IFT6285 - Devoir 3

Louis-Vincent Poellhuber - Youcef Barkat

10 novembre 2023

1 Temps d'Entraînement par Rapport au Nombre de Phrases Traitées

Configuration de la machine : Devoir complété sur un ordinateur Microsoft Surface Laptop 3. Les spécifications de cet ordinateur sont les suivantes :

— Processeur : Intel Core i7 (4 CORES)

— Mémoire vive (RAM) : 16 Go

Résultats : Dans cette section, nous présentons une figure illustrant l'impact de l'entraînement d'un modèle Word2Vec en utilisant une approche basée sur les tranches. Nous faisons varier le nombre de phrases traitées en termes de tranches pour comprendre les exigences en temps pour différentes configurations.

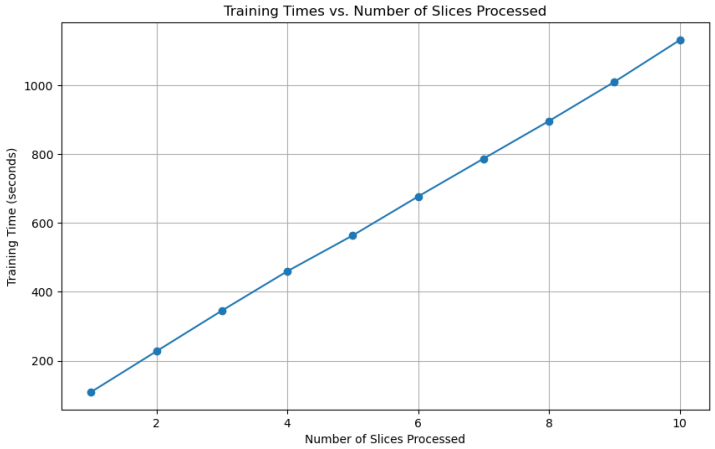


FIGURE 1 – Temps d'Entraînement par Rapport au Nombre de Tranches Traitées

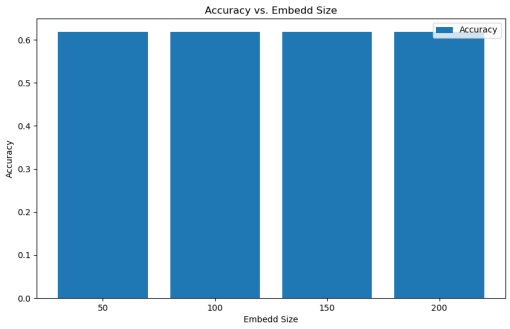
Les résultats montrent une augmentation significative du temps d'entraînement à mesure que le nombre de tranches traitées augmente. La relation entre les deux est clairement linéaire : ceci nous permet de facilement estimer le temps que prendra l'entraînement du modèle avec un certain nombre de tranches. Cette tendance doit être prise en compte lors de la configuration de l'entraînement du modèle Word2Vec.

2 Paramètres Étudiés

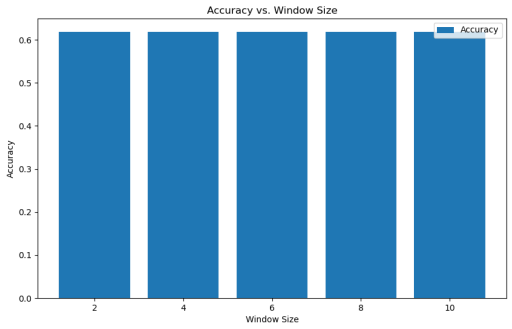
Les quatre meta-paramètres étudiés sont les suivants :

1. **Taille du Vecteur (Vector Size)** : La taille du vecteur détermine la dimension de l'espace vectoriel dans lequel les mots sont représentés. Nous avons testé différentes tailles, allant de 50 à 300.
2. **Taille de la Fenêtre (Window Size)** : La taille de la fenêtre définit le contexte dans lequel les mots sont considérés comme voisins. Nous avons varié la taille de la fenêtre de 2 à 10.
3. **Taille de l'Ensemble d'Entraînement (Training Data Size)** : La taille de l'ensemble d'entraînement influence la quantité de données utilisées pour former le modèle. Nous avons expérimenté avec des ensembles d'entraînement de tailles différentes.
4. **Paramètre min_count** : Le paramètre min_count indique le nombre minimum d'occurrences d'un mot pour qu'il soit pris en compte dans le modèle. Nous avons ajusté ce paramètre pour contrôler le vocabulaire du modèle.

2.1 Figures

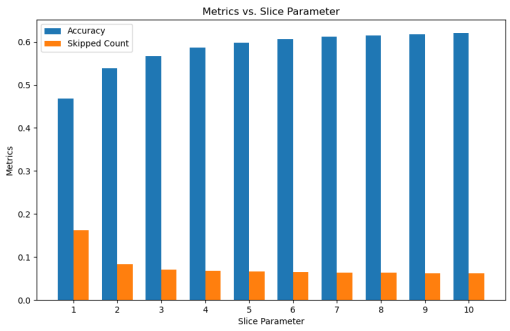


(a) Influence de la taille du vecteur sur la performance.

