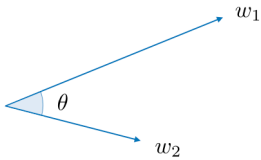
Remarque : les composantes individuelles de la représentation d'un mot n'est pas nécessairement facilement interprétable.

2.4 Comparaison de mots

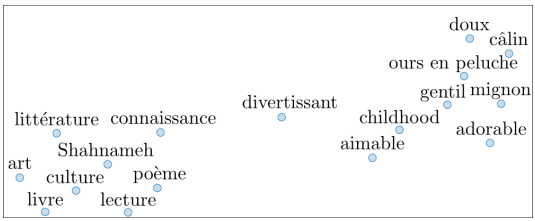
❒ **Similarité cosinus** – La similarité cosinus (en anglais *cosine similarity*) entre les mots w_1 et w_2 est donnée par :

$$\text{similarity} = \frac{w_1 \cdot w_2}{||w_1||\,||w_2||} = \cos(\theta)$$

Remarque : θ est l'angle entre les mots w_1 et w_2.



❒ **t-SNE** – La méthode t-SNE (en anglais *t-distributed Stochastic Neighbor Embedding*) est une technique visant à réduire une représentation dans un espace de haute dimension en un espace de plus faible dimension. En pratique, on visualise les vecteur-mots dans un espace 2D.



2.5 Modèle de langage

❒ **Vue d'ensemble** – Un modèle de langage vise à estimer la probabilité d'une phrase $P(y)$.

❒ **Modèle n-gram** – Ce modèle consiste en une approche naïve qui vise à quantifier la probabilité qu'une expression apparaisse dans un corpus en comptabilisant le nombre de son apparition dans le training data.

❒ **Perplexité** – Les modèles de langage sont communément évalués en utilisant la perplexité, aussi noté PP, qui peut être interprété comme étant la probabilité inverse des données normalisée par le nombre de mots T. La perplexité est telle que plus elle est faible, mieux c'est. Elle est définie de la manière suivante :

$$\text{PP} = \prod_{t=1}^{T} \left(\frac{1}{\sum_{j=1}^{|V|} y_j^{(t)} \cdot \widehat{y}_j^{(t)}} \right)^{\frac{1}{T}}$$

Remarque : PP est souvent utilisée dans le cadre du t-SNE.