1 Définition

La similitude par cosinus permet de déterminer le degré de similitude entre deux phrases par calcul du cosinus de l'angle formé par les deux vecteurs de chacune de ces phrases.
Soit les phrases suivantes :

1. Bonjour John.
2. Bonjour Doé.

On remarque que les deux phrases sont similaires mais comment le savoir avec un ordinateur.

2 Exemple Simple espace à 2 dimensions

Soit les phrases :

1. Bonjour John.
2. Bonjour.

On remarque qu'une des deux phrases est constituée de deux mots donc on peut les représenter dans un espace à deux dimensions.
On crée le tableau de similitude suivant :

Phrases	**Bonjour**	**John**
Bonjour John	1	1
Bonjour	1	0

TABLE 1 – Tableau de similitude simple exemple 1

La représentation dans un repère donne :

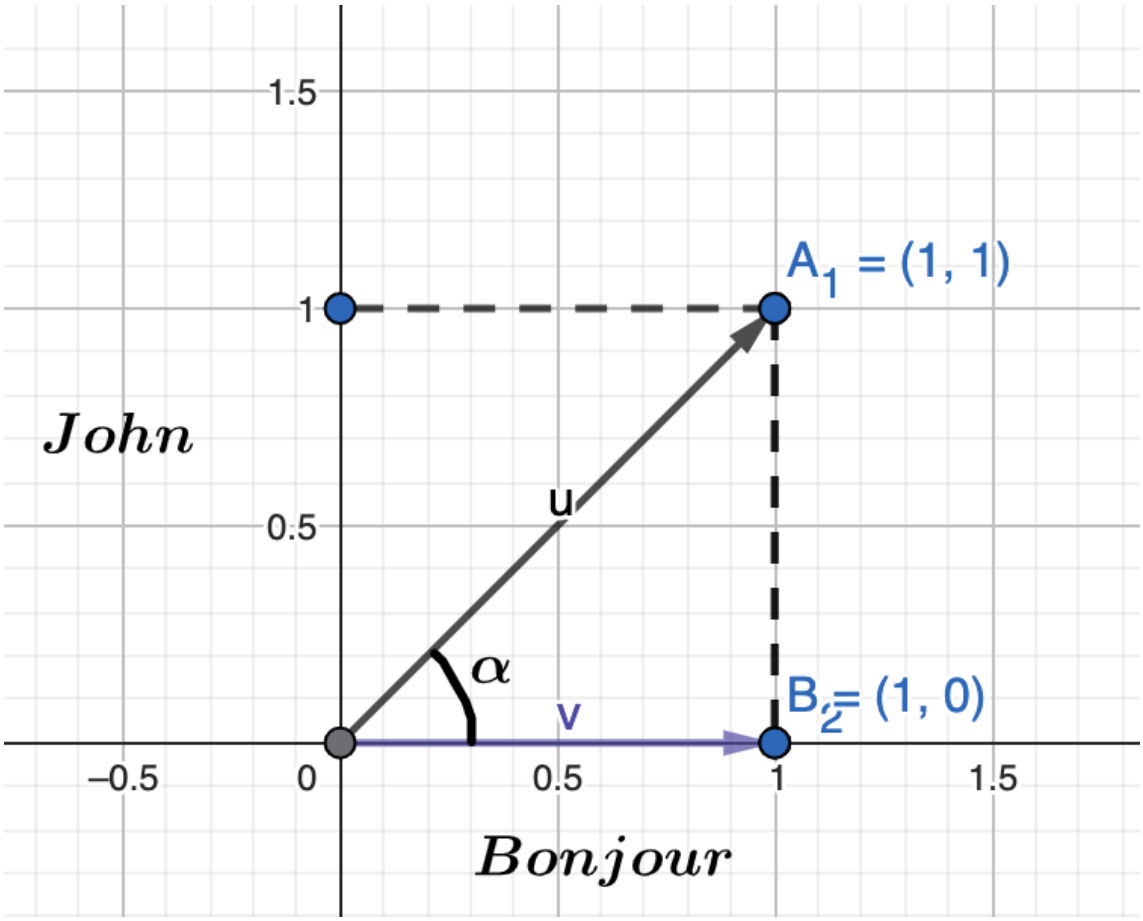

