

02-02

Représentation des documents et métriques de distance

**NOUS ÉCLAIRONS.
VOUS BRILLEZ.**

FORMATION CONTINUE
ET SERVICES AUX ENTREPRISES



Sommaire

1. Représentation des documents
2. Métriques de distance
3. Lectures et références

Sommaire

1. Représentation des documents
2. Métriques de distance
3. Lectures et références

Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

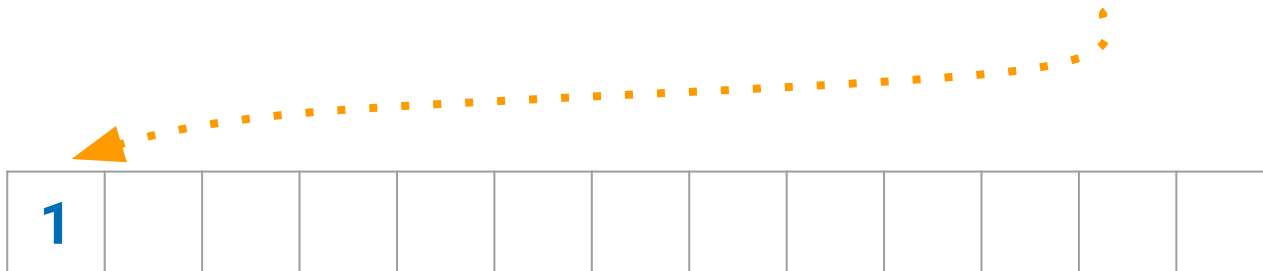
Messi est un joueur de foot de l'équipe
d'Argentine. Au Québec, le foot est
appelé "soccer"



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

Messi est un joueur de foot de l'équipe d'Argentine. Au Québec, le foot est appelé "soccer"



Messi

Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

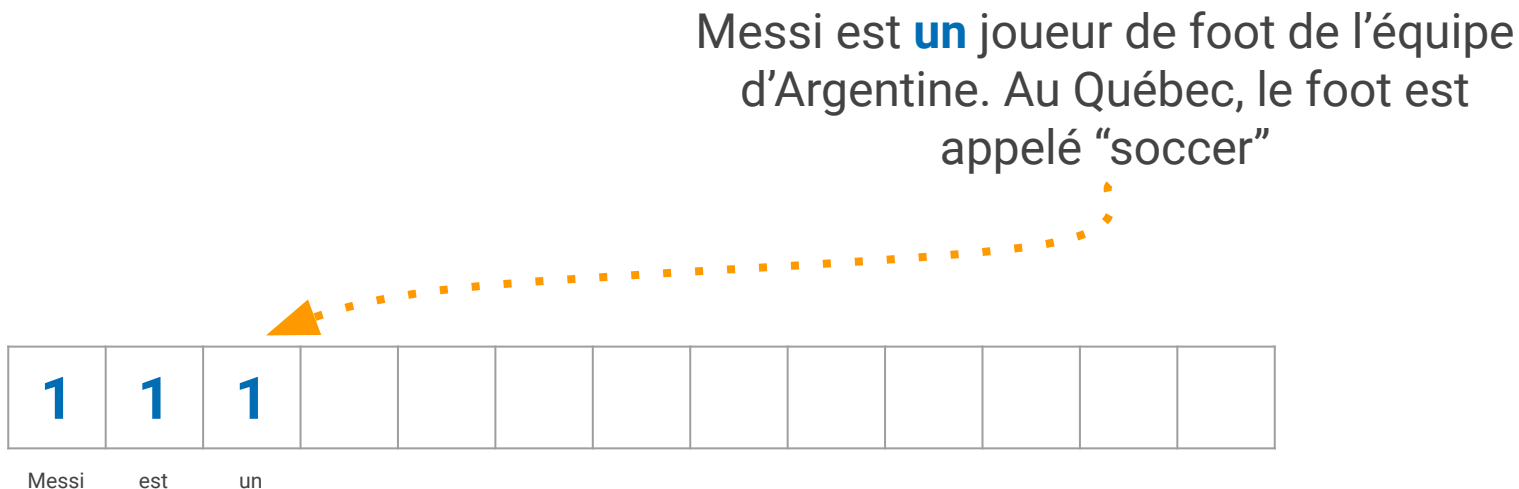
Messi **est** un joueur de foot de l'équipe d'Argentine. Au Québec, le foot est appelé "soccer"



