

Vecteurs, produit scalaire et embeddings

Comment une IA représente le sens des mots

À savoir avant de commencer

Ce document utilise les vecteurs et coordonnées (Seconde/Première), puis le produit scalaire et la similarité cosinus (Première/Terminale). Les vecteurs utilisés sont simplifiés à 2 ou 3 coordonnées pour rester calculables à la main — la réalité utilise des centaines, voire des milliers de coordonnées (voir l'encart en fin de document).

Niveau : Terminale — la version 2D (dans le plan) est accessible dès la Première

Introduction : donner un sens aux mots avec des nombres

Un ordinateur ne comprend que des nombres. Pour qu'une IA comme Claude puisse traiter du texte, chaque mot (ou plutôt chaque fragment de mot, appelé token) est d'abord transformé en une liste de nombres : un vecteur. On appelle cela un embedding (plongement, en français).

L'idée centrale : deux mots de sens proche doivent avoir des vecteurs proches dans cet espace.

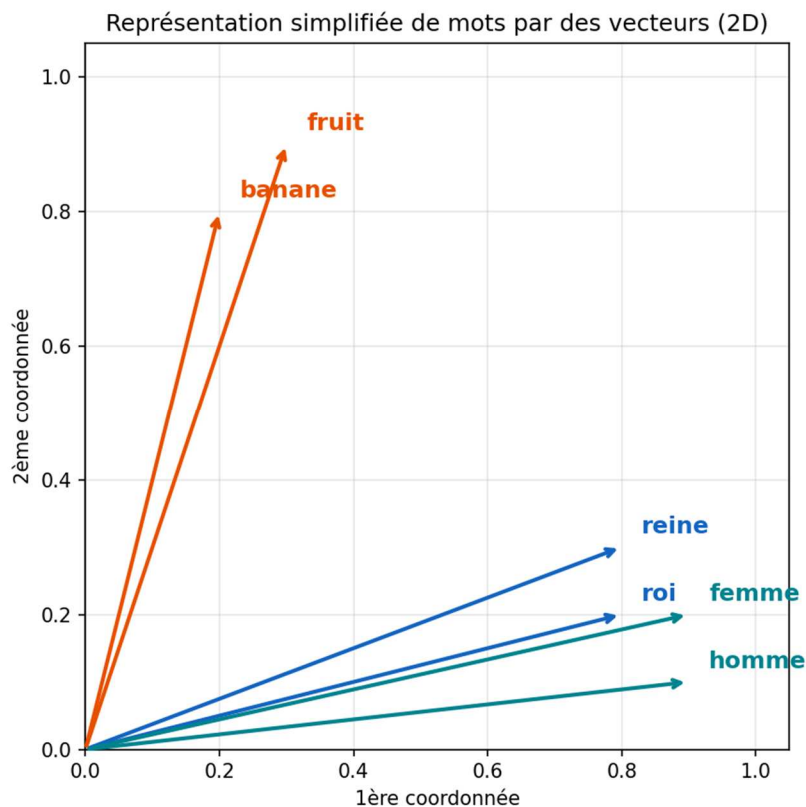
Par exemple, on s'attend à ce que les vecteurs de « roi » et « reine » soient très proches l'un de l'autre, alors que ceux de « roi » et « banane » soient très éloignés. Ce document explore comment on mesure cette proximité, et comment une IA s'en sert pour prédire le mot suivant dans une phrase.

Partie A — Les mots comme des points dans l'espace

Niveau : Terminale (vecteurs et distance dans l'espace)

Dans un modèle de langage réel, chaque mot est représenté par un vecteur à plusieurs centaines de coordonnées. Pour pouvoir calculer à la main, on utilise ici une version très simplifiée, à 3 coordonnées seulement. Tous les vecteurs partent de l'origine $O(0 ; 0 ; 0)$ d'un repère orthonormé de l'espace.

Mot	Coordonnées
roi	(0,8 ; 0,2 ; 0,9)
reine	(0,8 ; 0,3 ; 0,9)
homme	(0,9 ; 0,1 ; 0,7)
femme	(0,9 ; 0,2 ; 0,7)
banane	(0,2 ; 0,8 ; 0,3)
fruit	(0,3 ; 0,9 ; 0,4)



(Le graphique ci-dessus ne représente que les deux premières coordonnées de chaque vecteur, pour pouvoir le tracer sur une feuille.)

