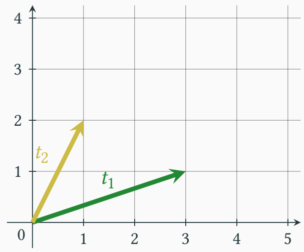
Similarités : exemple de travail

On considère deux points $t_1 = (v_{1,1}, \ldots v_{1,n})$ et $t_2 = (v_{2,1}, \ldots, v_{2,n})$ représentant des documents

Par exemple, si on utilise comme traits les fréquences de *orange* et *kiwi*, étant donné le fichier tabulaire suivant

Texte	orange	kiwi
t_1	3	1
t_2	1	2

On aura $t_1 = \binom{3}{1}$ et $t_2 = \binom{1}{2}$



Exemple de travail

Exemple

1. Représenter le corpus d'exemple sous forme tabulaire en utilisant comme attributs les fréquences des mots ayant trait au *cinéma* et à l'*économie*
2. En déduire une représentation du corpus comme un ensemble de points du plan

Pour la suite, choisir deux de ces points et calculer leur distance pour chacune des distances proposées.

Distance de Manhattan

Définition (distance de Manhattan)
On appelle *distance de Manhattan* entre t_1 et t_2 le nombre

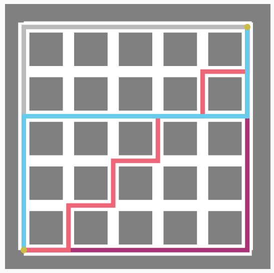
$$d_1(t_1, t_2) = \sum_{k=1}^{n} |v_{2,k} - v_{1,k}| \left(= |v_{2,1} - v_{1,1}| + \ldots + |v_{2,n} - v_{1,n}|\right)$$

Distance de Manhattan

Définition (distance de Manhattan)
On appelle *distance de Manhattan* entre t_1 et t_2 le nombre

$$d_1(t_1, t_2) = \sum_{k=1}^{n} |v_{2,k} - v_{1,k}| \left(= |v_{2,1} - v_{1,1}| + \ldots + |v_{2,n} - v_{1,n}|\right)$$

On l'appelle aussi *taxicab distance* : c'est la distance du chemin le plus court pour aller d'un point à un autre en se déplaçant sur une grille



Taxicab distance

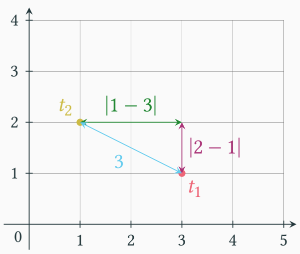
Distance de Manhattan

Définition (distance de Manhattan)
On appelle *distance de Manhattan* entre t_1 et t_2 le nombre

$$d_1(t_1, t_2) = \sum_{k=1}^{n} |v_{2,k} - v_{1,k}| \left(= |v_{2,1} - v_{1,1}| + \ldots + |v_{2,n} - v_{1,n}|\right)$$

Pour notre exemple

$$\begin{aligned} d_1(t_1, t_2) &= \sum_{k=1}^{n} |v_{2,k} - v_{1,k}| \\ &= |v_{2,1} - v_{1,1}| + |v_{2,2} - v_{1,2}| \\ &= |1 - 3| + |2 - 1| \\ &= 2 + 1 \\ &= 3 \end{aligned}$$



Distance de Manhattan

Distance euclidienne

Définition (distance euclidienne)
On appelle *distance euclidienne*[1] entre t_1 et t_2 le nombre

$$d_2(t_1, t_2) = \sqrt{\sum_{k=1}^{n} (v_{2,k} - v_{1,k})^2}$$

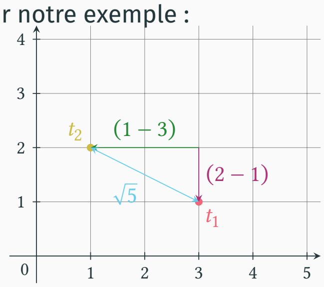
Distance euclidienne

Définition (distance euclidienne)
On appelle *distance euclidienne*[1] entre t_1 et t_2 le nombre

$$d_2(t_1, t_2) = \sqrt{\sum_{k=1}^{n} (v_{2,k} - v_{1,k})^2}$$

C'est la distance usuelle dans le plan. Pour notre exemple :

$$\begin{aligned} d_2(t_1, t_2) &= \sqrt{\sum_{k=1}^{n} (v_{2,k} - v_{1,k})^2} \\ &= \sqrt{(v_{2,1} - v_{1,1})^2 + (v_{2,2} - v_{1,2})^2} \\ &= \sqrt{(1-3)^2 + (2-1)^2} \\ &= \sqrt{2^2 + 1^2} \\ &= \sqrt{5} \end{aligned}$$