Messi est

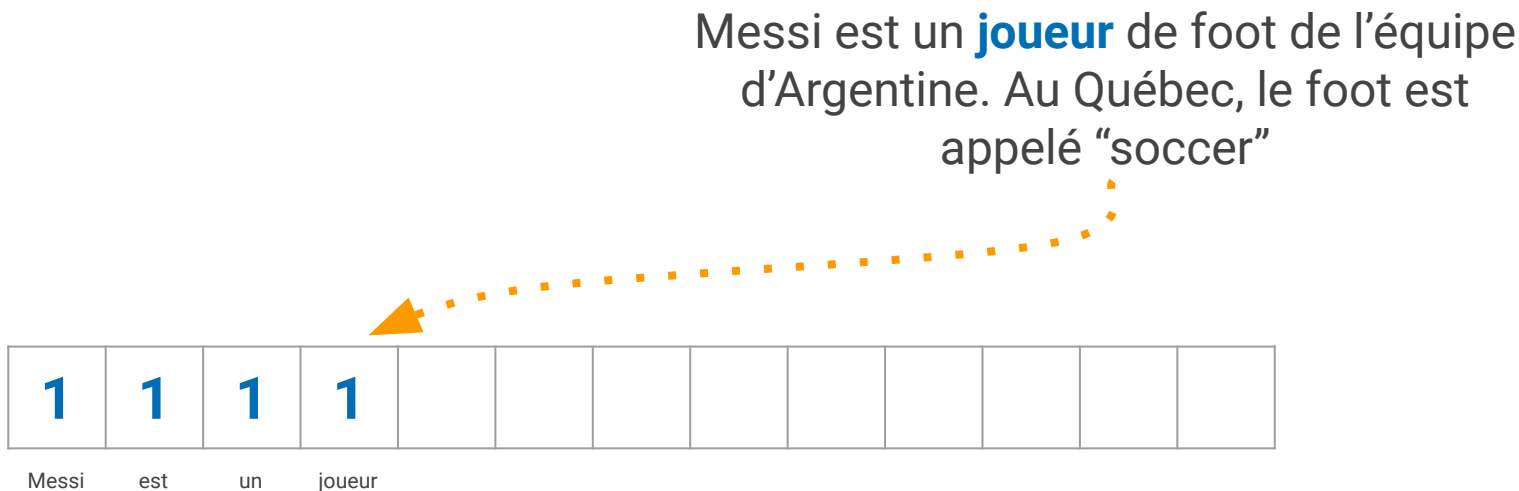
Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot



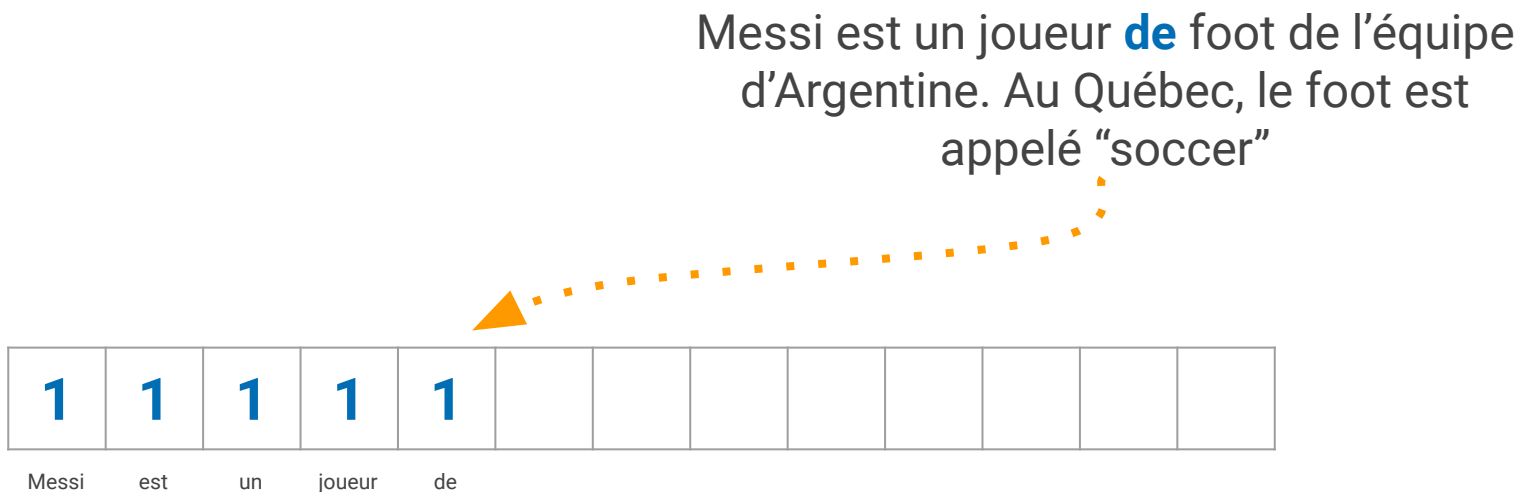
Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

