

Gestion des données et accès à l'information

Recherche d'information

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

Master Statistiques
Sorbonne Université

2024-2025

Définitions

Corpus

- ▶ Ensemble des documents considérés

Requête

- ▶ Entrée utilisateur
- ▶ Mots-clés
- ▶ Phrase en langue naturelle
- ▶ Document

Pertinence

- ▶ Document pertinent: document intéressant par rapport à la requête
- ▶ Pertinence d'un document: score d'adéquation entre le document et le requête

Recherche d'information

Recherche

- ▶ Extraction des documents les plus pertinents
- ▶ Éventuellement, de fragments des documents

Sortie

- ▶ Liste de documents pertinents
- ▶ Passages pertinents
- ▶ Trié par pertinence, et relativement court

Indexation

- ▶ Extraction des données
- ▶ Représentation des documents
- ▶ Stockage

Ambiguïtés



Créer un

Java

61 langues ▾

Article [Discussion](#) Lire [Modifier](#) [Modifier le code](#) [Voir l'histoire](#) [Outils](#) ▾

- Cette *page d'homonymie* répertorie les différents sujets et articles partageant un même nom.
- Cet article possède des *paronymes*, voir *Jara* et *Jawa*.

Toponyme [modifier | modifier le code]

Java est un nom de lieu notamment porté par :

- Java est une île **indonésienne** ;
 - les **Javanais** sont le **groupe ethnique** majoritaire de l'île ;
 - le **Javanais** est leur langue ;
 - le **café de Java** est un caféier qui provient de cette île ;
- Java est le nom translittéré d'un district de **Géorgie**, transcrit en **Djava** ;
- Java est également le nom de plusieurs villes des **États-Unis** ;
 - Java dans l'État de **New York** ;
 - Java dans le **Dakota du Sud** ;
 - Java ^(en) en **Virginie** dans le comté de **Pennsylvanie** ;
- Java, un hameau de **Bas-Oha** et une petite île sur la Meuse (**Belgique**) ;
- Java, village de **Sao Tomé-et-Principe**.

Informatique [modifier | modifier le code]

- Java, nom d'une technologie mise au point par **Sun Microsystems** (racheté par **Oracle** en 2010) qui permet de produire des logiciels indépendants de toute architecture matérielle. Cette technologie s'appuie sur différents éléments qui, par abus de langage, sont souvent tous appelés Java :

Sur les autres projets Wikimedia :
 [Java](#), sur le Wiktionnaire

Recherche d'information sur le web

Google et cie

Entrée

- ▶ List de mots-clés
- ▶ Éventuellement: formule logique

Sortie

- ▶ Liste de liens

Principes de fonctionnement

- ▶ Indexation des documents
- ▶ Analyse du contenu des pages
- ▶ Évaluation de la pertinence des documents
- ▶ Historiquement: *PageRank* (une page de référence est une page vers qui pointe beaucoup d'autres pages)

Autres exemples

Recherche d'image

- ▶ Entrée: image
- ▶ Sortie: liste d'images

Question answering

- ▶ Entrée: question en langue naturelle
- ▶ Sortie: réponse rédigée à partir des documents pertinents

Requête en langue naturelle

- ▶ Nécessite de comprendre la requête
- ▶ Peut-être un peu plus intuitif que les mots-clés
- ▶ Mais surtout, plus facile de préciser le contexte

Contre-exemples

Base de donnée classique

- ▶ Documents structurés: tables, colonnes, etc

ChatGPT

- ▶ Génération de la réponse la plus probable
- ▶ Pas de recherche
- ▶ Connaissances fixées lors de l'apprentissage

Modèles de recherche

Modèle booléen pour le texte

- ▶ Formule logiques entre les documents et la requête
- ▶ Pas d'ordonnancement possible

Modèle probabiliste

- ▶ Score de probabilité pour un document d'être pertinent sachant la requête
- ▶ Okapi BM25

Modèle vectoriel

- ▶ Documents et requête dans un espace vectoriel
- ▶ Distance (L2 ou cosinus)