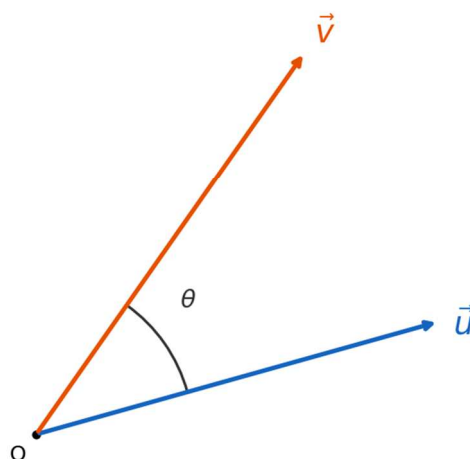
1. Calcule la distance, dans le repère $(O ; i, j, k)$, entre les points associés à roi et reine, à l'aide de la formule :
distance = $\sqrt{[(x_2-x_1)^2 + (y_2-y_1)^2 + (z_2-z_1)^2]}$.
2. Calcule de la même façon la distance entre roi et banane.
3. Compare ces deux distances. Ce résultat est-il cohérent avec le sens des mots en français ?

Partie B — Le produit scalaire, une autre mesure de proximité

Niveau : Terminale (produit scalaire dans l'espace)

La distance n'est pas le seul outil pour comparer deux vecteurs. En intelligence artificielle, on utilise très souvent le produit scalaire, lié à l'angle formé par les deux vecteurs.

$$\cos(\theta) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \times \|\vec{v}\|}$$



Rappel de cours

$$\mathbf{u} \cdot \mathbf{v} = x_u x_v + y_u y_v + z_u z_v \quad \text{et} \quad \cos(\theta) = (\mathbf{u} \cdot \mathbf{v}) / (\|\mathbf{u}\| \times \|\mathbf{v}\|)$$

où $\|\mathbf{u}\| = \sqrt{x_u^2 + y_u^2 + z_u^2}$ est la norme (longueur) du vecteur $\mathbf{u}$.

Plus $\cos(\theta)$ est proche de 1, plus les deux vecteurs pointent dans la même direction : les mots sont sémantiquement proches. Plus $\cos(\theta)$ est proche de 0, plus les vecteurs sont « indépendants » en direction. Un $\cos(\theta)$ proche de -1 indique des vecteurs de sens opposés.

4. Calcule le produit scalaire roi · reine, puis roi · banane.
5. Calcule les normes $\|\text{roi}\|$, $\|\text{reine}\|$ et $\|\text{banane}\|$ (arrondis à 10^{-3} près).
6. En déduire $\cos(\theta)$ pour les couples (roi, reine) et (roi, banane). Que peux-tu en conclure ?
7. Compare cette conclusion avec celle obtenue à l'aide des distances en partie A. Les deux méthodes (distance et produit scalaire) donnent-elles le même classement ?

Partie C — Le vecteur de contexte : deviner le mot suivant

Niveau : Terminale

Pour prédire un mot, une IA ne regarde pas un seul mot isolé, mais tout le contexte qui précède. Une des toutes premières idées historiques (utilisée par exemple dans le modèle Word2Vec, en 2013) consistait à calculer un vecteur de contexte en faisant simplement la moyenne des vecteurs des mots précédents.

Imaginons la phrase inachevée : « Le roi et la ... ». On donne le vecteur du mot « la » : $\text{la} = (0,7 ; 0,3 ; 0,9)$.

Formule du vecteur de contexte (moyenne)

$$\text{contexte} = (1/n) \times (\mathbf{v}_1 + \mathbf{v}_2 + \dots + \mathbf{v}_n)$$

8. Calcule le vecteur de contexte de la phrase « Le roi et la », en faisant la moyenne des vecteurs de « roi » et de « la ».
9. Calcule la distance entre ce vecteur de contexte et chacun des vecteurs suivants : reine (0,8 ; 0,3 ; 0,9), banane (0,2 ; 0,8 ; 0,3), fruit (0,3 ; 0,9 ; 0,4).
10. Quel mot, parmi ces trois candidats, est le plus proche du vecteur de contexte ? Que peut-on en conclure sur la prédiction que ferait cette IA simplifiée pour compléter la phrase ?
11. Cette conclusion te semble-t-elle cohérente avec ce que dirait un locuteur français ?

