Analyse Comparative des Vecteurs Distributionnels Évaluation: qualifier les k plus proches voisins obtenus

Xinhao ZHANG Weixuan CHENG

20 juin 2024

Table des matières

Ressource

Une collection de codes et résultats.
`https ://github.com/XINHAO-ZHANG/Comparaison_Evaluation_Vecteurs`

1 Introduction

Afin d'explorer les différences de représentation lexicale au sein d'un même corpus linguistique à travers différents modèles de langage, la vectorisation de texte est la méthode la plus courante pour fournir une représentation du langage que les ordinateurs peuvent comprendre. Parmi ces méthodes, la représentation distributionnelle permet d'ignorer l'ordre et le contexte des mots dans l'information textuelle, c'est-à-dire de maintenir l'indépendance entre les mots tout en créant un vocabulaire sous la forme d'une vectorisation multidimensionnelle.

Ce projet final vise à réaliser une analyse qualitative des k voisins les plus proches d'un ensemble spécifique de 25 mots, utilisant différents modèles d'embedding pour découvrir comment ces différents modèles capturent et reflètent les nuances sémantiques et contextuelles du langage. Ces méthodes comprennent le comptage direct des fréquences dans le corpus, suivi par une normalisation par l'Information Mutuelle Positive Pointwise (PPMI), une réduction de dimension, et l'utilisation du modèle FastText. Puis la similarité cosinus sera utilisée comme métrique pour identifier les voisins les plus proches. Cette méthode permet de mesurer la proximité sémantique entre les vecteurs de mots dans l'espace vectoriel.

2 Préparation des données

2.1 Choix du corpus

Pour notre projet, nous avons sélectionné le corpus issu du roman "Pride and Prejudice" de Jane Austen, considéré comme un pilier de la littérature anglo-saxon. Dans un premier temps, nous avons décidé de choisir un corpus de taille relativement réduite car il peut simplifier la gestion des données et accélérer le traitement et l'analyse en même temps, tout en évitant les complications liées à une matrice de mots trop épars.

En outre, d'un point de vue linguistique, ce roman utilise des représentations dramaturgiques, en dépeignant les personnages de manière objective à travers des dialogues abondants et variés (discours direct, discours indirect, discours indirect libre, etc.), conférant ainsi au style linguistique un caractère naturel et émotionnellement coloré. L'analyse vectorielle des mots basée sur ce corpus nous permettra également de saisir les caractéristiques stylistiques et expressives propres à l'œuvre de Jane Austen.

2.2 Prétraitement du texte

Dans le cadre du prétraitement du corpus, nous avons adopté une méthode minimaliste pour traiter tous les tokens. Afin de standardiser la casse, le texte a d'abord été converti en minuscules. Ensuite, nous avons employé le module NLTK pour la tokenisation, éliminant les caractères non-alphabétiques tels que les ponctuations. L'étape la plus cruciale a été la suppression des stopwords, ces mots courants qui pourraient compromettre la précision de notre analyse subséquente. Cette approche nous a permis de préparer efficacement le texte

pour les étapes ultérieures de notre projet, en nous concentrant sur les éléments les plus significatifs du corpus.

2.3 Sélection des mots-ciblés

Les mots que nous avons choisi pour étudier sont présentés ci-dessous :

sister	time	much	little	good
nothing	family	man	dear	great
mother	father	day	young	last
letter	room	friend	first	way
house	sure	manner	pleasure	aunt

Tab. 1: liste des 25 mots sélectionnés

En effet, nous avons essayé de pos-tagguer le texte intégral par l'intermédiaire de NTLK et ensuite conservé seuls les noms (NN) et les adjectifs (JJ). Nous pensons que cette sélection cible des mots qui sont plus susceptibles d'être pertinents pour l'expérience et ont plus de sûreté de redonner le résultat. Nous avons ainsi choisi quelques paires de mots synonymiques et antonymiques, comme 'mother' 'father', 'little' 'much' et 'good' 'great', qui pourraient être intéressant à traiter.

