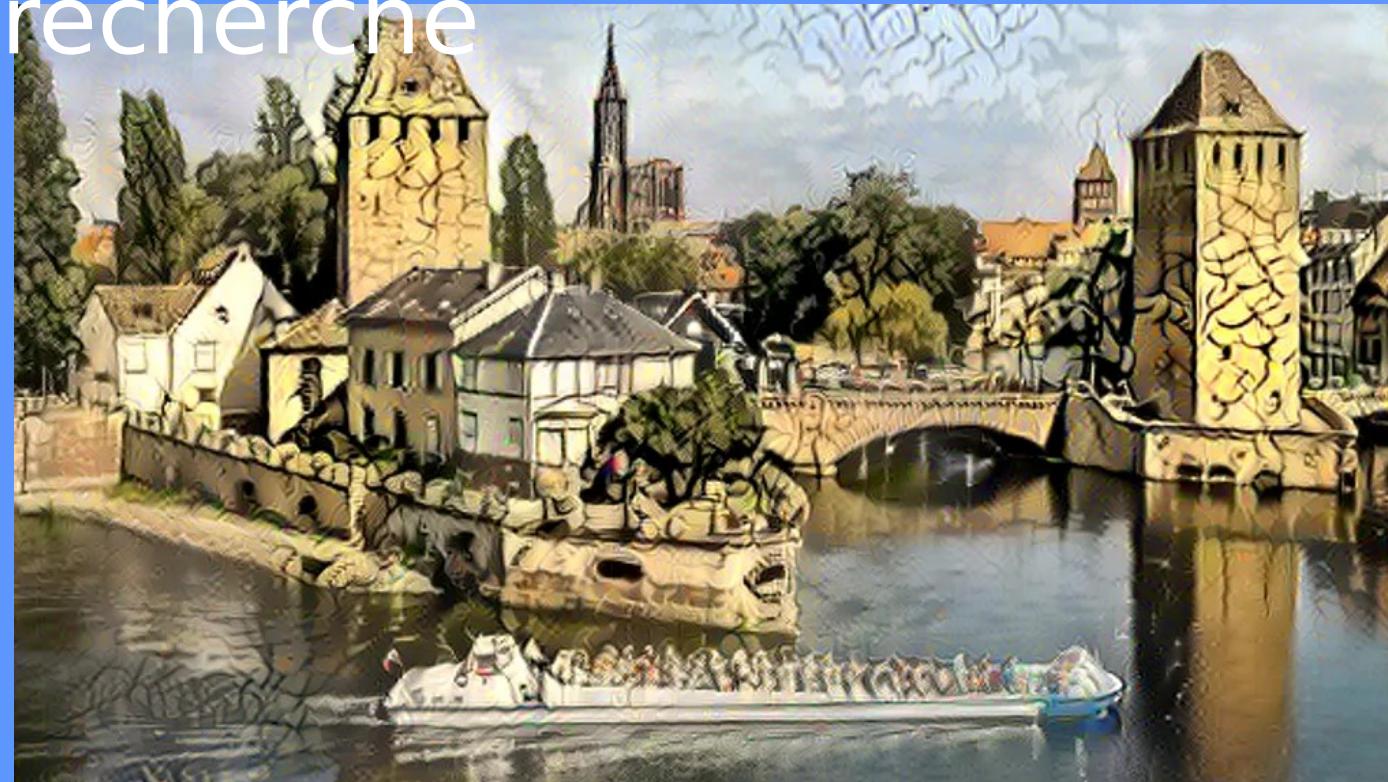


Strasbourg deep-learning Meetup

Learning to rank et
adaptation aux moteurs
de recherche



François Weber

AI engineer chez Qwant - bientôt chez
Soprema

