

(part 2/4)

- **Motivation**
 - ubiquity of information networks
 - applications (in particular to IR)
 - importance of indexing
- **Representation of documents**
 - sparse representations
 - dense representations
 - topic models
- **Network analysis**
 - spectral clustering, modularity
 - representation learning for graphs
- **Analyzing information networks**
 - Graph Neural Networks

2

- On appelle **document** un objet numérique qui véhicule un ensemble d'informations souvent structurées :



altre
medias

sans oublier les **méta-données** :

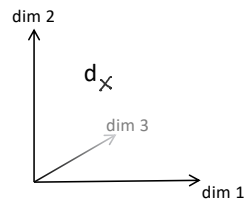
- auteur du document
- date de publication
- etc.

Network analysis for information retrieval, M2 MALIA-MIASHS, Julien Velcin

- On appelle **corpus** un ensemble de documents. Le corpus est souvent associé à une structure (par ex. hyperliens, citations, etc.).

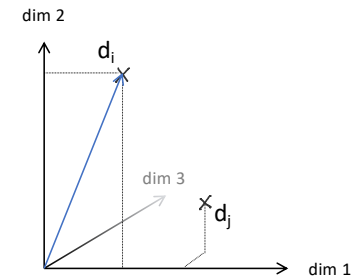
Représentation d'un document

- espace des mots (vocabulaire)
(avec différents types de pondération : TF, TFxIDF, OKAPI BM25)
- espace sémantique de faible dimension :
 - approche de plongement (sentence/document embedding)
 - approche thématique



5

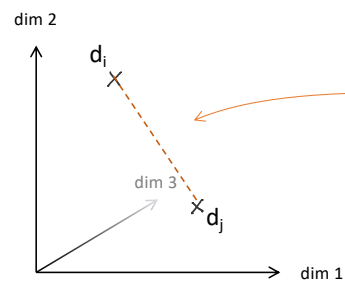
Représenter d dans un espace vectoriel



- Les axes peuvent être :
- des mots
 - des thématiques
 - des variables latentes

6

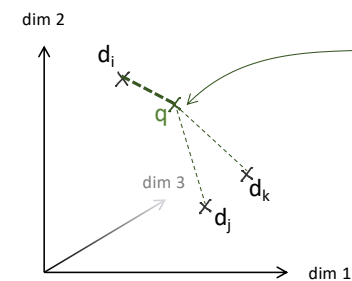
Comparer dans un espace vectoriel



- Différentes sortes de **mesures** :
- distance euclidienne
 - produit scalaire
 - cosinus
 - ...

7

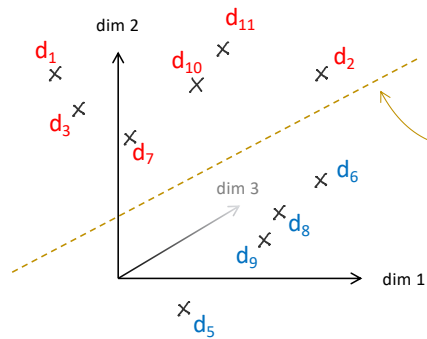
Comparer dans un espace vectoriel



- On peut considérer une **requête q** comme un pseudo-document et donc trouver les documents les plus proches

8

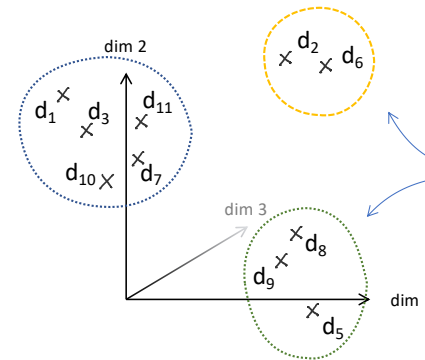
Classer dans un espace vectoriel



On peut chercher la **frontière** entre les classes dans cet espace (ici linéaire)

9

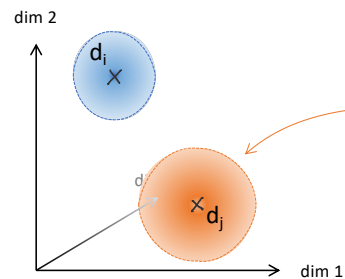
Catégoriser dans un espace vectoriel



On peut chercher des **catégories** de manière non supervisée (clustering)

10

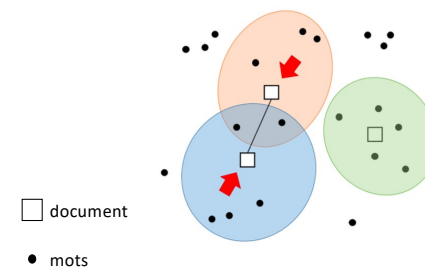
Aller plus loin : prise en compte de l'incertitude



On peut représenter un objet (document, auteur...) comme une **distribution de probabilité** dans l'espace

11

Aller plus loin : prise en compte de méta-données



On rapproche les représentations de documents qui sont **connectés** (relation de citation, même auteur...)

12

ML tasks with information networks

- Node / edge classification
- Link prediction
- Community detection