3 Vectorisation des mots

3.1 Matrice terme-terme normalisée - méthode A

La méthode de la matrice terme-terme, aussi connue sous le nom de matrice co-occurrence, est une approche de vectorisation qui consiste à créer des vecteurs de mots en mesurant la fréquence de co-occurrence de chaque paire de mots dans un contexte donné.

Concernant l'hyperparamètre du modèle, nous avons défini un window_size de 5. Cela signifie que pour chaque mot, nous prenons en compte les 5 mots précédents et suivants dans le texte. Cette taille de fenêtre ni trop grande ni trop petite aide à couvrir efficacement les mots liés au mot central. Après la construction de cette matrice, nous avons ensuite appliqué le PPMI (Positive Pointwise Mutual Information) pour capturer les relations sémantiques étroites. Il transforme les fréquences brutes en mesures qui soulignent les associations de mots les plus significatives, en éliminant les valeurs négatives qui pourraient fausser l'analyse.

3.2 Réduction de dimension - méthode B

Basée sur la méthode de la matrice terme-terme, nous avons appliqué une réduction de dimensionalité à nos vecteurs de mots pour les ramener à une dimension de 100. Afin d'évaluer l'efficacité de deux types de réduction dimensionnelle - linéaire et non-linéaire - dans le traitement d'un corpus complexe, nous avons choisi d'utiliser deux techniques différentes : PCA (Principal Component Analysis) et t-SNE (t-distributed Stochastic Neighbor Embedding).

Le PCA est une technique linéaire qui, malgré sa rapidité, présente certaines limites, notamment son incapacité à capturer la complexité et les relations non-linéaires entre les caractéristiques. En revanche, le t-SNE, une technique non-linéaire, est plus apte à illustrer ces relations complexes. Cependant, son principal inconvénient réside dans sa lenteur, ce qui peut être un obstacle pour les grands jeux de données.

3.3 Vecteur FastText - méthode C

Word2vec et FastText sont deux méthodes distinctes pour générer des vecteurs denses (complets, sans zéros). Contrairement à la méthode précédente qui comptait les occurrences de termes voisins, Word2vec cherche à prédire ces voisins. Les vecteurs sont obtenus en extrayant la matrice de transformation issue de l'entraînement du réseau neuronal.

FastText opère de façon similaire à Word2vec, mais il se concentre sur les n-grammes de caractères plutôt que sur des tokens entiers. Nous avons utilisé les implémentations de ces deux méthodes disponibles dans la bibliothèque Python gensim pour notre analyse.

4 Analyse et évaluation qualitative des résultats

4.1 Comparaison qualitative globale

Pour réaliser la comparaison qualitative de différents méthodes de vectorisation, nous avons trouvé les k (=10) voisins proches des mots-ciblés.

Dans nos analyses, nous avons classé les mots en différentes catégories : personnages, noms communs, adjectifs, et des termes qui sont à la fois noms et adjectifs. Cette classification a permis d'observer les nuances sémantiques entre les classes de mots. Pour les personnages, les voisins issus de la méthode A révélaient des associations directes dans le contexte des relations familiales et sociales du roman, tandis que la méthode B mettait en lumière des connections plus diversifiées, incluant des éléments narratifs et des traits de caractère. Par exemple, Le mot "sister" dans la méthode A est associé à des personnages et des termes familiaux comme 'mother', reflétant des liens familiaux directs. En méthode B, les voisins comme 'darcy' et 'bennet' montrent une extension aux relations et aux intrigues principales du roman. Ensuite, pour le mot "father", la méthode A révèle 'mother', 'family', indiquant une forte corrélation avec la cellule familiale, tandis que la méthode B ajoute des termes comme 'favourite' et 'godson', élargissant le contexte aux sentiments et aux relations extérieures.