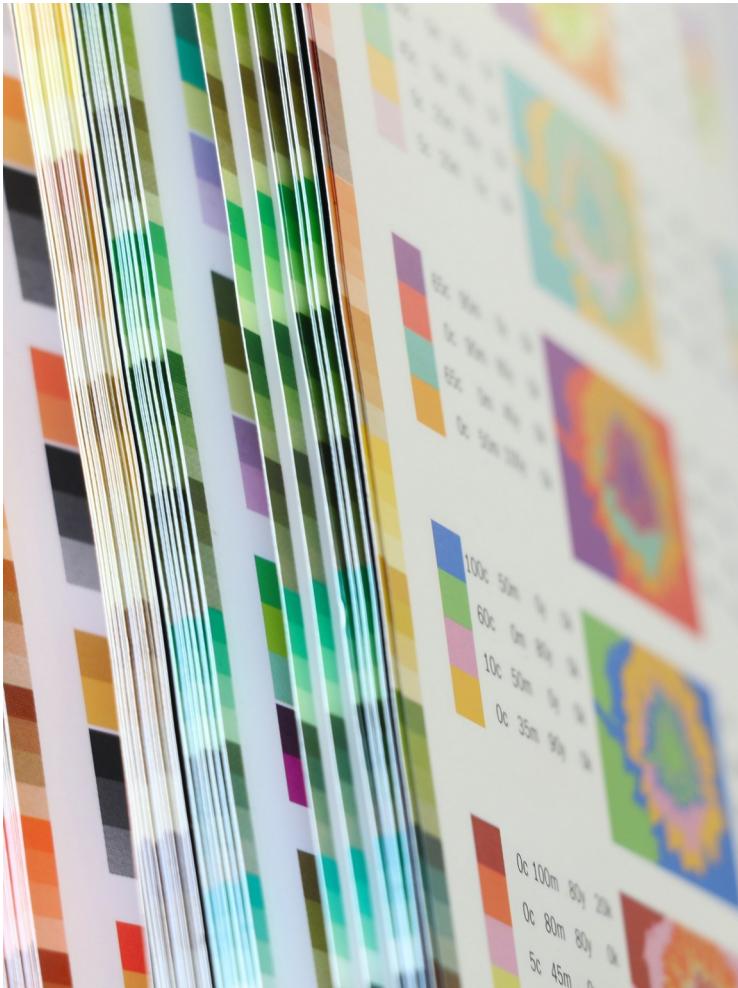
francois.weber@protonmail.com

<https://www.linkedin.com/in/fweber-89/>

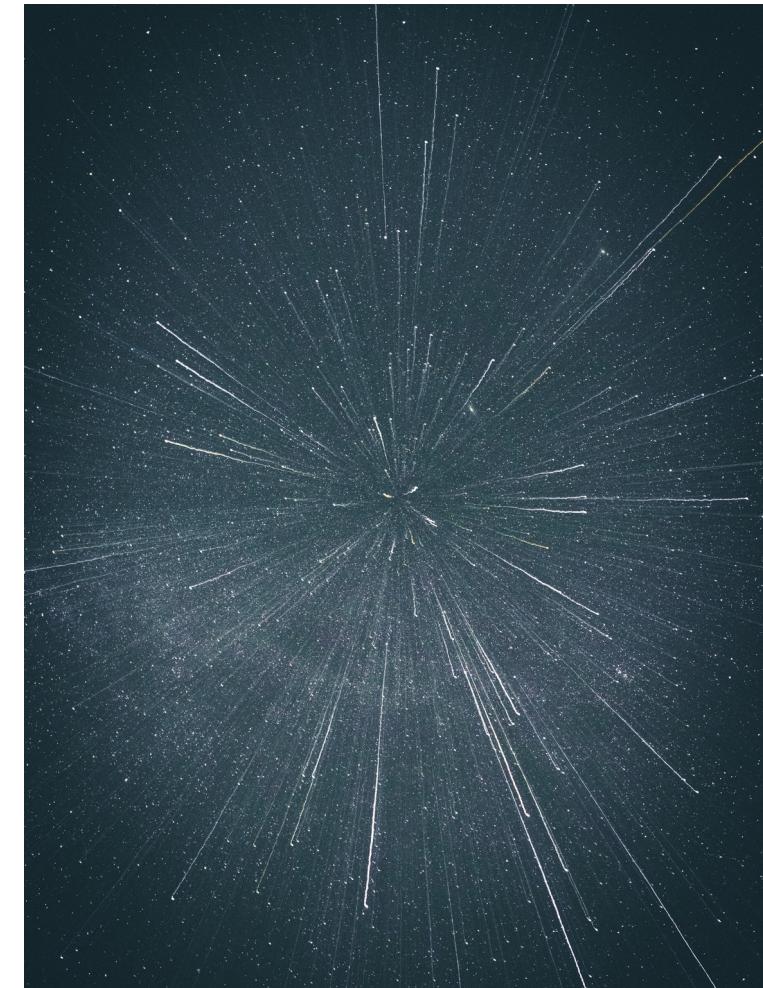




Architecture d'un moteur de recherche web



Nécessité
d'ordonnancement