FIGURE 1 – Représentation simple similitude exemple 1

La phrase 1 est représentée par le vecteur $A_1(1,1)$ et la phrase 2 par le vecteur $B_2(1,0)$.
L'angle formé par ces deux vecteurs est l'angle α dont le cosinus permet de calculer la similarité entre les deux phrases.
Dans l'exemple précédent, $\alpha = 45^o$ dont $cos(\alpha) = 0.71$. Soit 71% de chance que les deux phrases soient similaires.
Supposons que la phrase 2 soient **Bonjour, Bonjour, Bonjour.** alors on aura :

Phrases	**Bonjour**	**John**
Bonjour John	1	1
Bonjour, Bonjour, Bonjour	3	0

TABLE 2 – Tableau de similitude simple exemple 2

La représentation dans un repère donne :

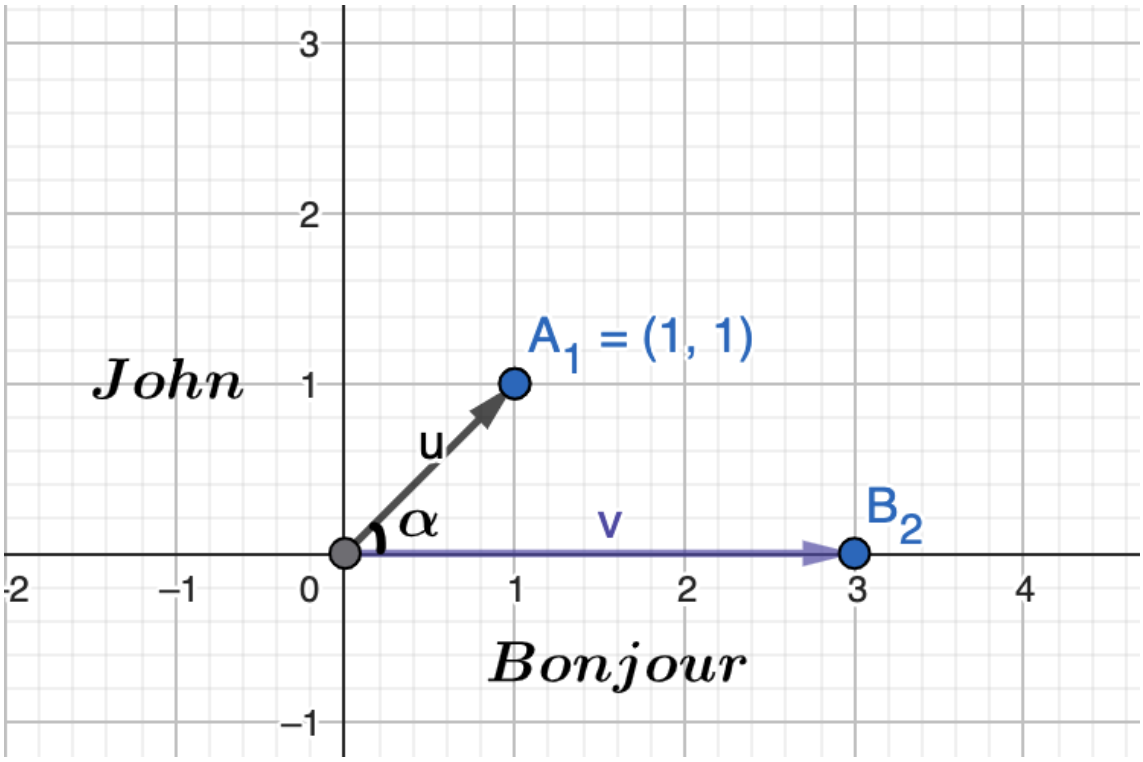

FIGURE 2 – Représentation simple similitude exemple 2

Peut importe la longueur de la phrase 2 le nombre d'ajout de *Bonjour* ne change pas le *cosinus* de α.

