Rapport TP Word2Vec vs GloVe

— Qian DU

En raison des limitations de taille des fichiers, les datasets et modèles entraînés ne sont pas inclus dans ce document.
Le code complet du projet est disponible sur GitHub à l'adresse suivante :
https://github.com/VanessaQianDU0320/TP1-Word2vec-avec-gensim/tree/main

1. Introduction

Dans la deuxième partie du projet de Fouille de données textuelles, nous poursuivons le travail réalisé dans le TP1 en introduisant un second modèle d'embeddings : GloVe (Global Vectors for Word Representation).

L'objectif principal de cette partie est de comparer les représentations vectorielles obtenues avec Word2Vec et celles produites par GloVe, en utilisant le même corpus que dans le TP1, à savoir le dataset Movies Overview.

L'implémentation de GloVe s'appuie sur le notebook fourni par l'enseignant via le dépôt GitHub(https://github.com/cr0wley-zz/Embeddings/blob/main/GloVe/Glove.ipynb?short_path=3f195ff). Nous avons adapté ce code afin de l'intégrer dans l'architecture du projet développé lors du TP1.

Ainsi, cette seconde partie s'inscrit dans la continuité du TP1, l'objectif est d'analyser les différences théoriques et empiriques entre un modèle prédictif (Word2Vec) et un modèle basé sur des statistiques globales (GloVe).

2. Description du modèle Word2Vec

2.1 Principes théoriques

Contrairement à Word2Vec, qui est un modèle prédictif basé sur un contexte local, GloVe adopte une approche fondée sur les statistiques globales du corpus.

Le principe fondamental de GloVe consiste à exploiter la fréquence de cooccurrence des mots dans l'ensemble du corpus. Autrement dit, le modèle analyse combien de fois deux mots apparaissent ensemble dans une fenêtre de contexte donnée. Ces informations statistiques sont ensuite utilisées pour apprendre des vecteurs numériques capables de refléter la structure globale du langage.

L'idée centrale est que les relations sémantiques entre les mots peuvent être capturées à partir des rapports de cooccurrence. Par exemple, si deux mots apparaissent fréquemment dans des contextes similaires, leurs vecteurs devraient être proches dans l'espace vectoriel.

Contrairement à Word2Vec, qui apprend à prédire un mot à partir de son contexte (ou inversement), GloVe cherche à reconstruire directement les relations statistiques observées

dans le corpus. Il s'agit donc d'un modèle basé sur une approximation des données globales plutôt que sur une tâche de prédiction locale.

Ainsi, les deux modèles poursuivent le même objectif — apprendre des représentations vectorielles pertinentes — mais reposent sur des mécanismes d'apprentissage fondamentalement différents.

2.2 Implémentation et paramètres choisis

Le modèle GloVe a été implémenté en Python à partir du notebook GitHub fourni par l'enseignant, adapté à l'architecture du TP1.

Dans un premier temps, nous avons testé plusieurs configurations similaires à celles utilisées pour Word2Vec. Cependant, la taille du vocabulaire et le nombre très élevé de cooccurrences rendaient l'entraînement extrêmement long sur un ordinateur personnel.

En effet :

vocab_size initial élevé

plus de 2 millions de paires de cooccurrence

entraînement par mise à jour individuelle (SGD simple)

Afin de rendre l'exécution possible dans un temps raisonnable, les paramètres finaux retenus sont :

- **embedding_size** = 100
- **iterations** = 3
- **learning_rate** = 0.001
- **max_pairs** = 30000
- **min_count** = 50

2.3 Résultats attendus

Théoriquement, nous attendons que :

- Les mots apparaissant fréquemment ensemble soient proches dans l'espace vectoriel
- Les relations thématiques soient capturées
- Des clusters sémantiques émergent (ex : amour, guerre, royauté)

Cependant, contrairement à Word2Vec :

- GloVe nécessite une couverture statistique importante
- Une convergence insuffisante peut entraîner des embeddings moins structurés

Nous anticipons donc que les performances de GloVe peuvent être sensibles :

- à la taille du corpus
- au nombre de paires de cooccurrence utilisées

- au nombre d'itérations

3. Mise en pratique et analyse des performances

3.1 Description des données

Le dataset utilisé est identique à celui du TP1 :

Dataset : Movies

Fichier : movies_metadata.csv

Texte utilisé : overview (description des films)

Corpus narratif contenant des thèmes variés (guerre, amour, famille, crime, etc.)

Le même prétraitement a été appliqué afin d'assurer une comparaison équitable entre Word2Vec et GloVe.

3.2 Architecture du projet

Le projet a été enrichi avec un nouveau fichier :

movies_GloVe.py

Ce fichier contient :

- Lecture des données
- Construction du vocabulaire
- Construction de la matrice de cooccurrence
- Entraînement du modèle GloVe
- Sauvegarde des embeddings

visualisation.py

Ce fichier contient les fonctions permettant de :

- Générer une réduction de dimension (t-SNE)
- Visualiser les embeddings Word2Vec et GloVe
- Sauvegarder automatiquement les figures dans le dossier rapport/

Le fichier **main.py** orchestre l'ensemble du processus

- Chargement ou entraînement des modèles
- Lancement des tests qualitatifs
- Appel des fonctions de visualisation