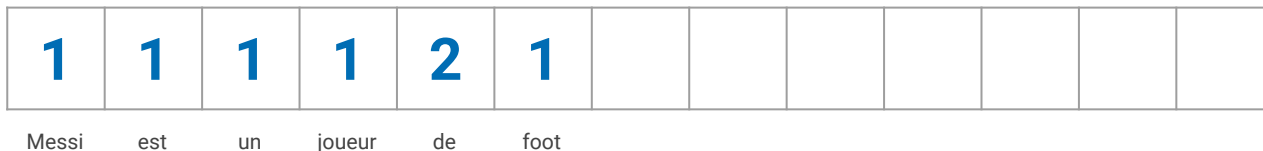
Messi est un joueur de **foot** de l'équipe d'Argentine. Au Québec, le foot est appelé "soccer"



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

Messi est un joueur de foot **de** l'équipe d'Argentine. Au Québec, le foot est appelé "soccer"



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

Messi est un joueur de foot de l'équipe
d'Argentine. Au Québec, le foot est
appelé "soccer"



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

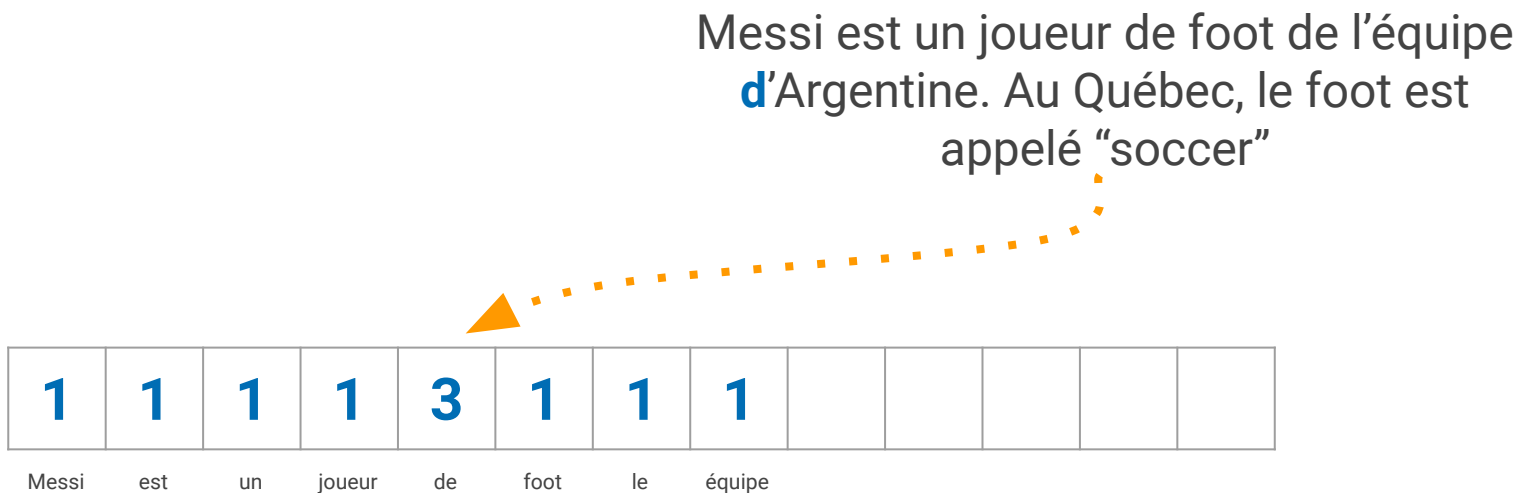
Messi est un joueur de foot de l'**équipe**
d'Argentine. Au Québec, le foot est
appelé "soccer"



