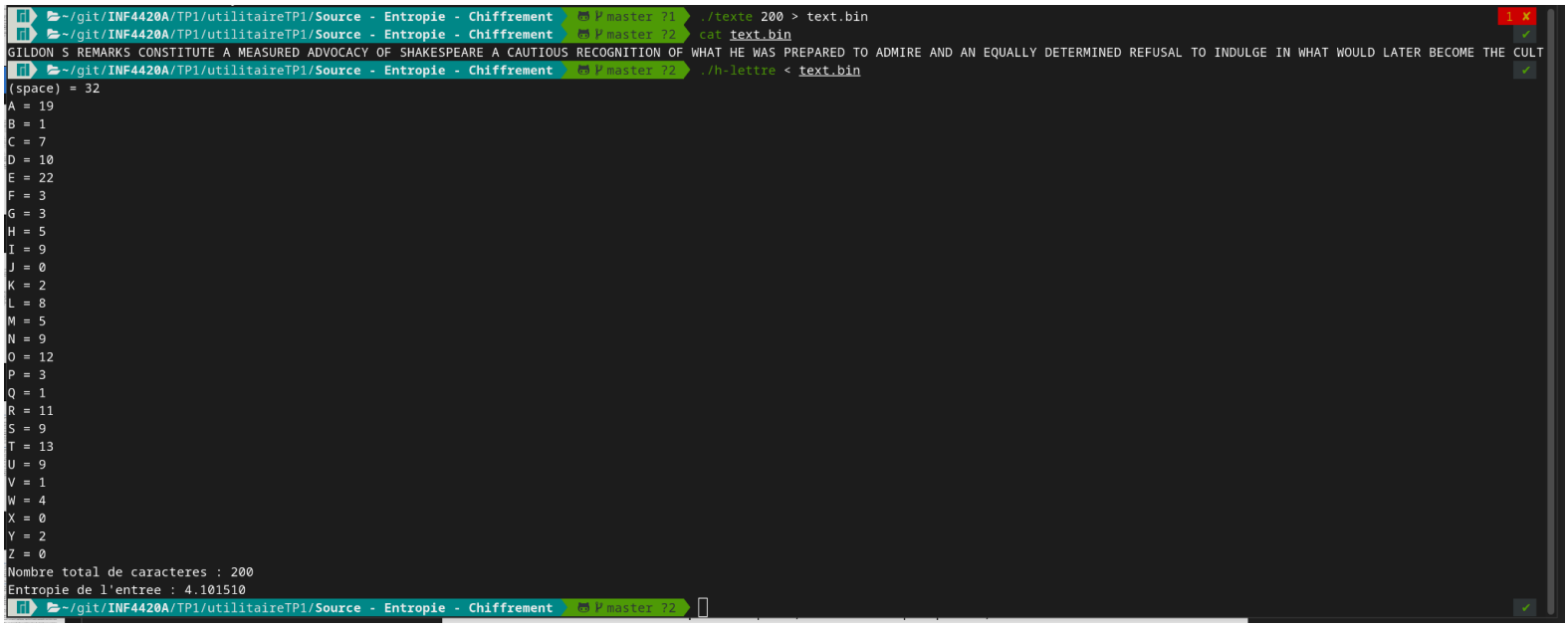
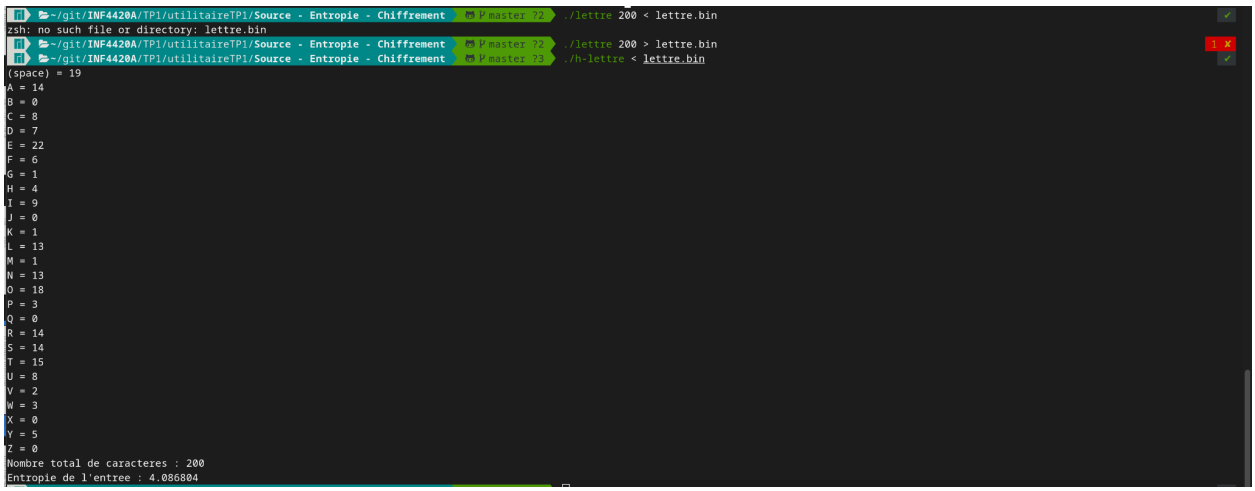
Partie A

