# IFT6285 - Devoir 1

Louis-Vincent Poellhuber - Youcef Barkat

3 octobre 2023

---

## a)

i **Moyenne, minimum et maximum de perplexité** :

(a) Moyenne : 412.93
(b) Minimum : 28.73
(c) Maximum : 21411.12

ii **Formule de perplexité utilisée** :

$$\text{perplexité}(phrase) = 10^{-\frac{-\text{score}(phrase)}{\text{mots}}} \tag{1}$$

## b)

| | Number of slices trained on | Average perplexity | CPU Time (s) | Disk space (kB) |
|---|---|---|---|---|
| **Bimodel 1** | 1 | 519.214705 | 10.4654 | 53543 |
| **Bimodel 2** | 2 | 459.769543 | 19.9174 | 87435 |
| **Bimodel 3** | 3 | 452.838032 | 24.4909 | 115548 |
| **Bimodel 4** | 4 | 445.563377 | 29.6741 | 140910 |
| **Bimodel 5** | 5 | 433.426625 | 39.1275 | 164432 |
| **Bimodel 6** | 6 | 432.019695 | 41.5106 | 186094 |
| **Bimodel 7** | 7 | 418.989377 | 47.1417 | 206837 |
| **Bimodel 8** | 8 | 416.189924 | 51.1339 | 226067 |
| **Bimodel 9** | 9 | 412.933038 | 56.4883 | 244980 |

FIGURE 1 – Impact du Nombre de Tranches d'entraînement sur les Performances du Modèle, l'espace disque et le Temps d'entraînement

La Figure 1 illustre l'influence de la variation du nombre de tranches d'entraînement sur les performances du modèle testé, ses besoins en espace disque et le temps nécessaire à l'entraînement. On y observe que plus le nombre de tranches augmente, plus la perplexité diminue, mais aussi plus le temps de calcul et l'espace sur le disque est plus grand.

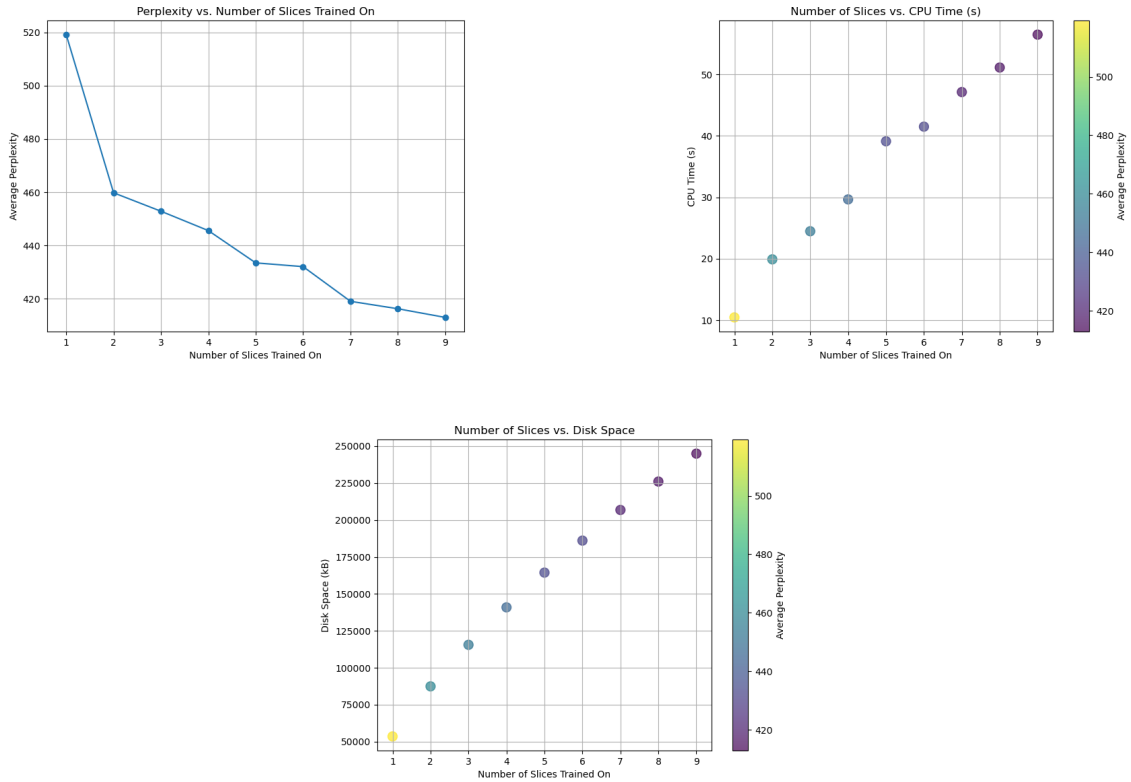


FIGURE 2 – De gauche à droite, haut en bas : Perplexité VS nombre de tranches, temps de calcul VS nombre de tranches, espace sur le disque VS nombre de tranches. Ces figures utilisent les mêmes données que celles dans la Figure 1.