Partie D — Bonus : l'analogie vectorielle

Niveau : Terminale

Une propriété étonnante de ces représentations vectorielles, observée sur de vrais modèles d'IA entraînés sur des milliards de phrases, est que certaines relations logiques entre mots se traduisent par de simples opérations vectorielles.

12. Calcule le vecteur roi – homme + femme, coordonnée par coordonnée.
13. Compare le résultat avec le vecteur de reine. Que remarques-tu ?
14. En une phrase, interprète ce que représente le calcul « roi – homme + femme » d'un point de vue du sens des mots.

(Cette analogie, roi – homme + femme $\approx$ reine, est un résultat historique célèbre observé pour la première fois sur le modèle Word2Vec en 2013.)

Ce que ce modèle simplifie, et ce que fait réellement une IA comme Claude

Ce que ce modèle simplifie	Ce que fait réellement un modèle comme Claude
Des vecteurs à 2 ou 3 coordonnées seulement, pour pouvoir calculer à la main. Un vecteur fixe par mot : « banane » a toujours exactement le même vecteur, quelle que soit la phrase. Un vecteur de contexte calculé par une simple moyenne des mots précédents, à poids égal. Une comparaison au vecteur de contexte faite « à la fin », sur des mots candidats fixés à l'avance.	Des vecteurs à plusieurs centaines, voire milliers de coordonnées, appris automatiquement pendant l'entraînement (par descente de gradient — voir document 0). Un vecteur contextuel : le mot « avocat » n'a pas le même vecteur dans « l'avocat a plaidé » et « j'ai mangé un avocat » — le sens exact est recalculé à chaque phrase. Un mécanisme appelé attention, qui pondère chaque mot du contexte différemment (et de façon apprise)

selon sa pertinence, plutôt que de tous les moyenner à poids égal.

Une distribution de probabilité calculée sur tout le vocabulaire (des dizaines de milliers de tokens), via un immense réseau de neurones, et non une simple recherche du plus proche voisin.

L'idée centrale — représenter le sens par des vecteurs, et mesurer la proximité sémantique par une distance ou un produit scalaire — reste, elle, parfaitement fidèle à la réalité : c'est exactement le principe des embeddings utilisés dans tous les modèles de langage actuels.

Exercices d'entraînement

On donne les vecteurs suivants, dans un espace simplifié à 3 dimensions :

Mot	Coordonnées
Paris	(0,90 ; 0,10 ; 0,20)
France	(0,85 ; 0,15 ; 0,25)
Rome	(0,90 ; 0,10 ; 0,60)
Italie	(0,85 ; 0,15 ; 0,65)
chaise	(0,10 ; 0,90 ; 0,50)

Exercice 1 — Distances

Niveau : Terminale (vecteurs et distance dans l'espace)

- Calcule la distance entre Paris et Rome, puis entre Paris et chaise.
- Que peux-tu en conclure sur la proximité sémantique entre ces mots ?

Exercice 2 — Produit scalaire et cosinus

Niveau : Terminale (produit scalaire dans l'espace)

- Calcule le produit scalaire Paris · France, puis les normes $\| \text{Paris} \|$ et $\| \text{France} \|$.
- En déduire $\cos(\theta)$ pour le couple (Paris, France). Que peux-tu en conclure ?
- Calcule $\cos(\theta)$ pour le couple (Paris, chaise) et compare avec la question précédente.

Exercice 3 — Analogie vectorielle (bonus)

Niveau : Terminale

- Calcule le vecteur Paris – France + Italie.
- Compare ce résultat avec le vecteur de Rome. Que remarques-tu ?
- Propose, avec tes mots, une explication de cette analogie « capitale – pays + autre pays $\approx$ autre capitale ».

Corrigé

Partie A

A.1. $\text{distance}(\text{roi}, \text{reine}) = \sqrt{(0,8-0,8)^2 + (0,3-0,2)^2 + (0,9-0,9)^2} = \sqrt{0,01} = 0,1.$

A.2. $\text{distance}(\text{roi}, \text{banane}) = \sqrt{(0,2-0,8)^2 + (0,8-0,2)^2 + (0,3-0,9)^2} = \sqrt{0,36+0,36+0,36} = \sqrt{1,08} \approx 1,039.$

A.3. La distance entre roi et reine (0,1) est bien plus petite que celle entre roi et banane ($\approx 1,039$) : les vecteurs traduisent correctement le fait que « roi » et « reine » sont sémantiquement très proches, contrairement à « roi » et « banane ».

