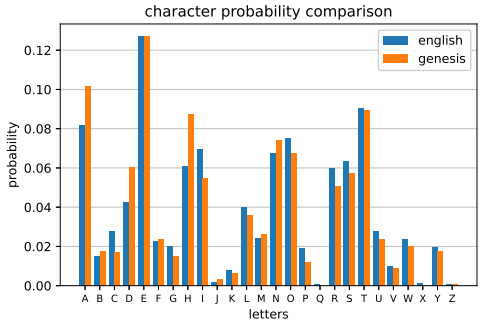
We zien dat deze verdelingen heel gelijkaardig zijn ondanks dat ze van verschillende genres zijn. We zien bijvoorbeeld dat in elke text het karakter *spatie* heel vaak voorkomt in vergelijking met andere tekens.

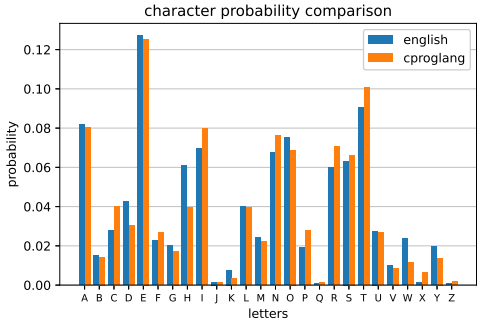
2.3. Vergelijking met Engels

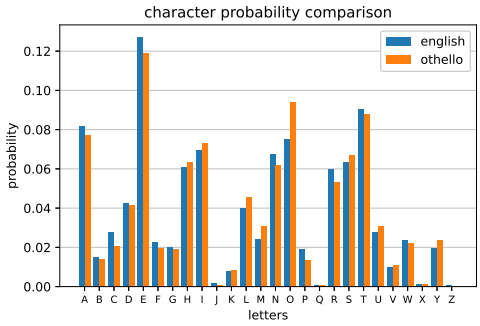
(vraag b)

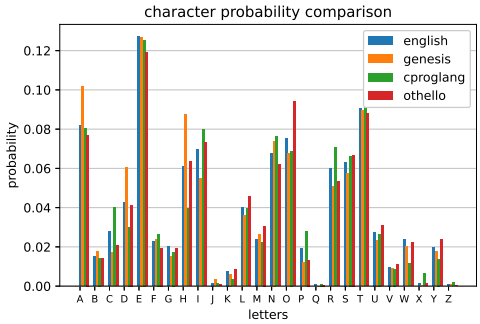
Tot slot vergelijken we de verdelingen van onze teksten met deze van Engels. Zoals eerder vermeld zullen we hiervoor onze teksten filteren tot enkel letters en geen onderscheid maken tussen hoofd en kleine letters.

Hiervoor zullen we eerst eens elke verdeling naast de verdeling van Engels plaatsen.

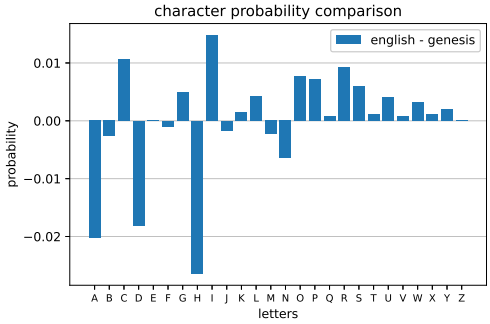


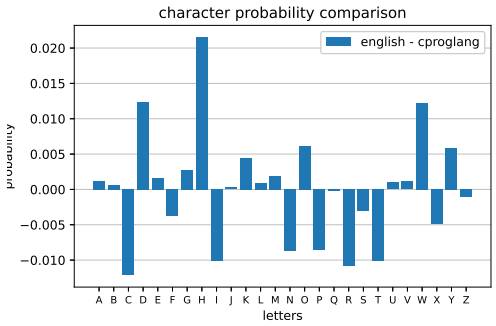


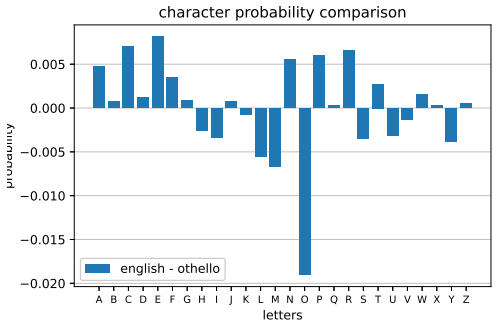


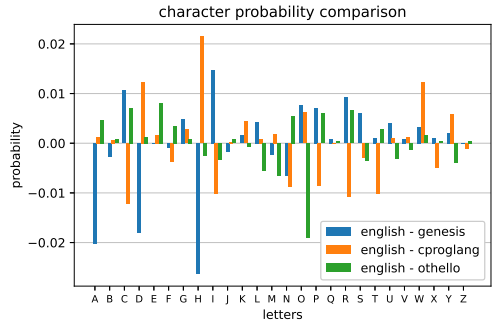


We kunen ook het verschil tussen de verdelingen en de verdeling van Engels bekijken.