1. L’entropie est de 4.10 comme on peut le voir sur la figure 1

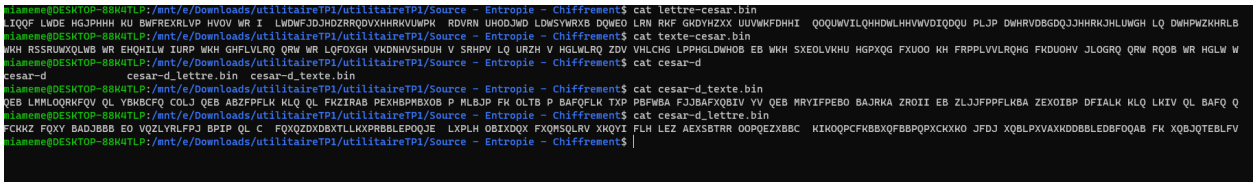


2. H(x), ici 4 est le nombre de bits minimum necessaire pour coder toute la chaine de caracteres.
3. Maths https://fr.wikipedia.org/wiki/Entropie_de_Shannon donne 4.75
4. On obtient 0.85, le facteur de compression
5. 

   La valeur n’est pas significative.
6. Dans les deux cas, les lettres sont tirées de l’alphabet anglais. Que cela soit dans un texte anglais (pour programme texte) ou bien dans tous les mots de la langue anglaise (programme lettre), les fréquences pour chacune des lettres vont être les mêmes (surtout si le mot tend vers l’infini. Ici 200 caractères est suffisant). Si la fréquence d’apparition des symboles est environ la même et le nombre de symboles total est le même (même alphabet) alors l’entropie va être très proche.

Question 2 - Histogrammes

a)



b)

c)On remarque que les histogrammes sont les mêmes pour le fichier lettre.bin avec le programme cesar et avec le programme cesar-d. De même avec le fichier texte.bin. Le nombre d'occurrence est le même pour des lettres différentes ce qui fait du sens car le chiffrement de césar fait seulement décaler les lettres vers la droite donc on va avoir les mêmes fréquences mais sur des lettres décalées. De manière générale les 4 histogrammes sont très identiques ce qui fait du sens car la langue de référence pour les quatres fichiers est toujours l'anglais.

Dans le cas où les fréquences seraient comptabilisées sur deux lettres à la fois, les fréquences seraient alors bien plus faibles. Dans notre cas, si l'on parle de la fréquence du "th" qui existe beaucoup en anglais, on va remarquer que la fréquence est relativement grande dans le fichier texte.bin que dans le fichier lettre.bin. Cela est dû au fait que dans le fichier lettre.bin ce sont des lettres pris aléatoirement (il n'y a **pas** d'ordre)donc on a peu de chance d'avoir t et h d'affilé plusieurs fois. Par contre, les lettres qui reviennent souvent en anglais par exemple e, t ou o vont plus souvent se retrouver collée. Il y aura donc une fréquence plus élevée de "ee" pour le fichier lettre.bin.

d) Cela faciliterait effectivement le déchiffrement d'un vrai texte (texte français avec un ordre respectant la langue française) car nous pourrions identifier les diagrammes les plus présents au lieu des lettres les plus présentes. En anglais on vient de parler du "th" qui revenait souvent. Cela est aussi le cas de "he" ou de "in" par exemple. On pourrait alors voir les fréquences de ces diagrammes et déterminer plus facilement la langue du texte.

Dans le cas du fichier lettre.bin il n'y aurait aucune différence puisque les lettres sont tirées aléatoirement donc il n'y aurait pas de diagramme de n-lettres identifiables dans le tas de lettres collées. Pour augmenter la facilité du déchiffrement d'un texte aléatoire (créé avec le programme lettre par exemple) alors il faudrait avec les données des n-diagrammes pour un texte aléatoire tiré de la probabilité qu'une lettre déterminée soit générée correspond à sa fréquence dans la langue anglaise. Par exemple, au lieu de regarder la fréquence de "th" il faudrait regarder la fréquence de "ee" (et autres fréquences pas seulement celle-là).