Partie B

B.1. $\text{roi} \cdot \text{reine} = 0,8 \times 0,8 + 0,2 \times 0,3 + 0,9 \times 0,9 = 0,64 + 0,06 + 0,81 = 1,51.$ $\text{roi} \cdot \text{banane} = 0,8 \times 0,2 + 0,2 \times 0,8 + 0,9 \times 0,3 = 0,16 + 0,16 + 0,27 = 0,59.$

B.2. $\| \text{roi} \| = \sqrt{0,64+0,04+0,81} = \sqrt{1,49} \approx 1,221.$ $\| \text{reine} \| = \sqrt{0,64+0,09+0,81} = \sqrt{1,54} \approx 1,241.$ $\| \text{banane} \| = \sqrt{0,04+0,64+0,09} = \sqrt{0,77} \approx 0,877.$

B.3. $\cos(\text{roi}, \text{reine}) = 1,51 / (1,221 \times 1,241) \approx 1,51 / 1,515 \approx 0,996$. $\cos(\text{roi}, \text{banane}) = 0,59 / (1,221 \times 0,877) \approx 0,59 / 1,071 \approx 0,551$. Le cosinus proche de 1 pour (roi, reine) confirme une très forte proximité de direction ; celui de (roi, banane), plus modéré ($\approx 0,55$), indique une proximité bien plus faible.

B.4. Les deux méthodes s'accordent : reine est bien plus proche de roi que banane, que ce soit en distance ou en produit scalaire (cosinus). C'est rassurant, mais il faut savoir que dans des espaces à de nombreuses dimensions, distance et cosinus ne donnent pas toujours exactement le même classement — c'est pourquoi les IA choisissent en général l'une ou l'autre selon le contexte (le cosinus est souvent préféré car il ignore la « longueur » du vecteur, qui peut varier pour des raisons sans lien avec le sens).

Partie C

C.1. vecteur de contexte = $((0,8+0,7)/2 ; (0,2+0,3)/2 ; (0,9+0,9)/2) = (0,75 ; 0,25 ; 0,9)$.

C.2. $\text{distance}(\text{contexte}, \text{reine}) = \sqrt{[(0,8-0,75)^2 + (0,3-0,25)^2 + (0,9-0,9)^2]} = \sqrt{0,0025+0,0025} = \sqrt{0,005} \approx 0,071$.

$\text{distance}(\text{contexte}, \text{banane}) = \sqrt{[(0,2-0,75)^2 + (0,8-0,25)^2 + (0,3-0,9)^2]} = \sqrt{0,3025+0,3025+0,36} = \sqrt{0,965} \approx 0,982$.

$\text{distance}(\text{contexte}, \text{fruit}) = \sqrt{[(0,3-0,75)^2 + (0,9-0,25)^2 + (0,4-0,9)^2]} = \sqrt{0,2025+0,4225+0,25} = \sqrt{0,875} \approx 0,935$.

C.3. Le mot le plus proche du vecteur de contexte est reine (distance $\approx 0,071$, bien plus petite que les deux autres). L'IA simplifiée prédirait donc « reine » pour compléter « Le roi et la ... ».

C.4. Oui : « Le roi et la reine » est effectivement une suite naturelle et très plausible en français, contrairement à « Le roi et la banane » ou « Le roi et la fruit » (qui n'est même pas grammaticalement correct).

Partie D

D.1. $\text{roi} - \text{homme} + \text{femme} = (0,8-0,9+0,9 ; 0,2-0,1+0,2 ; 0,9-0,7+0,7) = (0,8 ; 0,3 ; 0,9)$.

D.2. Ce vecteur est exactement égal au vecteur de reine (0,8 ; 0,3 ; 0,9) !