1	1	1	1	2	1	1	1					
Messi	est	un	joueur	de	foot	le	équipe					

Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot



Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

Messi est un joueur de foot de l'équipe d'**Argentine**. Au Québec, le foot est appelé "soccer"



1	1	1	1	3	1	1	1	1				
Messi	est	un	joueur	de	foot	le	équipe	Argentine				

Représentation word count

- Modèle **sac de mots** (bag of words)
 - Ignore l'ordre des mots
 - Compte le nombre d'occurrence de chaque mot

Messi est un joueur de foot de l'équipe d'Argentine. Au Québec, le foot est appelé "soccer"

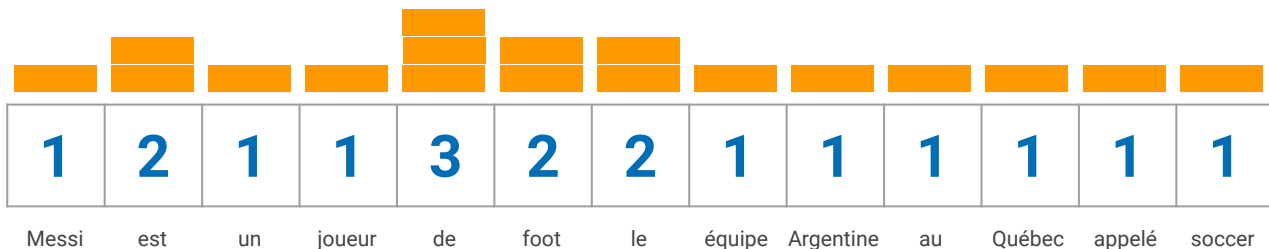
1	2	1	1	3	2	2	1	1	1	1	1	1
Messi	est	un	joueur	de	foot	le	équipe	Argentine	au	Québec	appelé	soccer

Représentation word count

■ Modèle **sac de mots** (bag of words)

- Ignore l'ordre des mots
- Compte le nombre d'occurrence de chaque mot
- Équivalent à un **histogramme**

Messi est un joueur de foot de l'équipe d'Argentine. Au Québec, le foot est appelé "soccer"



Représentation word count - Inconvénients

- Les mots **fréquents** dominent les mots **rares**
- **le / de / est / foot**
- **Messi / soccer / etc ...**

Messi est un joueur de foot de l'équipe d'Argentine. Au Québec, le foot est appelé "soccer" ...

1	2	1	1	3	2	2	1	1	1	1	1	1
Messi	est	un	joueur	de	foot	le	équipe	Argentine	au	Québec	appelé	soccer

Représentation TF-IDF

- La représentation TF-IDF met l'accent sur **les mots importants**
 - **Term Frequency** (TF) - Fréquence d'apparition dans le document (local)

TF =

					word	count					
--	--	--	--	--	------	-------	--	--	--	--	--

Représentation TF-IDF

- La représentation TF-IDF met l'accent sur **les mots importants**
 - **Term Frequency (TF)** - Fréquence d'apparition dans le document (local)

$$\text{TF} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|} \hline & & & & & \text{word count} & & & & & & & \\ \hline \end{array}$$



- **Inverse Document Frequency (IDF)** - Fréquence d'apparition (inverse) dans le corpus (global)

$$\text{IDF} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|} \hline & & & \log & \frac{\# \text{documents}}{1 + \# \text{documents utilisant le mot}} & & & & & & & \\ \hline \end{array}$$



Représentation TF-IDF

- Term Frequency Inverse Document Frequency (TF-IDF)

$$\text{TF-IDF} = \text{word count} \times \log \frac{\# \text{documents}}{1 + \# \text{documents utilisant le mot}}$$