Résumé : La similitude par cosinus (*cosine similarity en anglais*) est le cosinus de l'angle formé entre deux vecteurs et permet de révéler la similitude entre ceux-ci. Sa valeur est toujours entre 0 et 1.

3 Cas de similitude exacte

Soient les deux phrases suivantes :

1. Bonjour John
2. Bonjour John

On a comme comme tableau de similitude :

Phrases	Bonjour	John
Bonjour John	1	1
Bonjour John	1	1

TABLE 3 – Tableau de similitude exacte

La représentation dans un repère donne :

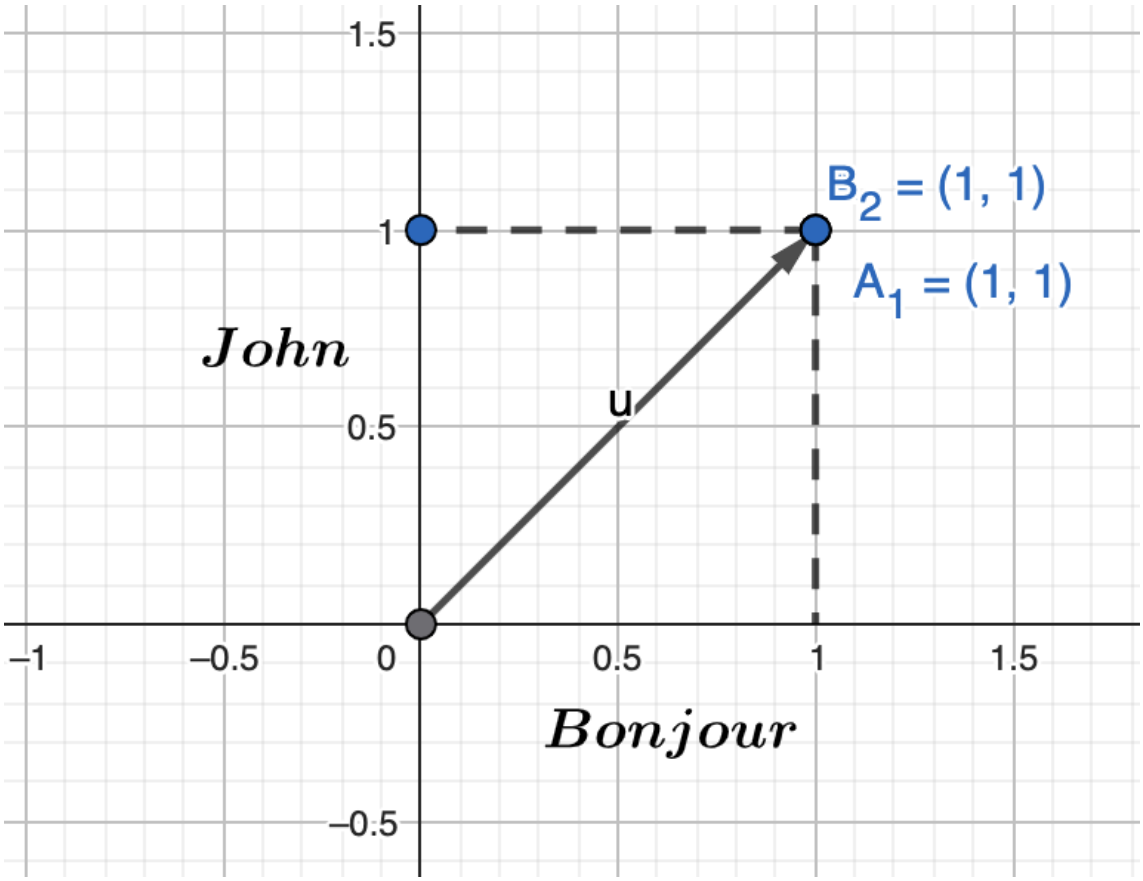

FIGURE 3 – Représentation similitude exacte

Les deux vecteurs sont confondus donc $\alpha = 0^o$ soit $cos(\alpha) = 1$ soit 100% de chance que les deux vecteurs soient similaires. On peut conclure que les deux phrases sont totalement similaires.