optimal : le learning to rank



L'usage du NLP dans les moteurs modernes



Architecture d'un moteur de recherche web

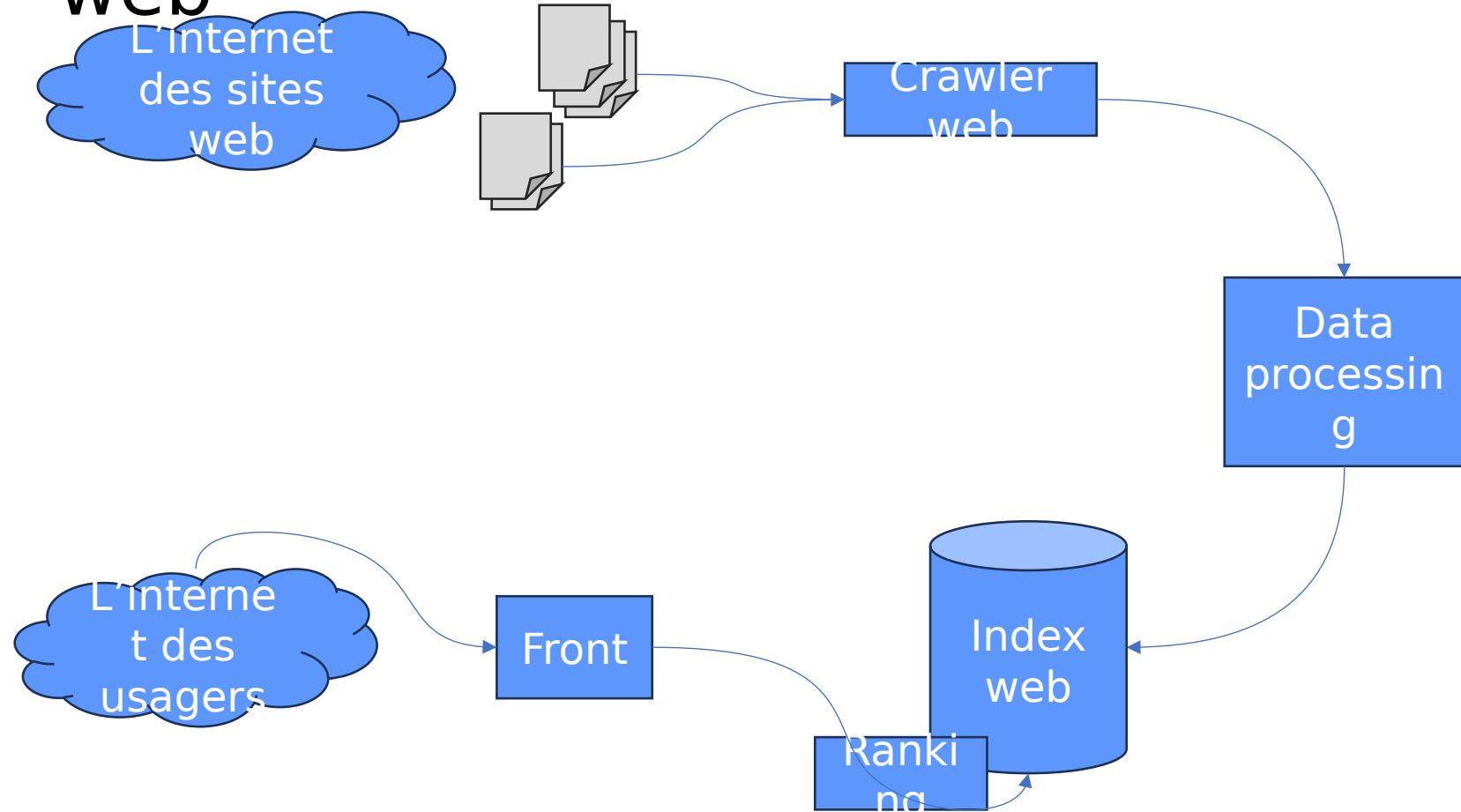
Nécessité d'ordonnancement optimal : le learning to rank