TF =

--	--	--	--	--	--	--	--	--	--	--	--	--



x

IDF =

--	--	--	--	--	--	--	--	--	--	--	--	--



=

TF-IDF =

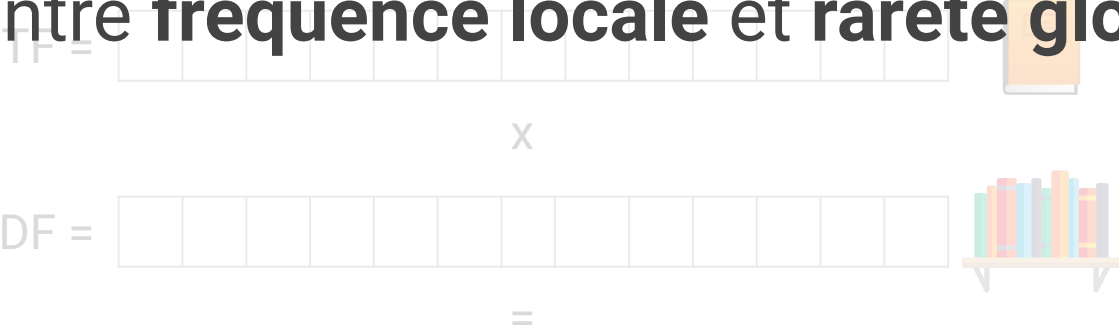
--	--	--	--	--	--	--	--	--	--	--	--	--

Représentation TF-IDF

- Term Frequency Inverse Document Frequency (TF-IDF)

$$\text{TF-IDF} = \text{word count} \times \log \frac{\# \text{documents}}{1 + \# \text{documents utilisant le mot}}$$

Compromis entre **fréquence locale** et **rareté globale**



TF-IDF =

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--



<https://github.com/mswawola-cegep/420-a58-sf-gr-12060.git>

Télécharger les données depuis **Teams**

Class Materials/people_wiki.zip

02-02-A1

Sommaire

1. Représentation des documents
2. Métriques de distance
3. Lectures et références

Métriques de distance: notion de “plus proche”

- En 1D, **distance euclidienne**

$$\text{distance}(x_i, x_q) = |x_i - x_q|$$

- Dans le cas de plusieurs dimensions:
 - Il existe plusieurs **fonctions de distance** intéressantes
 - Fléau de la dimension
 - Il peut être intéressant d'appliquer un **poids différent** à chaque dimension (feature weight)

Feature weight (1/3)

- Tout simplement parce que certaines dimensions / variables peuvent être plus **importantes** que d'autres !



Nombre de chambres
Nombre de salles de bain
Superficie habitable
Superficie du terrain
Nombre d'étages
Vue panoramique
Année de construction

....



Feature weight (1/3)

- Tout simplement parce que certaines dimensions / variables peuvent être plus **importantes** que d'autres !



Nombre de chambres

Nombre de salles de bain

Superficie habitable

Superficie du terrain

Nombre d'étages

Vue panoramique

Année de construction

....



Feature weight (2/3)

- Tout simplement parce que certaines dimensions / variables peuvent être plus **importantes** que d'autres !



Titre
Abstract
Texte principal
Sous-titres
Conclusion
....



Feature weight (2/3)

- Tout simplement parce que certaines dimensions / variables peuvent être plus **importantes** que d'autres !



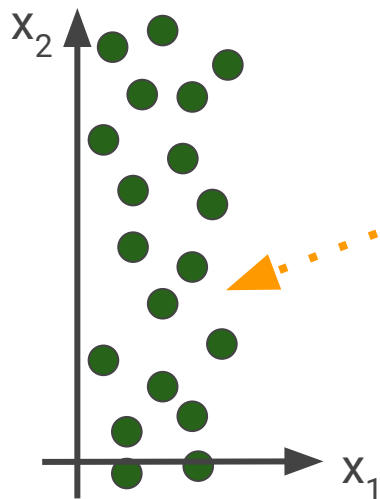
Titre
Abstract
Texte principal
Sous-titres
Conclusion

....



Mise à l'échelle revisitée

- Tout simplement parce que certaines dimensions / variables peuvent être plus **importantes** que d'autres !
- Également, certaines variables varient plus que d'autres



Les faibles variations de x_1 “comptent plus” que les fortes variations de x_2

Choisir des poids en fonction de l'étendue des variables

Distance euclidienne pondérée

- Formellement:

$$\text{distance}(x_i, x_q) = \sqrt{a_1(x_{i1} - x_{q1})^2 + \dots + a_n(x_{in} - x_{qn})^2}$$




a_1, a_2, \dots, a_n sont les poids appliqués aux différentes variables.
Permet de définir l'importance relative