FastText, en se basant sur des n-grams de caractères, tend à produire des résultats qui reflètent davantage une similarité morphologique qu'une concordance sémantique. Par exemple, des mots comme 'daughter', 'together', 'altogether','rather' sont souvent associés en raison de leur suffixe commun 'er'. Il semble que dans le cas de 'personnage' la méthode C se focalise particulièrement sur ce suffixe. Bien que la méthode C ici surpasse les méthodes basées uniquement sur le comptage de co-occurrences et leur réduction dimensionnelle pour ces mots spécifiques, FastText a tout de même tendance à s'égarer et à devenir moins pertinent dans certains contextes.

Mots	Voisins (A)	Voisins (B)	Voisins (C)
sister	'bingley', 'brother', 'elizabeth', 'much', 'must', 'would', 'could', 'still', 'jane', 'mother'	'bingley', 'jane', 'elizabeth', 'much', 'darcy', 'must', 'could', 'bennet', 'miss', 'would'	'sisters', 'daughter', 'altogether', 'master', 'together', 'daughters', 'father', 'larger', 'another', 'matters"
aunt	'uncle', 'phillips', 'letter', 'city', 'forster', 'visits', 'thanked', 'dear', 'go', 'came'	'uncle', 'phillips', 'richard', 'colonel', 'forster','came', 'already', 'alternate', 'city', 'honour'	'language', 'arisen', 'pursuit', 'voluntarily', 'horror', 'became', 'christmas','militia', 'absurd', 'morrow'
mother	'father', 'sisters', 'elizabeth', 'though', 'heard', 'lydia', 'sister', 'affectionate', 'home', 'equivocal'	'elizabeth', 'father', 'jane', 'though', 'heard', 'daughters', 'could', 'attended', 'welcomed', 'lydia'	'brother', 'another', 'weather', 'others', 'rather', 'father', 'together', 'whether', 'altogether', 'thither'
father	'mother', 'lydia', 'good', 'dear', 'one', 'darcy', 'miserly', 'though', 'family', 'brother'	'mother', 'good', 'done', 'lydia', 'favourite', 'godson', 'library', 'daughters', 'one', 'man'	'weather', 'another', 'rather', 'others', 'thither', 'brother', 'mother', 'altogether', 'farther', 'together'

Tab. 2: Mots voisins pour la classe "Personnage"

Pour les noms communs et les adjectifs, les différences entre les listes de voisins obtenues par les deux méthodes soulignaient comment chaque approche capte des aspects variés du texte. Les noms communs trouvés par la méthode A tendaient à être associés à des contextes immédiats et des interactions directes, alors que la méthode B dévoilait des liens qui pourraient refléter des thèmes plus larges et des motifs récurrents. Par exemple, le mot "House" en méthode A est associé à des termes concrets comme 'road' et 'door', représentant l'espace physique. La méthode B révèle des mots comme 'building', indiquant une réflexion sur la structure ou le statut social qu'elle représente. Pour "Letter", il montre des mots liés à l'acte de communication ('wrote', 'read') en méthode A , alors qu'en B, on observe 'impatience', suggérant l'émotion associée à l'attente d'une lettre.

Les résultats générés par la méthode C pour les mots "house" et "letter" tendaient toujours à privilégier la similarité morphologique. Par exemple, des mots tels que "housekeeper" partagent un lien sémantique évident avec "house", tandis que "horror" ne présentent pas de lien sémantique direct malgré une ressemblance morphologique. De même, "letter" et "better" sont similaires en forme, mais sont très éloignés de "letter" en termes de sens.