Recherche directe

Version simpliste

- ▶ Parcourir tous les documents
- ▶ Comparer chaque document à la requête
- ▶ Prendre ceux avec le meilleur score
- ▶ Variante de k -plus-proche-voisin

Limites

- ▶ Choix de la similarité ?
- ▶ Pas forcément adapté pour des documents longs

Index inversé

Représentation directe

- ▶ Dictionnaire de documents
- ▶ Clé: un document
- ▶ Valeur: liste des termes contenus dans le document

Index inversé

- ▶ Dictionnaire de termes
- ▶ Clé: un terme
- ▶ Valeur: list des documents qui contiennent ce terme
- ▶ Adapté pour la recherche à partir de mot-clé: filtrage
- ▶ On ne regarde que les documents qui contiennent les mots-clés de la requête

Traitement automatique des langues (TAL)

- ▶ Natural Language Processing (NLP)
- ▶ Langue naturelle

Objectifs

- ▶ Compréhension de l'écrit
- ▶ Génération de documents textuels
- ▶ Interaction homme-machine en langue naturelle
- ▶ Souvent un peu tout ça en même

(écrit uniquement, pas de voix)

Représenter des mots

Ordre arbitraire

Le	chat	mange	des	croquettes
0	1	2	3	4

Comparer les mots

- ▶ Pas vraiment de distance pertinente
- ▶ "le" et "chat" presque à la même distance
- ▶ "chat" et croquette" très éloignés
- ▶ Pas de sémantique intéressante

One-hot encoding

Toujours la numérotation arbitraire

Vecteur binaire

Le	1	0	0	0	0
chat	0	1	0	0	0
mange	0	0	1	0	0
des	0	0	0	1	0
croquettes	0	0	0	0	1

Comparer les mots

- ▶ Même distance entre chaque mot
- ▶ Un peu mieux qu'avant
- ▶ Mais toujours pas vraiment de sémantique

Au niveau d'un document

Comment représenter un texte entier ?

Sac de mots

- ▶ Compter les mots

	Mot 1	...	Mot j	...	Mot p
Document 1					
...					
Document j			s_{ij}		
...					
Document N					

s_{ij} Nombre d'apparitions du mot j dans le document i

Variantes

- ▶ Fréquence au lieu du comptage

TF-IDF

Term Frequency TF

$tf(t_i, d)$ = nombre d'occurrences de t_i dans le document d

Inverse Document frequency IDF

$$idf(t_i) = \log \frac{1 + N}{1 + df(t_i)}$$

- ▶ $df(t_i)$: nombre de documents contenant t_i
- ▶ N : nombre de documents

TF-IDF

Pondération de chaque mot:

$$tf(t_i, d)idf(t_i)$$

Propriétés de TF-IDF

- ▶ Term frequency: dépend de la taille du document
- ▶ idf: fréquence inverse, tend vers 0 si t_i apparaît dans tous les documents

Conséquences

- ▶ Mot qui apparaît partout: négligeable
- ▶ Document plus long: scores plus élevés

Sémantique

- ▶ Mettre en lumière les mots les plus déterminants pour le sens du document

Pré-traitements

Mots inutiles

- ▶ Articles et autres ?
- ▶ Ponctuation ?
- ▶ Mots extrêmement fréquents ?

Variantes d'un mot

- ▶ Singulier/Pluriel, masculin/féminin
- ▶ Conjugaison
- ▶ Chiffres

Lemmatisation

- ▶ Recherche d'une unité lexicale élémentaire

Spécificités des prétraitements

Dépend de la langue

- ▶ Allemand: *Arbeiterunfallversicherungsgesetz* (loi sur l'assurance des accidents du travail)

Dépend de la tâche

- ▶ Thème général du document: regarder juste les radicaux
- ▶ Traduction: besoin des pluriels et des conjugaison