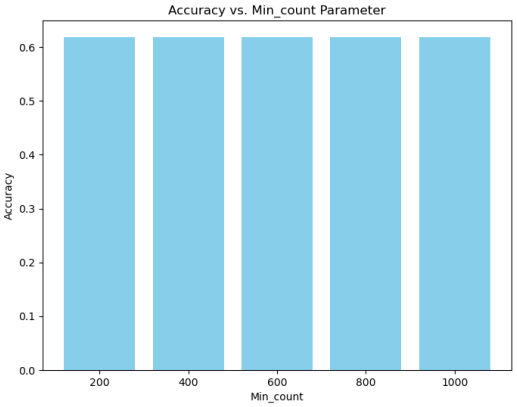


(b) Influence de la taille de la fenêtre sur la performance.

FIGURE 2 – Figures 1 et 2 : Résultats des variations de la taille du vecteur et de la fenêtre.



(a) Influence de la taille de l'ensemble d'entraînement sur la performance.



(b) Influence du paramètre min_count sur la performance.

FIGURE 3 – Figures 3 et 4 : Résultats des variations de la taille de l'ensemble d'entraînement et du paramètre min_count.

2.2 Résultats

Nous avons constaté que la modification des paramètres de taille du vecteur, de la fenêtre et du paramètre min_count n'a pas eu d'impact significatif sur la performance du modèle sur le benchmark TOEFL. Cependant, la variation de la taille de l'ensemble d'entraînement a montré une amélioration de la précision du modèle, notamment avec des ensembles d'entraînement plus importants.

Ainsi, nous avons choisi d'aussi étudier le temps d'entraînement selon chaque hyperparamètre visé, pour voir si certains mènent à des temps d'entraînemen plus longs que d'aures. Comme on peut le voir sur la Figure 4, minimum_count a un gros impact lorsque le minimum est petit. Les deux autres hyperparamètres ont moins de variation dans leur temps d'entraînement.

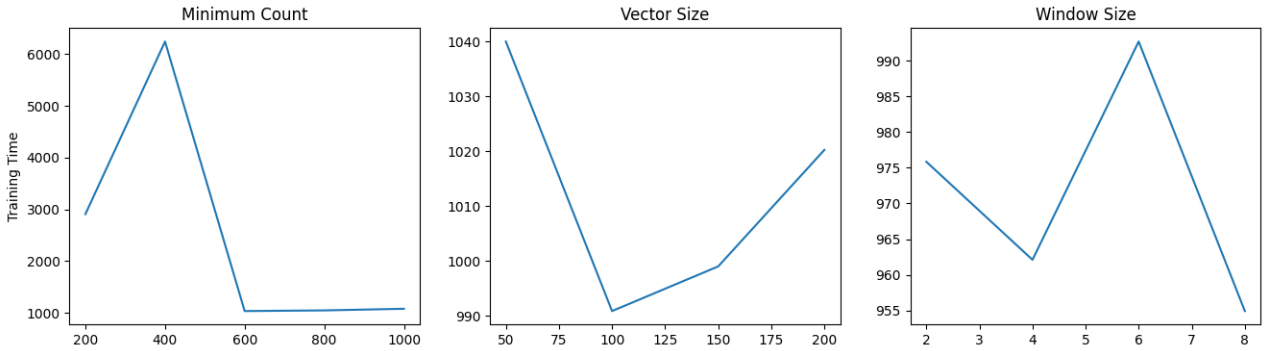


FIGURE 4 – Temps d'entraînement selon l'hyperparamètre visé

3 Paramètres d'Entraînement

Le modèle Word2Vec a été entraîné en utilisant les paramètres par défaut. Comme on peut voir sur la Figure 4, le choix de la taille du vecteur et de la taille de la fenêtre n'a peu d'impact sur le temps d'entraînement.

- Taille de la fenêtre (window size) : 5
- Dimension du vecteur (vector size) : 200
- Algorithme de l'optimiseur : Skip-gram (par défaut)
- Nombre de négatifs (negative samples) : 5
- Min_count : 10

Ensemble de Données : Le modèle a été entraîné sur l'ensemble de données 1 Billion Word Short Benchmark Dataset.

Calcul des Mots Proches : Pour répondre à la question spécifiée dans le rapport, nous avons calculé les 10 mots les plus proches pour une liste de mots donnée. Le résultat a été stocké dans un fichier texte au format demandé.

4 Pourcentage de synonymes

Nous estimerions le nombre de synonymes à 34.8%. Nous avons estimé cette valeur en comparant les différences de distances entre chaque voisins du fichier. Si la différence entre deux voisins est moins que 0.02, ces mots sont considérés comme des synonymes. Ce seuil a été choisi arbitrairement en observant le fichier des voisins. Une approche moins heuristique aurait été d'obtenir une liste de synonymes pour tous les mots du dictionnaire anglais et comparer nos voisis à chaque synonyme de ce dictionnaire.