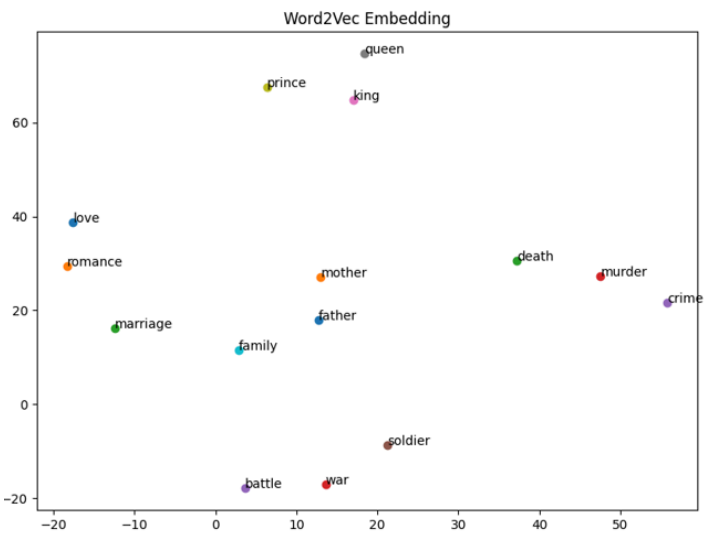
3.3 Analyse des performances

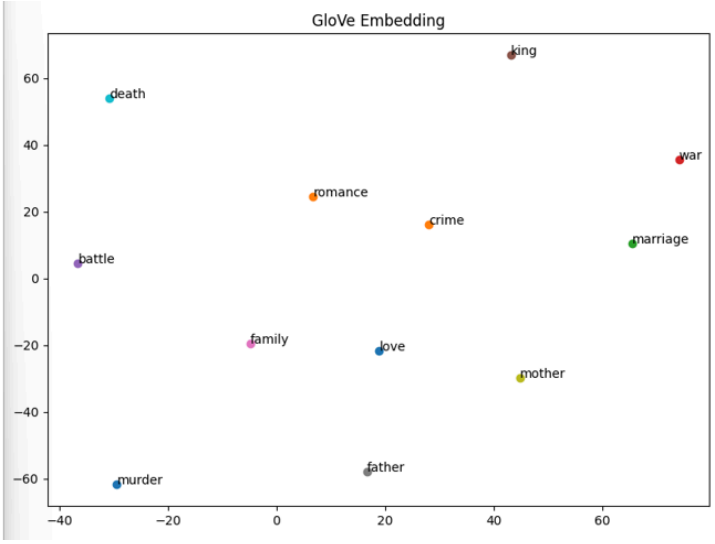
Résultats qualitatifs

Les résultats obtenus avec Word2Vec montrent une forte cohérence sémantique :

- "king" proche de "queen", "prince"
- "war" proche de "civil", "vietnam"
- "love" proche de "romance", "passion"

Les visualisations t-SNE montrent des clusters relativement structurés.
En revanche, les résultats obtenus avec GloVe apparaissent moins cohérents.
Les mots similaires retournés sont parfois dispersés et moins thématiquement liés.





Analyse des différences

Plusieurs facteurs peuvent expliquer ces écarts :

- Word2Vec est un modèle prédictif local, qui converge rapidement.
- GloVe repose sur des statistiques globales nécessitant une couverture importante.
- L'implémentation utilisée ici est simplifiée et entraîne les paires individuellement.
- Le nombre de paires utilisées (max_pairs) reste une fraction du total.

Ainsi, le modèle GloVe n'a probablement pas atteint une convergence suffisante pour produire une structure sémantique aussi claire que Word2Vec.

Cela illustre une différence fondamentale :

- Word2Vec apprend efficacement même avec un échantillonnage partiel
- GloVe nécessite une estimation plus complète de la structure globale

4. Conclusion et limites

Cette deuxième partie du projet a permis d'explorer et de comparer deux approches majeures de représentation vectorielle des mots : Word2Vec et GloVe. Alors que Word2Vec repose sur un apprentissage prédictif local basé sur le contexte immédiat des mots, GloVe adopte une approche statistique globale fondée sur l'analyse des cooccurrences dans l'ensemble du corpus. Cette différence conceptuelle se reflète dans les résultats obtenus.

Dans notre expérimentation sur le dataset Movies, Word2Vec a produit des représentations sémantiques relativement cohérentes, avec des regroupements thématiques clairs et des mots similaires appartenant au même champ lexical. En revanche, les embeddings générés par GloVe se sont révélés moins structurés. Bien que la fonction de coût converge et que l'entraînement soit stable, la qualité sémantique observée reste inférieure à celle de Word2Vec dans notre implémentation.

Cette différence ne remet pas en cause la validité théorique de GloVe, mais souligne l'importance des conditions d'entraînement et des ressources computationnelles. Le modèle

GloVe, reposant sur des statistiques globales, nécessite une couverture suffisante des cooccurrences et un volume d'itérations plus important pour atteindre une convergence optimale. Les contraintes techniques liées à l'exécution sur un ordinateur personnel ont limité le nombre de paires utilisées et le nombre d'itérations, ce qui a probablement impacté la qualité des représentations apprises.

Ainsi, cette comparaison met en évidence que le choix d'un modèle d'embeddings dépend non seulement de ses fondements théoriques, mais également de son implémentation pratique et des ressources disponibles. Word2Vec apparaît plus robuste dans un cadre expérimental à ressources limitées, tandis que GloVe requiert un entraînement plus complet pour exploiter pleinement son potentiel statistique. Ce travail permet donc de mieux comprendre les différences fondamentales entre apprentissage prédictif et modélisation statistique globale dans la construction des représentations vectorielles.

.