Mots	Voisins (A)	Voisins (B)	Voisins (C)
house	'oaks', 'front', 'spanish', 'chestnuts', 'intermediate', 'road', 'standing', 'door', 'scattered', 'lawn'	'spanish', 'front', 'chestnuts', 'oaks', 'mile', 'lawn', 'road', 'standing', 'intermediate', 'building'	'housekeeper', 'horror', 'close', 'breakfast', 'nearer', 'amuse', 'curious', 'earliest', 'story', 'purchase'
letter	'wrote', 'contents', 'written', 'received', 'write', 'soon', 'finished', 'jane', 'morning', 'read'	'wrote', 'written', 'impatience', 'write', 'reading', 'jane', 'contents', 'contain', 'read', 'arrived'	'letters', 'better', 'utter', 'written', 'monday', 'forster', 'sunday','quarter', 'chapter', 'answer'

Tab. 3: Mots voisins pour la classe "Noms communs"

Ensuite pour les adjectifs, le mot "Good" en méthode A est relié à des termes comme 'humour' et 'qualities', reflétant des attributs personnels. En B, on trouve 'love', ce qui peut indiquer la présence de thèmes relationnels ou moraux. Puis, pour le mot "young", la méthode A montre des voisins comme 'ladies' et 'woman', probablement dans un contexte descriptif, alors que B présente 'accomplished', 'woman', mettant en lumière les attentes sociales relatives à la jeunesse.

En ce qui concerne l'évaluation des adjectifs voisins, la méthode C se comporte assez bien car le modèle FastText permet de retourner les formes comparatives et superlatives des

adjectifs, par exemple pour "young", "younger" et "youngest" étaient identifiés comme les mots les plus proches.

Mots	Voisins (A)	Voisins (B)	Voisins (C)
good	'humour', 'opinion', 'deal', 'one', 'lydia', 'father', 'much', 'always', 'qualities', 'enough'	'sure', 'father', 'said', 'married', 'dear', 'much', 'well', 'always', 'great', 'love'	'uneasy', 'likelihood', 'goodness', 'easy', 'world', 'clergyman', 'purpose', 'little', 'debts', 'domestic'
young	'ladies', 'man', 'men', 'woman', 'lady', 'accomplished', 'women', 'handsome', 'spleen', 'nothing'	'man', 'men', 'ladies', 'accomplished', 'woman', 'women', 'sex', 'officers', 'assure', 'crowded'	'youngest', 'younger', 'eldest', 'elder','larger', 'neighbours', 'lay','neighbourhood', 'lucases', 'window'

Tab. 4: Mots voisins pour la classe "Adjectifs"

Enfin, pour les mots pouvant être classés à la fois comme noms et adjectifs, les analyses révélaient des tendances intéressantes en matière de polyvalence contextuelle et de variabilité sémantique. Les voisins identifiés par les deux méthodes reflétaient les multiples façons dont ces mots sont utilisés dans l'œuvre, soulignant ainsi l'importance de la méthode de vectorisation choisie pour capturer la complexité du langage littéraire. par exemple, "Dear" dans les deux méthodes est entouré de termes d'affection comme 'lizzy' et 'jane', mais la méthode B inclut aussi des mots qui indiquent des interactions sociales comme 'madam' et 'think', suggérant une utilisation plus formelle ou réfléchie du mot.

Mots	Voisins (A)	Voisins (B)	Voisins (C)
dear	'oh', 'lizzy', 'jane', 'sir', 'replied', 'think', 'suppose', 'well', 'cried', 'go'	'lizzy', 'oh', 'cried', 'sure', 'said', 'know', 'jane', 'well', 'madam', 'think'	'bear', 'clear', 'third', 'lizzy', 'near', 'year', 'thin', 'delightful', 'share', 'afraid'

Tab. 5: Mots voisins pour la classe "Termes à la fois noms et adjectifs"

4.2 Comparaison des méthodes PCA et t-SNE de réduction de dimensionalité

Dans cette section, nous avons exploré les résultats obtenus via deux méthodes de réduction de dimensionalité : PCA et t-SNE, en réduisant à deux dimensions.