Poids binaires

- Formellement:

$$\text{distance}(x_i, x_q) = \sqrt{a_1(x_{i1} - x_{q1})^2 + \cdots + a_n(x_{in} - x_{qn})^2}$$



Choisir 0 ou 1 comme poids permet de réaliser une sélection de variables

Poids binaires

- Formellement:

$$\text{distance}(x_i, x_q) = \sqrt{a_1(x_{i1} - x_{q1})^2 + \dots + a_n(x_{in} - x_{qn})^2}$$

L'ingénierie de données est couverte par le cours
420-A56-SF - Transformation et manipulation des données
Choisir 0 ou 1 comme poids pour réaliser une sélection de variables

Distance euclidienne (non pondérée)

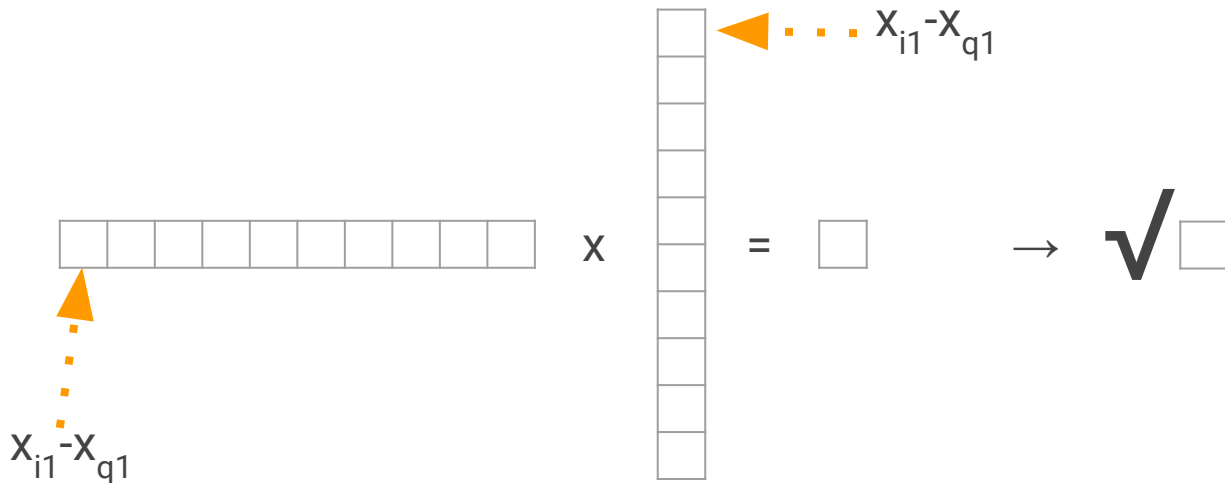
- La distance euclidienne peut-être définie par un produit de deux vecteurs:

$$\text{distance}(x_i, x_q) = \sqrt{(x_{i1} - x_{q1})^2 + \dots + (x_{in} - x_{qn})^2}$$

Distance euclidienne (non pondérée)

- La distance euclidienne peut-être définie par un produit de deux vecteurs:

