Network analysis for information retrieval

Julien Velcin

Master MALIA-MIASHS

2023-2024

(part 2/4)

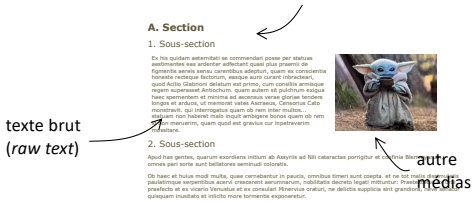
Outline

- 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

Representation of documents

Quelques définitions

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



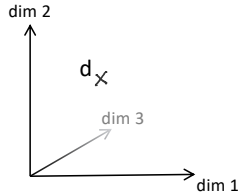
sans oublier les **méta-données** :
- auteur du document
- date de publication

etc.

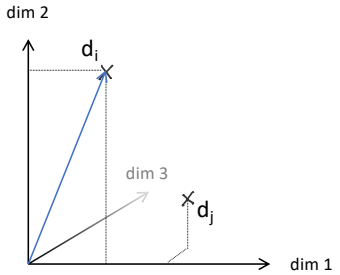
- On appelle **corpus** un ensemble de documents. Le corpus est souvent associé à une structure (par ex. hyperliens, citations, etc.).

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

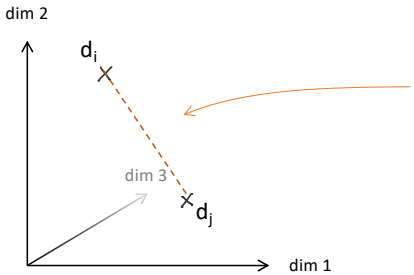


Représenter **d** dans un espace vectoriel



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

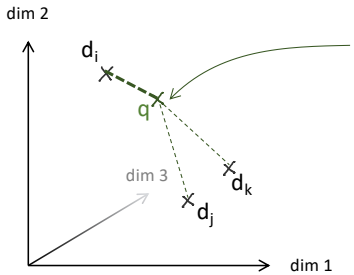
Comparer dans un espace vectoriel



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

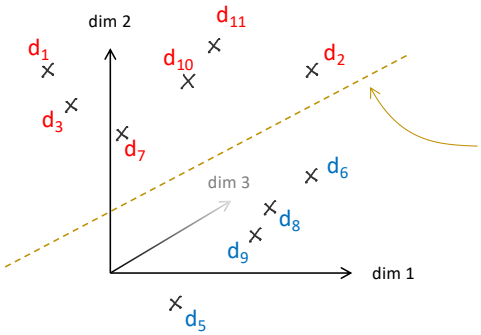
...

Comparer dans un espace vectoriel



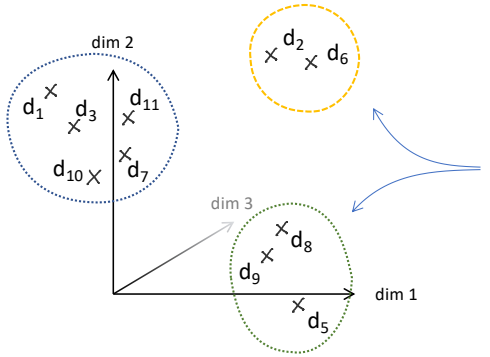
On peut considérer une **requête q** comme un pseudo-document et donc trouver les documents les plus proches

Classer dans un espace vectoriel



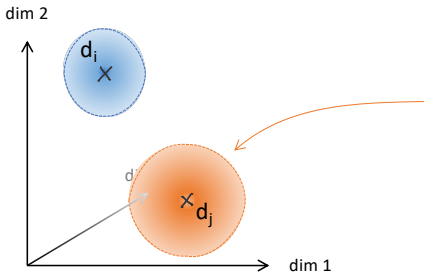
On peut chercher la **frontière** entre les classes dans cet espace (ici linéaire)

Catégoriser dans un espace vectoriel



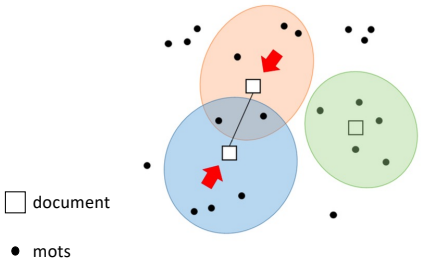
On peut chercher des **catégories** de manière non supervisée (clustering)

Aller plus loin : prise en compte de l'incertitude



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

Aller plus loin : prise en compte de méta-données



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

Document Network Projection in Pretrained Word Embedding Space
A. Gourru, J. Velcin, J. Jacques and A. Guille. ECIR 2020. https://github.com/AntoineGourru/DNEmbedding

ML tasks we usually solve

- Node / edge classification
- Link prediction
- Community detection