Hiervan kunnen we ook de gemiddeldes bekijken. Voor Genesis komt dit neer op *2.27e-05* , voor The C Programming Language op *2.27e-05* en voor Othello op *2.27e-05*. Opmerkelijk is dat deze alle drie hetzelfde zijn.

2.4. Conclussie en vooruitblikken

We zien dat de karakterdistributies in de teksten heel gelijkaardig zijn en dat deze ook heel weinig verschillen van de verdeling van Engels. Meer nog, wanneer we het gemiddelde verschil beschouwden van de teksten tegenover Engels, bleek dit voor alle drie de teksten hetzelfde te zijn.

Vermits de karakterverdelingen van de teksten ook zeer dicht aanleunen bij Engels, vermoeden we dat de entropie van onze teksten ook zeer dicht zal aanleunen bij de gemiddelde entropie in een Engelse tekst, dewelke neerkomt op 4.11 (bij geheugen 0).

3. Entropie in de verschillende teksten

Na onze basis analyse gedaan te hebben, bekijken we nu de entropie van de teksten. Voordat we dit kunnen doen, moeten we eerst een programma schrijven die deze kan berekenen.

3.1. Berekenen van de entropie

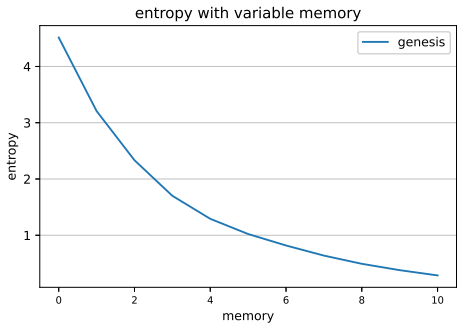
Voor het berekenen van de entropie werd een programma geschreven in python met behulp van de python library numpy.

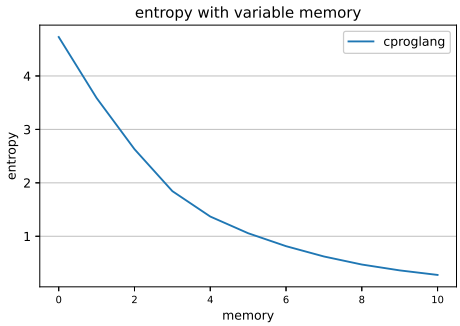
De code hiervoor kan worden teruggevonden in zowel de notebook als het uitvoerbare python bestand.

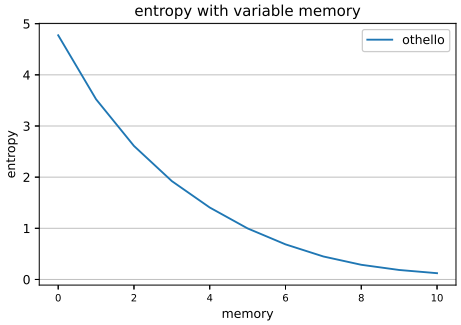
3.2. Entropie in de teksten

(vraag c)

We beelden nu voor elke tekst de entropie af met een variabel geheugen dat van 0 tot 10 varieert.







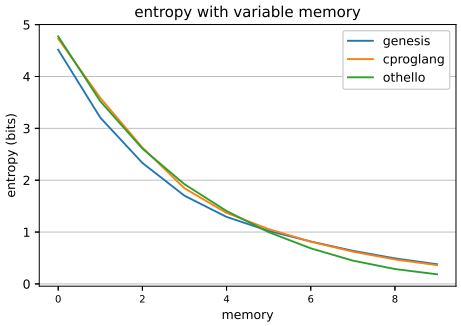
3.3. Bevindingen

(vraag d)

Het verloop van de entropie van de teksten is zeer gelijkaardig. Zoals we al voordien vermoedden hebben de teksten een entropie van iets meer dan 4.11 bij een geheugen van 0. We zien ook duidelijk dat de entropie daalt naarmate het geheugen groter wordt, meer nog, het convergeert naar nul. Als we even ons inbeelden dat we een geheugen van grootte *lengte - 1* hebben, dan is meteen duidelijk dat de hoeveelheid onzekerheid nul is, er is immers maar 1 karakter dat kan volgen. Dit geeft intuitief aan waarom de entropie naar 0 daalt.