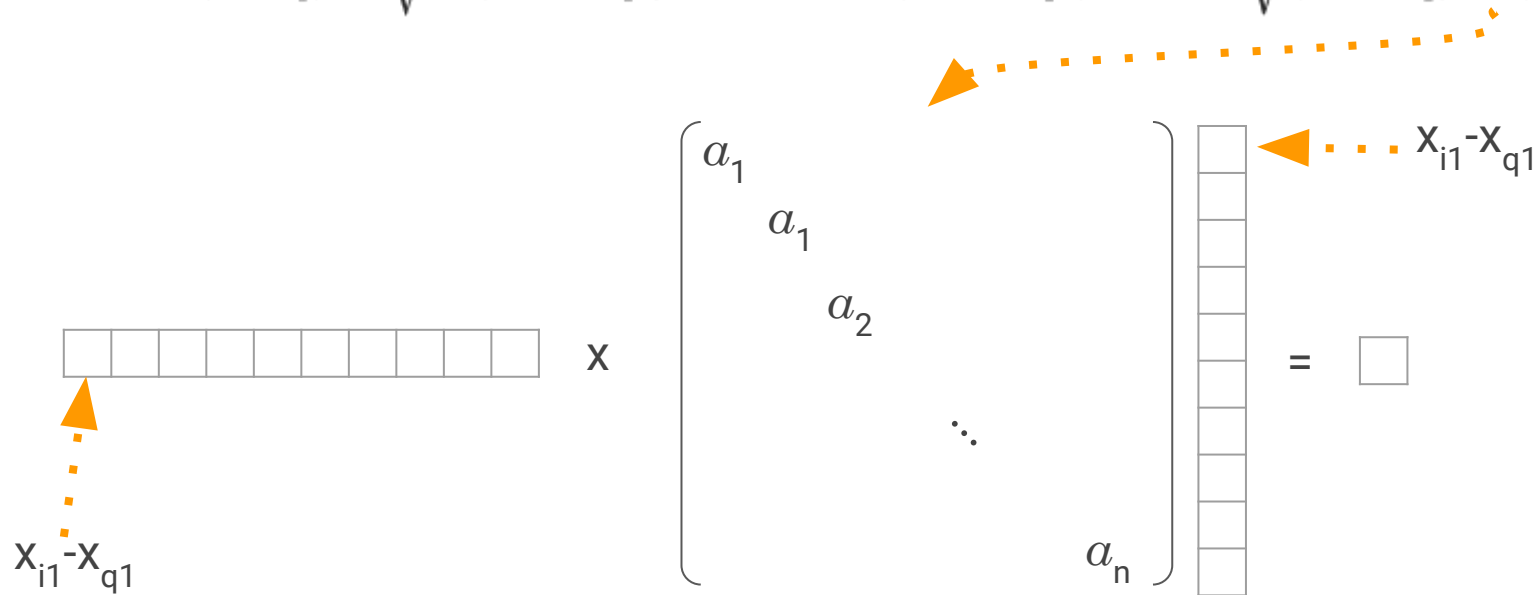
$$\text{distance}(x_i, x_q) = \sqrt{(x_{i1} - x_{q1})^2 + \cdots + (x_{in} - x_{qn})^2} = \sqrt{(x_i - x_q)^T (x_i - x_q)}$$



Distance euclidienne pondérée

- La dist. euclidienne pondérée peut-être définie par un produit de deux vecteurs:

$$\text{distance}(x_i, x_q) = \sqrt{a_1(x_{i1} - x_{q1})^2 + \dots + a_n(x_{in} - x_{qn})^2} = \sqrt{(x_i - x_q)^T A (x_i - x_q)}$$



Mesure de similarité

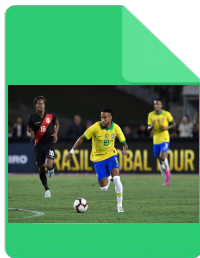


X_q

1	0	0	0	3	5	0	0	1	2
---	---	---	---	---	---	---	---	---	---



$$\text{Similarité} = \mathbf{x}_i^T \mathbf{x}_q$$



X_i

3	0	0	0	0	2	0	0	1	1
---	---	---	---	---	---	---	---	---	---

Que vaut est la similarité ?

Mesure de similarité



x_q

1	0	0	0	3	5	0	0	1	2
---	---	---	---	---	---	---	---	---	---



$$\text{Similarité} = x_i^T x_q$$



x_i

0	0	4	7	1	0	2	0	1	0
---	---	---	---	---	---	---	---	---	---

Que vaut est la similarité ?

Normalisation de la similarité

$$\begin{aligned} \text{Similarité} = \mathbf{x}_i^T \mathbf{x}_q &\xrightarrow{\text{Normalisation}} \frac{\mathbf{x}_i^T \mathbf{x}_q}{\mathbf{x}_i^T \mathbf{x}_i \mathbf{x}_q^T \mathbf{x}_q} \quad \longleftrightarrow \quad \left(\frac{\mathbf{x}_i}{\mathbf{x}_i^T \mathbf{x}_i} \right)^T \frac{\mathbf{x}_q}{\mathbf{x}_q^T \mathbf{x}_q} \\ &\quad \downarrow \\ a^T b = \|a\| \times \|b\| \times \cos(\theta) &\xrightarrow{\quad} \frac{\mathbf{x}_i^T}{\|\mathbf{x}_i\|} \frac{\mathbf{x}_q}{\|\mathbf{x}_q\|} = \cos(\theta) \\ &\quad \text{Similarité cosinus !} \end{aligned}$$