D.3. Ce calcul montre que la direction « masculin $\rightarrow$ féminin » (obtenue en soustrayant homme et en ajoutant femme) semble être apprise de façon cohérente dans l'espace des vecteurs : partir de « roi », retirer ce qui est masculin et ajouter ce qui est féminin, ramène très précisément à « reine ». Cela illustre que ces espaces vectoriels capturent, dans une certaine mesure, des relations logiques entre les mots, et pas seulement leur proximité de sens brute.

Exercice 1

1. $\text{distance}(\text{Paris}, \text{Rome}) = \sqrt{[(0,9-0,9)^2 + (0,1-0,1)^2 + (0,2-0,6)^2]} = \sqrt{0,16} = 0,4$. $\text{distance}(\text{Paris}, \text{chaise}) = \sqrt{[(0,9-0,1)^2 + (0,1-0,9)^2 + (0,2-0,5)^2]} = \sqrt{0,64+0,64+0,09} = \sqrt{1,37} \approx 1,170$.

2. Paris est bien plus proche de Rome (deux capitales) que de chaise (un objet du quotidien, sans lien sémantique) : la distance confirme la proximité sémantique attendue.

Exercice 2

1. $\text{Paris} \cdot \text{France} = 0,9 \times 0,85 + 0,1 \times 0,15 + 0,2 \times 0,25 = 0,765 + 0,015 + 0,05 = 0,83$. $\|\text{Paris}\| = \sqrt{(0,81+0,01+0,04)} = \sqrt{0,86} \approx 0,927$. $\|\text{France}\| = \sqrt{(0,7225+0,0225+0,0625)} = \sqrt{0,8075} \approx 0,899$.

2. $\cos(\text{Paris}, \text{France}) = 0,83 / (0,927 \times 0,899) \approx 0,83 / 0,833 \approx 0,996$. Un cosinus très proche de 1 confirme une direction quasiment identique : « Paris » et « France » sont fortement associés (une capitale et son pays).

3. $\text{Paris} \cdot \text{chaise} = 0,9 \times 0,1 + 0,1 \times 0,9 + 0,2 \times 0,5 = 0,09 + 0,09 + 0,1 = 0,28$. $\|\text{chaise}\| = \sqrt{(0,01+0,81+0,25)} = \sqrt{1,07} \approx 1,034$. $\cos(\text{Paris}, \text{chaise}) = 0,28 / (0,927 \times 1,034) \approx 0,28 / 0,959 \approx 0,292$. Ce cosinus, bien plus faible, confirme que « Paris » et « chaise » sont sémantiquement bien plus éloignés que « Paris » et « France ».

Exercice 3

1. $\text{Paris} - \text{France} + \text{Italie} = (0,90-0,85+0,85 ; 0,10-0,15+0,15 ; 0,20-0,25+0,65) = (0,90 ; 0,10 ; 0,60)$.

2. Ce vecteur est exactement égal à celui de Rome (0,90 ; 0,10 ; 0,60) !

3. En partant de la capitale Paris, en retirant ce qui caractérise son pays (France) et en ajoutant ce qui caractérise un autre pays (Italie), on obtient précisément la capitale correspondante (Rome). Cela montre que la relation « capitale de » est, elle aussi, encodée de façon cohérente comme une direction quasiment constante dans cet espace vectoriel, qu'on peut « appliquer » d'un pays à un autre par une simple addition/soustraction de vecteurs.

Mathématiques du programme mobilisées dans ce document

Seconde / Première — le principe, en 2D

Repérage dans le plan, coordonnées de points, distance entre deux points (Seconde).

Vecteurs et produit scalaire dans le plan (Première) : tout ce document fonctionnerait à l'identique avec des vecteurs à 2 coordonnées seulement.

Terminale (spécialité mathématiques) — la version utilisée ici, en 3D

Vecteurs de l'espace, coordonnées d'un vecteur (extension à 3 coordonnées du programme de Première).

Produit scalaire dans l'espace, norme d'un vecteur, cosinus de l'angle entre deux vecteurs.

Moyenne de plusieurs vecteurs, interprétation géométrique d'opérations vectorielles (addition, soustraction).

Point de vigilance

Ce document utilise des vecteurs à 3 coordonnées pour rester proche de la réalité (un vrai embedding a des centaines de coordonnées). Le produit scalaire et la distance dans l'espace (3 coordonnées) relèvent du programme de Terminale ; le même travail avec 2 coordonnées seulement (dans le plan) est accessible dès la Première.