L'usage du NLP dans les moteurs modernes

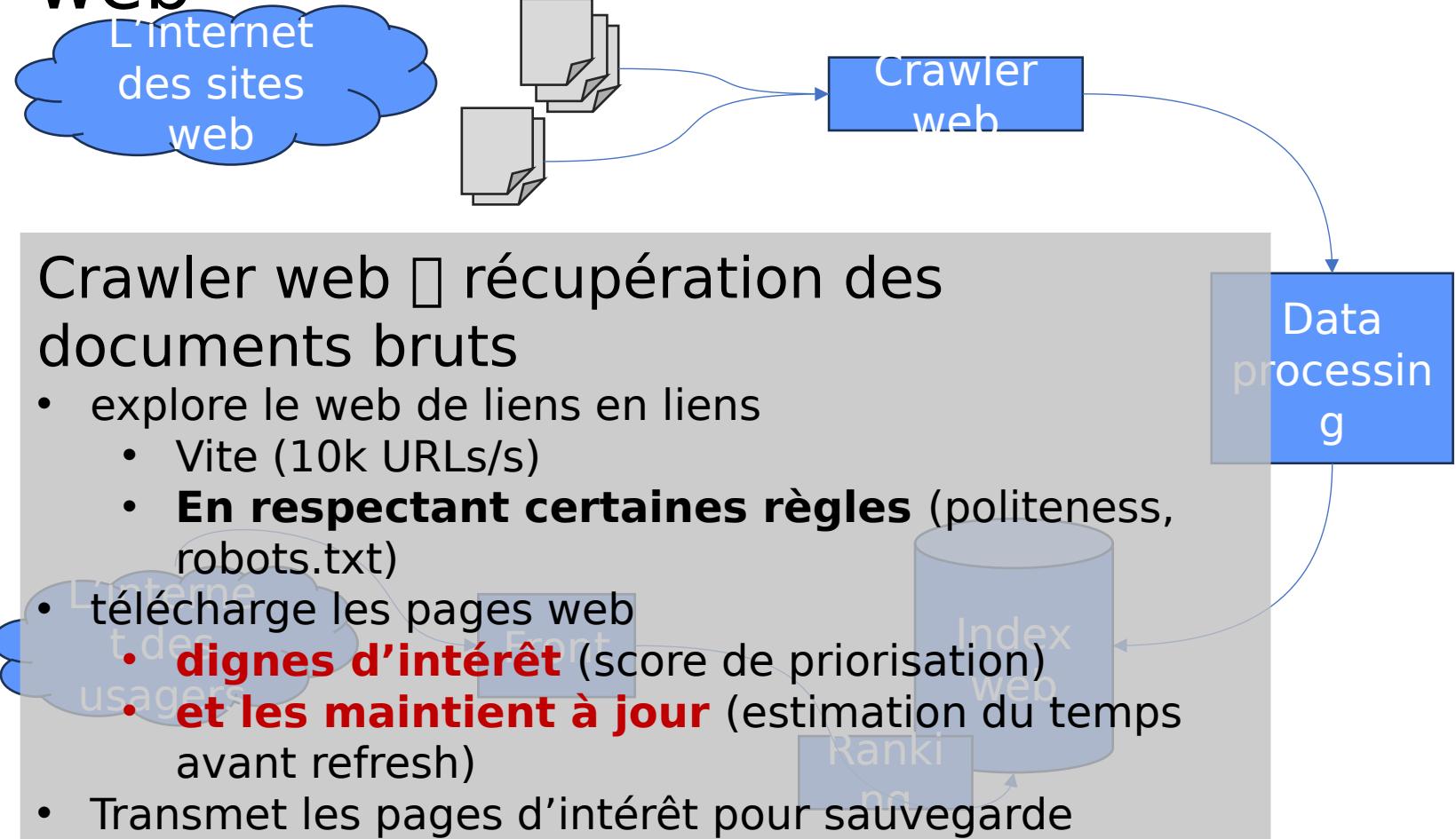


Architecture d'un moteur de recherche web





Architecture d'un moteur de recherche web

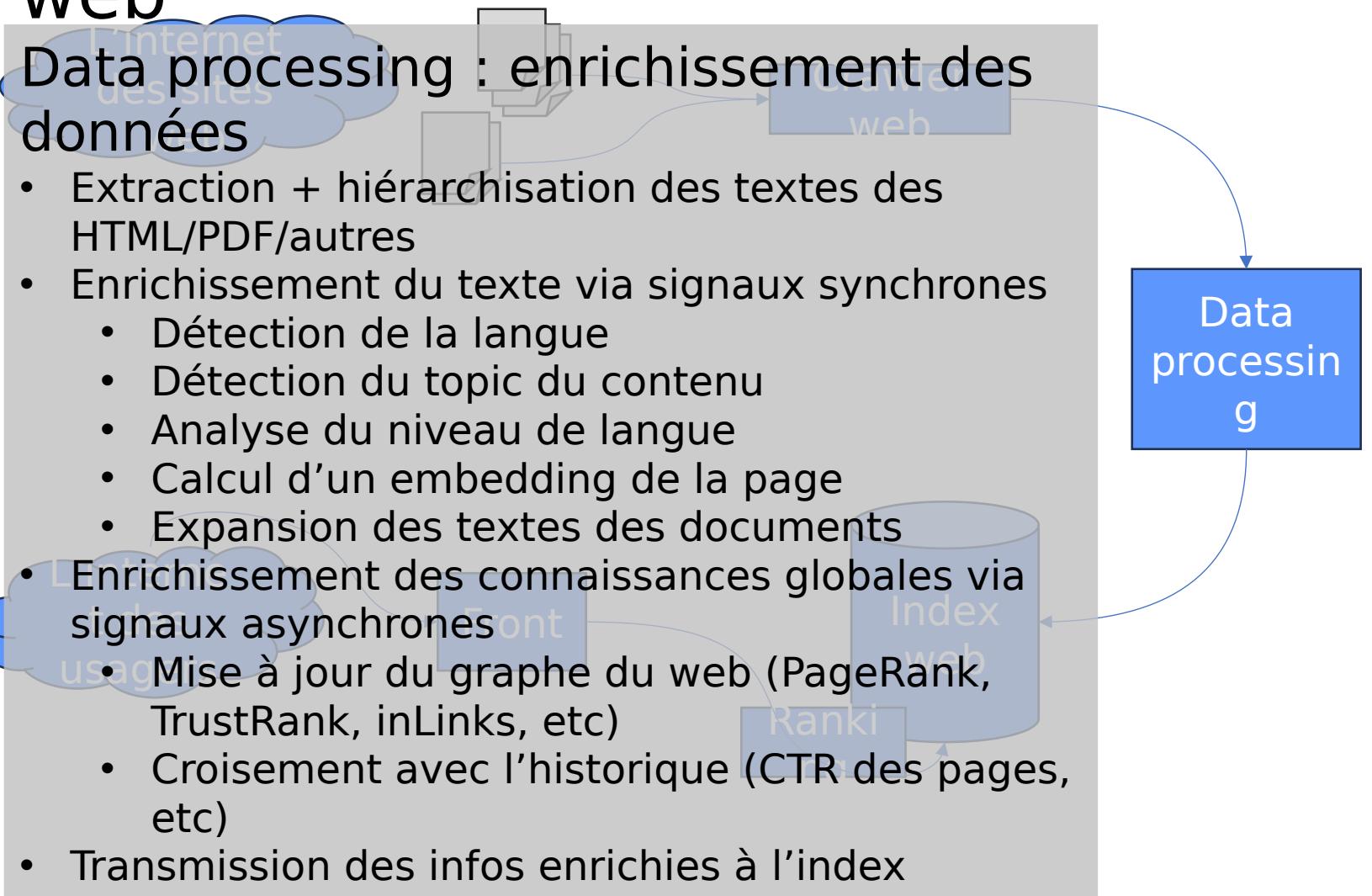




Architecture d'un moteur de recherche web

Data processing : enrichissement des données

- Extraction + hiérarchisation des textes des HTML/PDF/autres
- Enrichissement du texte via signaux synchrones
 - Détection de la langue
 - Détection du topic du contenu
 - Analyse du niveau de langue
 - Calcul d'un embedding de la page
 - Expansion des textes des documents
- Enrichissement des connaissances globales via signaux asynchrones
 - Mise à jour du graphe du web (PageRank, TrustRank, inLinks, etc)
 - Croisement avec l'historique (CTR des pages, etc)
 - Transmission des infos enrichies à l'index