Exercice: normaliser le vecteur suivant



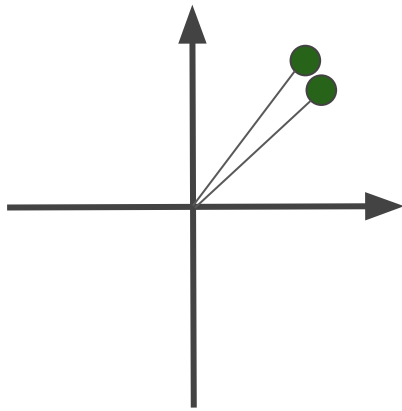
1	0	0	0	3	5	0	0	1	0
---	---	---	---	---	---	---	---	---	---



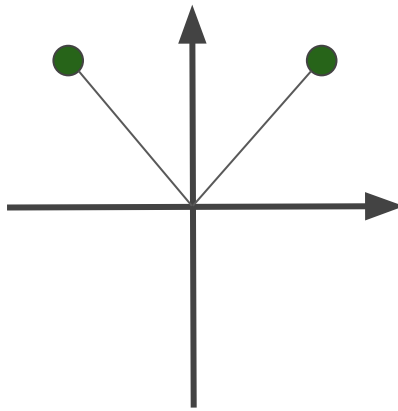
$1/6$	0	0	0	$3/6$	$5/6$	0	0	$1/6$	0
-------	---	---	---	-------	-------	---	---	-------	---

Similarité cosinus

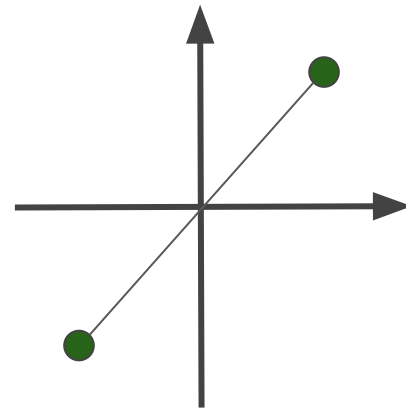
- D'une manière générale, la similarité cosinus est comprise entre -1 et 1



θ très petit



$\theta \sim 90^\circ$

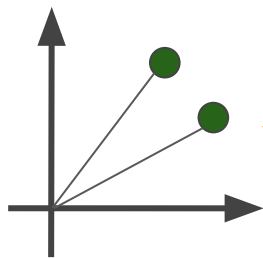


$\theta \sim 180^\circ$

Quels sont les cosinus de ces angles ?

Similarité cosinus

- D'une manière générale, la similarité cosinus est comprise entre -1 et 1



... Pour TF-IDF, θ dans ce cadran
similarité comprise entre 0 et 1

**Question
pourquoi ?**

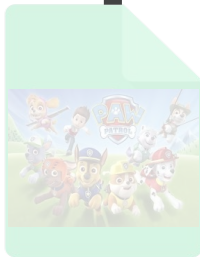
- Nous pouvons aussi définir la distance cosinus: **distance = 1 - similarité**

Normaliser ou pas ?



1 0 0 0 3 5 0 0 1 0

4



0 0 4 7 1 0 2 0 1 0

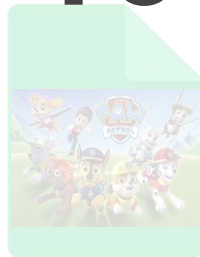
x2

Oui !



2 0 0 0 6 10 0 0 2 0

16



0 0 8 14 2 0 4 0 2 0

x2

Normaliser ou pas ?



Document
volumineux



Tweet



La normalisation peut rendre des
objets différents similaires

Non ...



Document
volumineux



Document
volumineux

Autres métriques de distance

- Basée sur la corrélation
- Mahalanobis
- Rank-based
- Manhattan
- Jaccard
- Hamming
- ...

Combinaison de métriques - Exemple

- Texte des documents
 - Distance cosinus
- Nombre de lectures des documents
 - Distance euclidienne
- Affecter un poids différent à chaque métrique



<https://github.com/mswawola-cegep/420-a58-sf-gr-12060.git>

02-02-A2

Sommaire

1. Représentation des documents
2. Métriques de distance
3. Lectures et références

Lectures et références

[1] Machine Learning: Clustering and Retrieval - Emily Fox & Carlos Guestrin - University of Washington