La figure 2 révèle qu'à mesure que le nombre de tranches utilisées pour l'entraînement augmente, on observe une tendance générale à la hausse de la perplexité moyenne, du temps CPU et de l'utilisation de l'espace disque. Cela suggère un compromis entre l'amélioration de la compréhension du modèle (moins de perplexité) et les besoins en ressources. Bien que l'entraînement sur davantage de données puisse améliorer les performances du modèle, il exige davantage de ressources informatiques et d'espace de stockage.

## c)

| | N-gramme | Average perplexity | CPU Time (s) | Disk space (kB) | Loading Time (s) |
|---|---|---|---|---|---|
| **Bimodel** | 2 | 412.933038 | 56.4883 | 244980 | 8.3803 |
| **Trimodel** | 3 | 317.593202 | 154.0450 | 1217636 | 41.4838 |
| **Quadmodel** | 4 | 303.863015 | 485.1350 | 3235823 | 110.1083 |
| **Pentamodel** | 5 | 301.977247 | 843.0900 | 5970728 | 189.1594 |

FIGURE 3 – Statistiques de performance selon la complexité du modèle.

Pour choisir le meilleur modèle, nous avons mené des expériences en faisant varier le paramètre $n$ dans les modèles n-grammes. Nous avons augmenté progressivement $n$ tout en surveillant de près les métriques de performance et les ressources informatiques nécessaires à l'exécution et au chargement du modèle. Les résultats de ces expériences sont présentés dans le tableau de la Figure 3. Ainsi, dans le contexte du choix du modèle optimal parmi les options disponibles, le Trimodèle se démarque comme le choix préféré. Bien qu'il n'obtienne pas le score de perplexité absolu le plus bas par rapport au Pentamodèle, il parvient à établir un équilibre louable entre les performances du modèle et l'efficacité des ressources. Le Trimodèle offre un compromis favorable en affichant une perplexité nettement plus faible (317,59) que le Bimodèle, tout en étant considérablement plus efficace en termes de temps CPU (154,05 secondes) et d'utilisation de l'espace disque (1 217 636 ko). De plus, le Trimodèle présente un temps de chargement relativement rapide (41,48 secondes), ce qui renforce sa praticité pour le déploiement et une utilisation en temps réel. En résumé, les performances impressionnantes du Trimodèle, associées à une utilisation efficace des ressources et à des temps de chargement rapides, en font le choix pragmatique pour la création d'une solution robuste et pratique de modèle linguistique.

# d)

Pour cette section, tout le code a été exécuté sur Python.

i **Nombre de lignes** : 6
**Temps de calcu**l : 20.2s
Il y a 2 756 013 phrases dans le corpus. Les 6 lignes servent à concaténer les différentes slices.

ii **Nombre de lignes** : 3
**Temps de calcu**l : 2.5s
Ceci utilise le code précédent. Il n'existe pas de phrases ayant été vues 4 fois, mais il y en a une seule qui a été vue 3 fois :
"*American negotiator Watson said the Bush administration is planning probably four more meetings in the Major Economies series before a " leaders ' meeting " in mid-2008 presents a final outcome .*"

iii **Nombre de lignes** : 5
**Temps de calcu**l : 1.0s
Ceci utilise le code en i. Oui, il existe 3 182 phrases ayant exactement 20 caractères et 16 100 phrases ayant exactement 120 caractères.

iv **Nombre de lignes** : 1
**Temps de calcu**l : 0.0s
Ceci utilise le code précédent. La phrase la plus courte est de 2 caractères, alors que la phrase la plus longue est de 9 572 caractères.

v **Nombre de lignes** : 3
**Temps de calcu**l : 0.0s
Ceci utilise le code en i. Les 100 phrases les plus fréquentes ayant moins de 21 caractères sont sauvegardées dans le fichier **frequent_short_sentences.csv**.

vi **Nombre de lignes** : 18
**Temps de calcu**l : 65.9s
Voici les 6 mots les plus fréquents après **continue to** et leur nombre d'occurences :

| Mots | Occurences |
|---|---|
| be | 621 |
| work | 209 |
| do | 165 |
| grow | 139 |
| rise | 126 |
| have | 114 |

De plus, la chaîne de 6 mots la plus fréquente après **continue to** correspond aux quatre chaînes suivantes, avec 3 occurences chaque :

(a) "serve as chairman of the board"
(b) "act to strengthen and stabilize our"
(c) "complement fiscal stimulus with strong government"
(d) "show restraint in dealing with those"

vii **Nombre de lignes** : 13
**Temps de calcu**l : 9.8s
Les 10 premiers mots les plus fréquents dans une phrase sont : "*A version of this article appeared in print on March*", apparu 113 fois.

viii **Nombre de lignes** : 13
**Temps de calcu**l : 31.6s
Voici les 5 mots de 4 lettres majuscules les plus fréquents. La liste complète se trouve dans le fichier **most_common_4maj_words.csv**. Veuillez noter que nous avons enlevé la ponctuation pour trouver les mots de 4 lettres majuscules.

| Mots | Occurences |
|---|---|
| YORK | 5 512 |
| NATO | 3 133 |
| NASA | 2 238 |
| NYSE | 1 720 |
| AIDS | 1 437 |