Modèles de mots

Représentation dans un espace vectoriel

- ▶ Un mot \rightarrow un point
- ▶ Word2vec, Glove, FastText

Avec de la sémantique

- ▶ Mots proches d'un point de vue sémantiques proches dans l'espace de représentation

Méthode

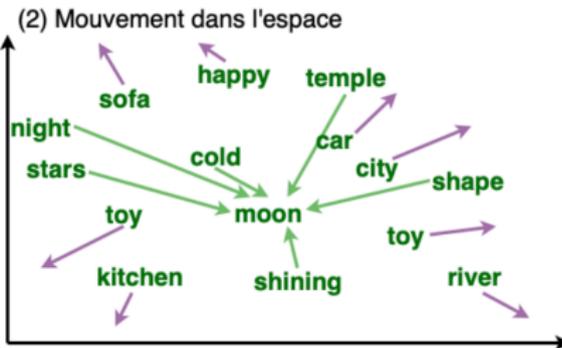
- ▶ Apprentissage de représentation: construction d'un espace latent
- ▶ Loss: rapprocher les paires de mots au sens similaire, éloigner les autres
- ▶ Sens similaire: le contexte, les mots autour

Contexte

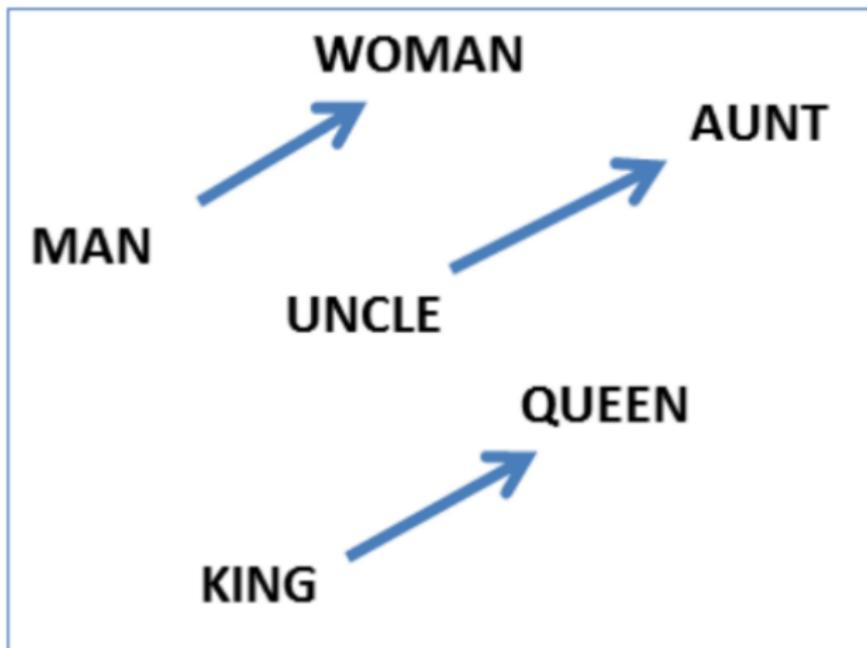
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The splash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

Apprentissage

he curtains open and the moon sh: (2) Mouvement dans l'espace
ars and the cold , close moon "
rough the night with the moon sh:
made in the light of the moon . :
surely under a crescent moon , t
sun , the seasons of the moon ? l
m is dazzling snow , the moon ha:
un and the temple of the moon , c
in the dark and now the moon ri:
bird on the shape of the moon ove
But I could n't see the moon OR the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The splash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind



Opérations dans l'espace latent



Modèles de langue

Entraînement

- ▶ Apprentissage supervisé: étiquetage coûteux
- ▶ Auto-supervision: étiquetage gratuit

Tâche d'entraînement

- ▶ Prédiction des mots masqués
- ▶ *Le chat X des croquettes*

Modèle génératif

- ▶ Prédiction des mots suivants
- ▶ Compléter des phrases
- ▶ Dialogue: la réponse complète ce qui précède

Représentation d'un document

- ▶ Espace de très grande dimension 100 - 10000