Distance euclidienne

1. Euclide d'Alexandrie, IIIè s AEC

Distance de Tchebychev

Définition (distance de Tchebychev)
On appelle *distance de Tchebychev*[1] entre t_1 et t_2 le nombre

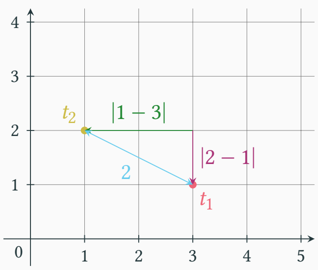
$$d_\infty(t_1, t_2) = \max_{1 \leqslant k \leqslant n} |v_{2,k} - v_{1,k}|$$

Distance de Tchebychev

Définition (distance de Tchebychev)
On appelle *distance de Tchebychev*[1] entre t_1 et t_2 le nombre

$$d_\infty(t_1, t_2) = \max_{1 \leqslant k \leqslant n} |v_{2,k} - v_{1,k}|$$

$$\begin{aligned} d_\infty(t_1, t_2) &= \max_{1 \leqslant k \leqslant n} |v_{2,k} - v_{1,k}| \\ &= \max_{1 \leqslant k \leqslant n} \left(|v_{2,1} - v_{1,1}|, |v_{2,2} - v_{1,2}|\right) \\ &= \max_{1 \leqslant k \leqslant n} \left(|1-3|, |2-1|\right) \\ &= \max_{1 \leqslant k \leqslant n} (2, 1) \\ &= 2 \end{aligned}$$



Distance de Tchebychev

1. Пафну́тий Льво́вич Чебышёв, 1821–1894

Distances de Minkowski

Définition (distance de Minkowski)
Pour tout $p \geqslant 1$, on appelle *distance de Minkowski*[1] *de paramètre* p entre t_1 et t_2 le nombre

$$d_p(t_1 - t_2) = \left(\sum_{k=1}^{n} |v_{2,k} - v_{1,k}|^p\right)^{\frac{1}{p}}$$

Distances de Minkowski

Définition (distance de Minkowski)
Pour tout $p \geqslant 1$, on appelle *distance de Minkowski*[1] *de paramètre* p entre t_1 et t_2 le nombre

$$d_p(t_1 - t_2) = \left(\sum_{k=1}^{n} |v_{2,k} - v_{1,k}|^p\right)^{\frac{1}{p}}$$

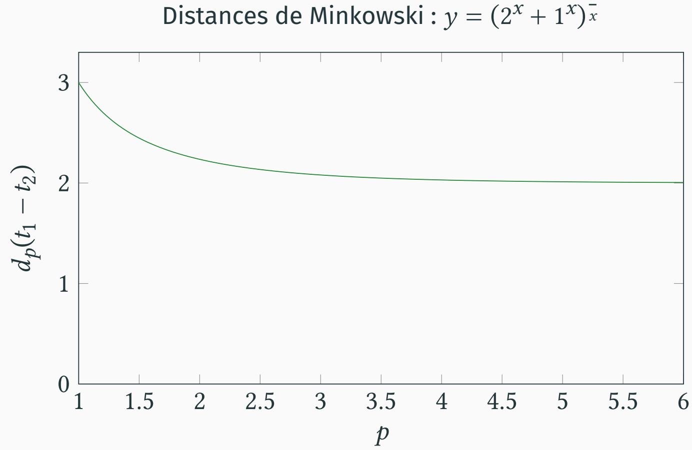
Pour $p = 1$ on retrouve la distance de Manhattan, pour $p = 2$ la distance euclidienne, et pour $p \to +\infty$, la distance de Minkowski tend vers la distance de Tchebychev.

Intuitivement, pour $p = 1$, elle traite également tous les écarts entre coordonnées, pour $p = \infty$, elle ne conserve que le plus grand écart, et les autres p donnent une interpolation entre ces deux extrêmes.

1. Hermann Minkowski, 1864–1909

Distances de Minkowski

Distances de Minkowski : $y = (2^x + 1^x)^{\frac{1}{x}}$



Distances de Minkowski pour notre exemple