En observant la Tableau 6, nous avons constaté que le PCA, qui est une technique linéaire, présente des voisins pour le mot "sister" tels que 'success', 'kind', et 'flattered', qui semblent refléter une perspective positive et émotionnelle. Cependant, lorsque nous examinons les résultats du t-SNE, nous avons trouvé des termes tels que 'notions', 'discrimination', et 'diversified', qui indiquent une variété thématique plus large et des associations plus abstraites. Concernant le mot "young", le PCA associe des mots comme 'tolerably', 'therefore', et 'enough', qui pourraient dénoter une certaine modération ou suffisance. Avec t-SNE, les mots voisins incluent 'liberally', 'least', et 'severe', indiquant un contraste dans les attitudes ou les comportements. Néanmoins, nous avons également observé que les noms directement liés à cet adjectif, tels que "man", "woman", qui étaient présent dans la réduction à 100 dimensions, ne le sont plus.

Mots	PCA	T-SNE
sister	success', 'kind', 'flattered', 'remember', 'comparison', 'much', 'future', 'uncomfortable', 'present', 'settle'	'notions', 'discrimination', 'arch', 'diversified', 'walks', 'frighten', 'moment', 'smile', 'plain', 'alighted'
mother	'many', 'astonishment', 'intelligence', 'fact', 'seldom', 'made', 'observation', 'forced', 'property', 'thanked'	'jealousy', 'interrupting', 'reduced', 'accept', 'thinking', 'absence', 'mistake', 'became', 'secrecy', 'brothers'
young	avoid', 'tolerably', 'therefore', 'change', 'rather', 'enough', 'unexpected', 'something', 'views', 'fear'	'qualities', 'bestowed', 'weeks', 'popular', 'degenerate', 'liberally', 'least', 'lose', 'inclined', 'severe'

TAB. 6: Mots voisins de deux méthodes de réduction de dimensionalité

5 Conclusion

En conclusion, ce projet a démontré les différences de la représentation vectorielle des mots en analysant les voisins identifiés par différentes méthodes d'embedding. L'étude a confirmé que différentes méthodes de vectorisation peuvent révéler divers aspects d'un même corpus, qui pourrait offrir une vue d'ensemble riche et nuancée de sa représentation sémantique.

Par rapport aux vecteurs issus de matrices de cooccurrence plus simples, les résultats obtenus après l'application de la PPMI (Positive Pointwise Mutual Information) et des techniques de réduction dimensionnelle sont relativement meilleurs. Les vecteurs améliorés par PPMI et réduits par PCA ou t-SNE ont révélé des associations de mots plus pertinentes et ont permis une meilleure compréhension des relations sémantiques complexes au sein du texte. Quant à FastText, sa performance dans la tâche de recherche de mots voisins est modérée, principalement parce que l'algorithme se concentre sur la similarité morphologique plutôt que sur la pertinence sémantique.

En ce qui concerne la réduction de dimensionnalité, chaque méthode a révélé des aspects distincts des données sémantiques. Alors que le PCA préserve les relations linéaires et directes entre les mots, le t-SNE a excellé à mettre en évidence des mots moins évidents et des structures locales sémantiquement plus fines. Cela souligne l'importance de choisir la méthode de réduction dimensionnelle en fonction de l'objectif de l'analyse.

De plus, il est également crucial de considérer le nombre de dimensions à maintenir. Bien que la réduction à deux dimensions facilite la visualisation, elle peut entraîner une perte d'informations significatives. Ainsi, conserver un nombre plus élevé de dimensions peut être avantageux pour préserver davantage de caractéristiques vectorielles. La dimension plus haute va mieux représenter les relations sémantiques inhérentes au corpus.

Références

[1] ArcGIS Pro 3.1. *Fonctionnement de la réduction des dimensions* sur le site ersi. Consulté le 11 jan, 2024.
`https ://pro.arcgis.com/fr/pro-app/3.1/tool-reference/spatial-statistics/how-dimension-reduction-works.htm/`

[2] *FasText Word vectors for 157 languages*. Consulté le 12 jan, 2024.
`https ://fasttext.cc/docs/en/crawl-vectors.html`

[3] Meta Research *facebook research/FasTtext*.
`https ://github.com/facebookresearch/fastText`