4 Cas d'aucune similitude

Soient les deux phrases suivantes :

1. Bonjour.
2. John.

Le tableau de similitude donne :

Phrases	**Bonjour**	**John**
Bonjour	1	0
John	0	1

TABLE 4 – Tableau de non similitude

La représentation dans un repère donne :

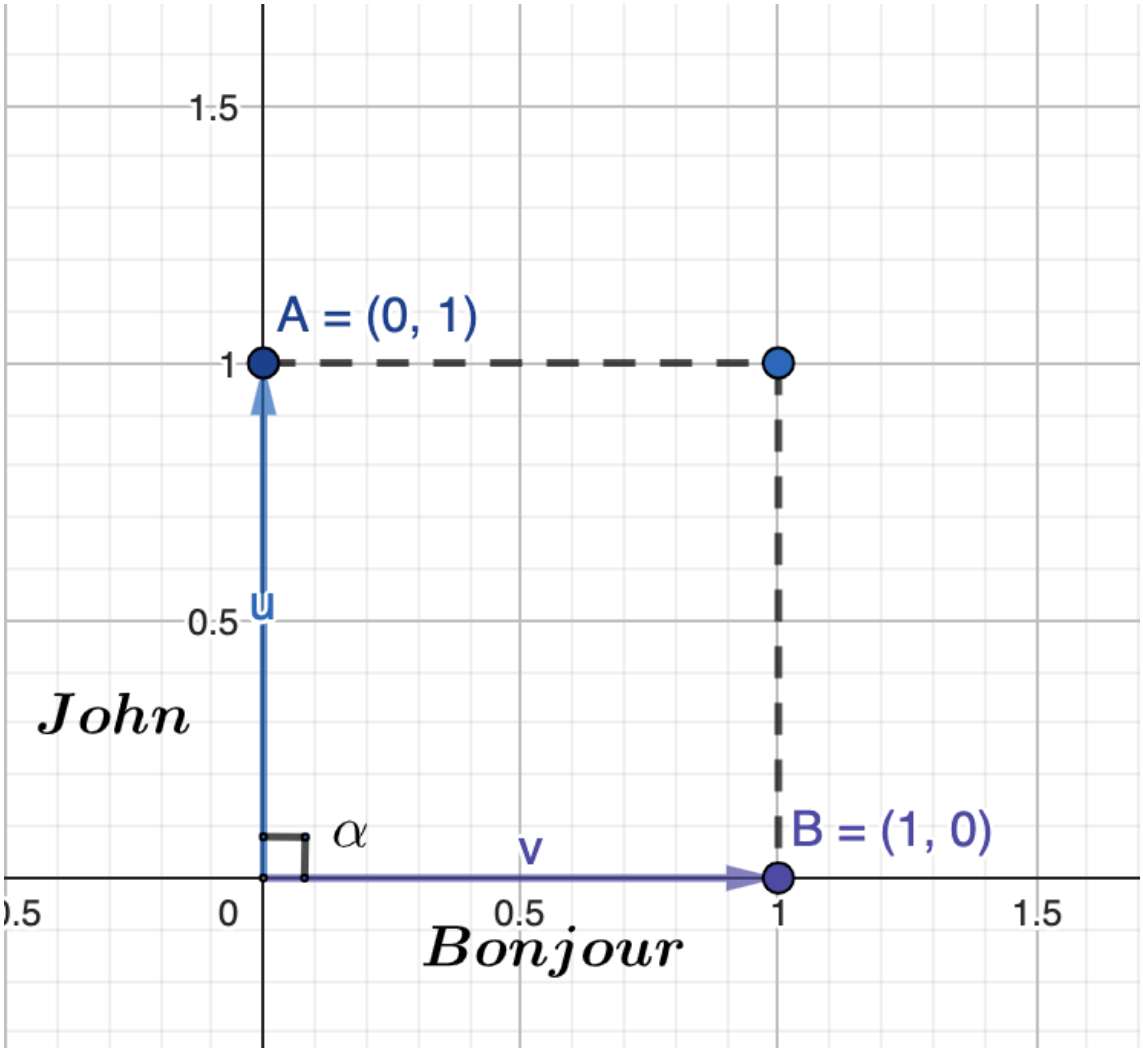

FIGURE 4 – Représentation non similitude

L'angle formé par les deux vecteurs est $\alpha = 90^o$ donc $cos(\alpha) = 0$ soit 0% de chance que les deux vecteurs soient similaires. On peut conclure que les deux phrases sont totalement différentes.