(vraag e)

Om beter de verschillen te kunnen zien tussen de verschillende teksten, beelden we de hoeveelheden entropie af op 1 grafiek.



We zien dat de entropie zeer gelijkaardig loopt. Wat opvalt is dat de entropie in *Othello* het hoogst begint maar daarna veel sneller daalt waardoor het bij geheugen tien het laagst staat. De hoge start komt doordat de verdeling van de letters meer uniform zijn bij deze tekst dan bij de andere en er dus meer onzekerheid is. De snelle daling komt vermoedelijk doordat het taalkgebruik in *Othello* simpeler is en dus bepaalde combinaties van letters sneller voorspelbaar worden.

4. Entropie voor en na compressie in de verschillende teksten

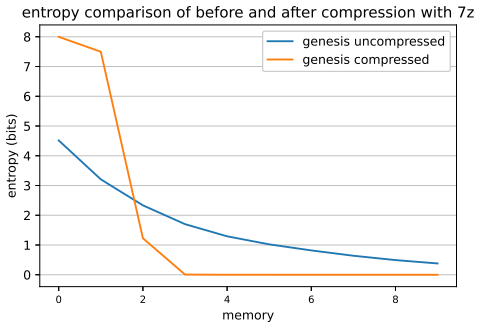
(vraag f)

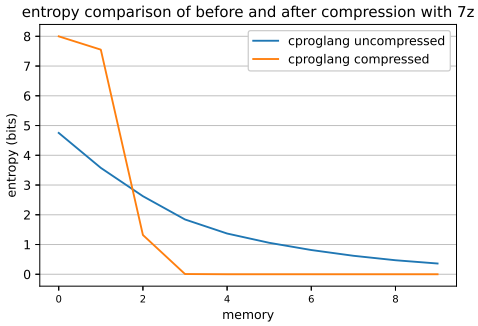
Tot slot zou het interessant zijn om de entropie te gaan vergelijken met de entropie na compressie. Voor het comprimeren passen we het volgende commando toe op onze teksten:

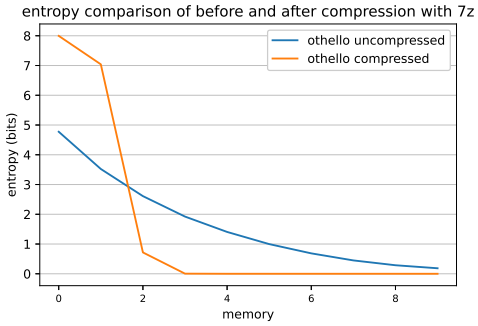
```
7z a {output}.7z {input}.txt
    -mx=9
```

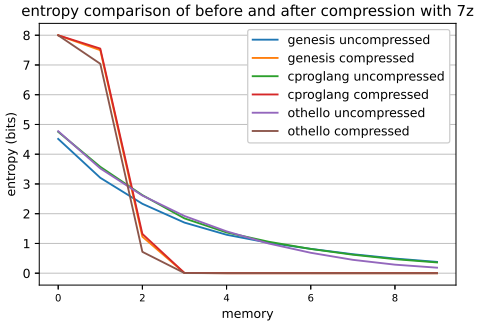
We zullen nu ook onze teksten moeten inlezen als bytes, zonder deze te interpreteren. We moeten dit doen om een eerlijke vergelijking te hebben met de gecomprimeerde teksten.

We krijgen dan volgende resultaten:









Ook hier zien we terug dat de drie teksten zich zeer gelijkaardig gedragen op het vlak van entropie. Opmerkelijk is dat de entropie met geheugen nul van een tekst na compressie telkens op 8 start en pas minder entropie heeft vanaf dat het geheugen de waarde 2 aanneemt.

5. Slotnoot

We kunnen besluiten dat de entropie'en in de drie tekst bestanden zich zeer gelijkaardig gedragen. Vermoedelijk gedragen de meeste Engelse teksten zich op deze manier.

References

1. http://www.stewarton-bibleschool.org/bible/text/genesis.txt.
2. https://archive.org/details/TheCProgrammingLanguageFirstEdition.
3. https://shakespeare.folger.edu/shakespeares-works/othello/download/.
4. https://gist.github.com/randallmorey/dea827d6f1c48374bdea0d2f5a320a16.