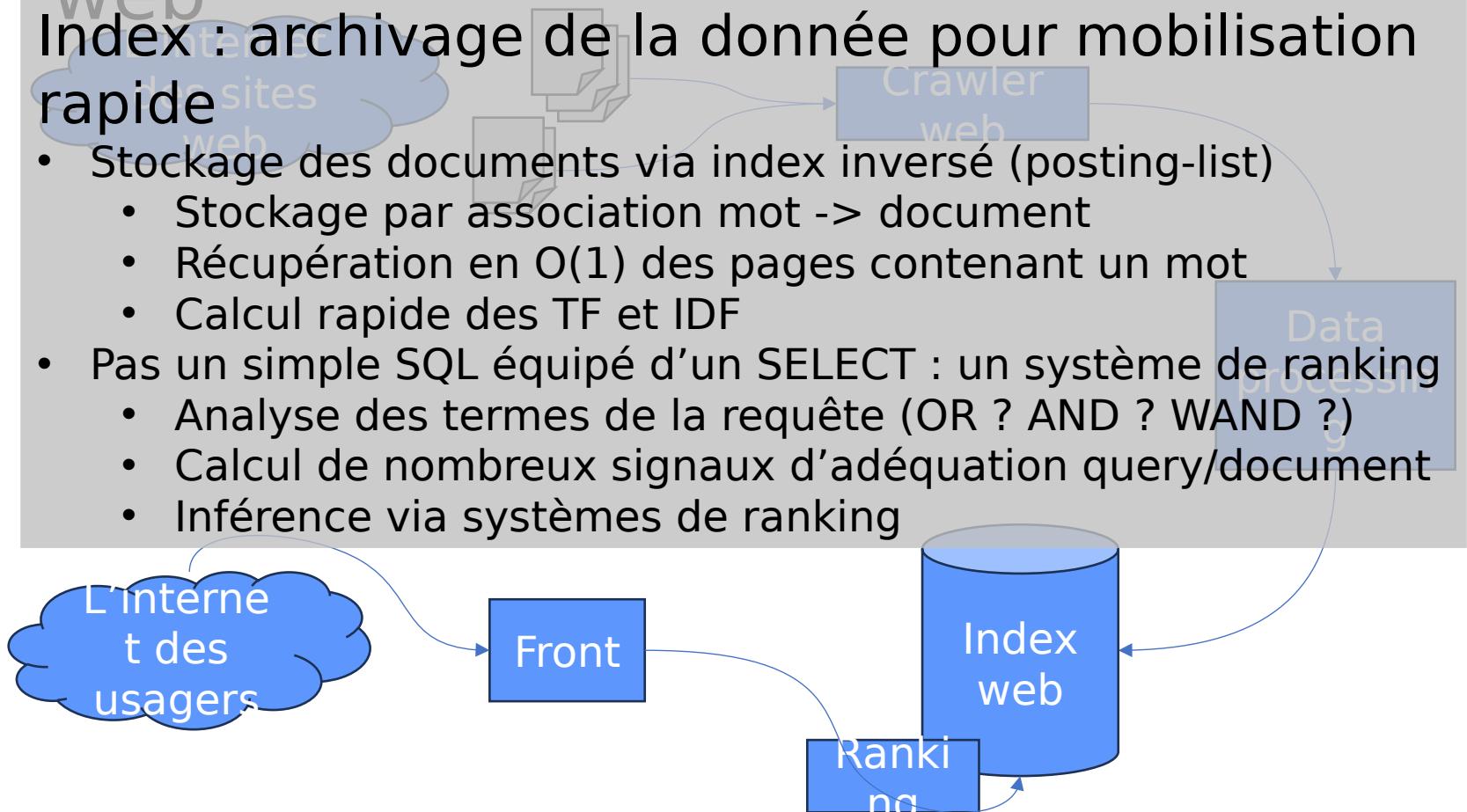




Architecture d'un moteur de recherche web

Index: archivage de la donnée pour mobilisation rapide

- Stockage des documents via index inversé (posting-list)
 - Stockage par association mot -> document
 - Récupération en $O(1)$ des pages contenant un mot
 - Calcul rapide des TF et IDF
- Pas un simple SQL équipé d'un SELECT : un système de ranking
 - Analyse des termes de la requête (OR ? AND ? WAND ?)
 - Calcul de nombreux signaux d'adéquation query/document
 - Inférence via systèmes de ranking