5 Résumé

$$cos(\alpha) = \begin{cases} 0 & \text{Si aucune similitude entre les deux phrases.} \\ 1 & \text{Similitude totalement exacte.} \\ \in]0,1[& \text{Sensiblement similaires quand les deux phrases ont en commun des mots.} \end{cases}$$

Pour avoir le **cosinus** de la similitude entre deux phrases suivre les étapes suivantes :

1. Faire un tableau de fréquence des mots (compter le nombre d'occurrences de chaque mot dans chacune des phrases).
2. Afficher les points.
3. Trouver l'angle entre Vecteurs.
4. Calculer le cosinus de l'angle formé.

6 Cas complexes

Les cas précédents marchent pour des phrases avec deux mots c'est à dire représentables dans un espace à 2 dimensions ℝxℝ.
Les phrases de la vie réelle sont de plusieurs mots soit représentables dans un espace $\mathbb{R}^n$, dans ce cas on utilise la formule suivante :

$$cos(\alpha) = \frac{\sum_{i=1}^{n} A_i Bi}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}, avec\ i = \text{ indice des mots}$$

Soient les phrases suivantes :

1. Bonjour John
2. Bonjour

Phrases	**Bonjour**	**John**
Bonjour John	1	1
Bonjour	1	0

TABLE 5 – Tableau de similitude avec formule

Le calcul de la similitude donne :

$$cos(\alpha) = \frac{(1*1)+(1*0)}{\sqrt{1^2+1^2} * \sqrt{1^2+0^2}} = \frac{1}{\sqrt{2}*1} = 0.7071$$

Soit les $cos(45)$ trouvé précédemment.
Soient deux phrases :

1. Bonjour tout le monde.
2. Bonjour John

Le tableau de similitude donne :

Phrases	**Bonjour**	**tout**	**le**	**monde**	**John**
Bonjour tout le monde.	1	1	1	1	0
Bonjour John	1	0	0	0	1

TABLE 6 – Tableau de similitude plusieurs dimensions

Le calcul de la similitude donne :

$$cos(\alpha) = \frac{(1*1)+(1*0)+(1*0)+(1*0)+(0*1)}{\sqrt{1^2+1^2+1^2+1^2+0^2} * \sqrt{1^2+0^2+0^2+0^2+1^2}} = \frac{1}{2*\sqrt{2}} = 0.35$$

Soit 35% de chance d'être similaires.

7 Aller plus loin

Le soucis avec cette méthode est que la similitude se base sur la construction des mots (plan syntaxique/grammatical) donc deux mots **Bonjour** et **Hello** seront considérés comme non similaires car le sens implicite n'est pas pris en compte.
De même **Beau** et **Belle** seront considérés comme non similaires.
Dans les projets, on utilisera le lemme des mots c'est à dire la forme de base du mot. Avec l'utilisation du lemme on aura :

- beau => lemme beau
- belle => lemme beau

En utilisant les lemmes **Beau** et **Belle** seront représentés par leur forme lemmatisée et donc similaire.
De plus, cette méthode ne prend pas en compte le sens sémantique entre deux phrases c'est à dire deux phrases **Le chat mange la souris** et **La souris mange le chat** seront considérées comme similaires au plan syntaxique et grammatical mais sur le plan sémantique (du sens) elles sont totalement différentes.
Pour aller plus loin sur le sujet, on peut utiliser les librairies de NLP (Natural Language Processing) comme SpaCy et NLTK qui contiennent des outils de traitement automatique des langues.