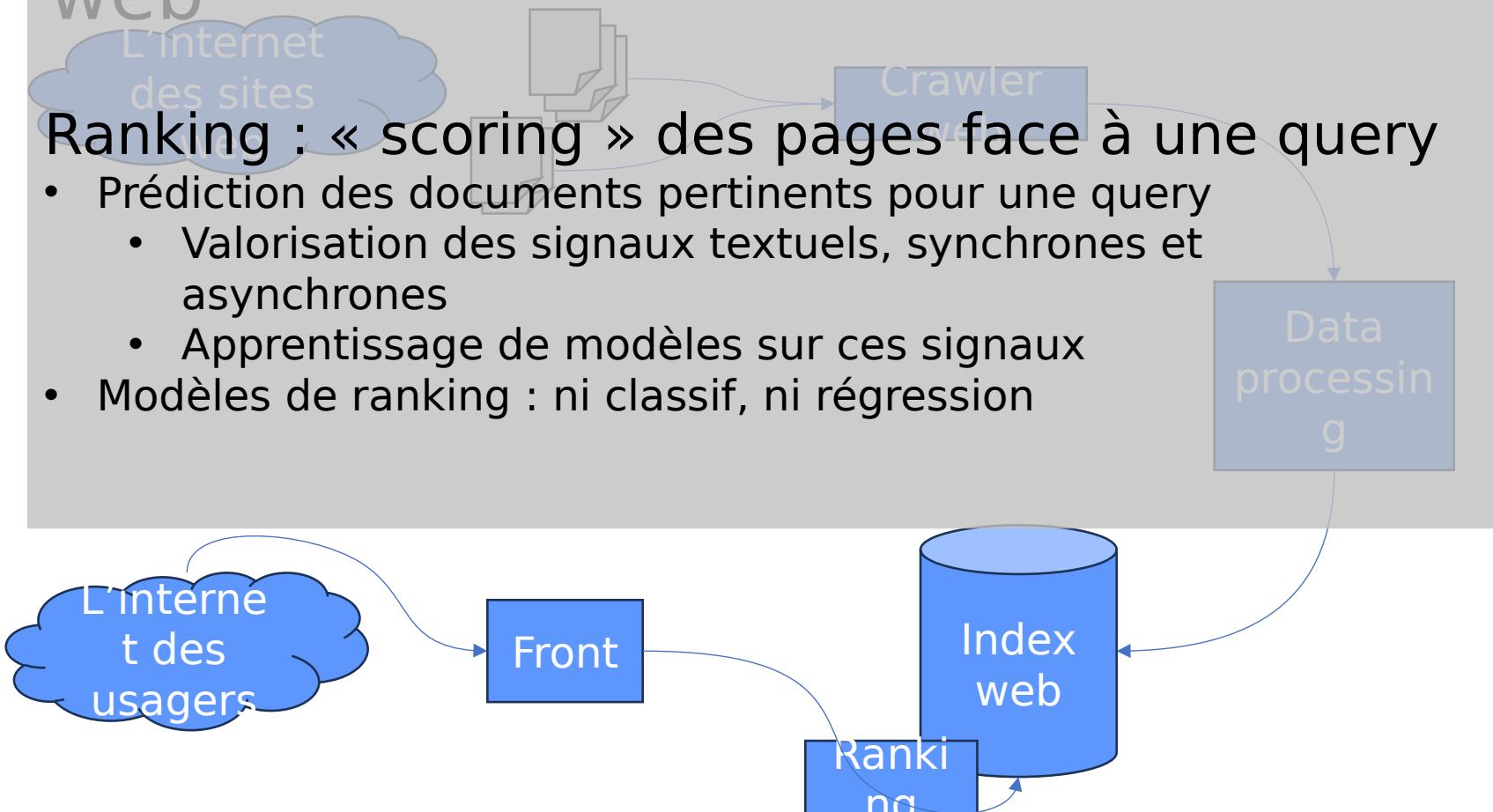


Architecture d'un moteur de recherche

web
L'internet
des sites

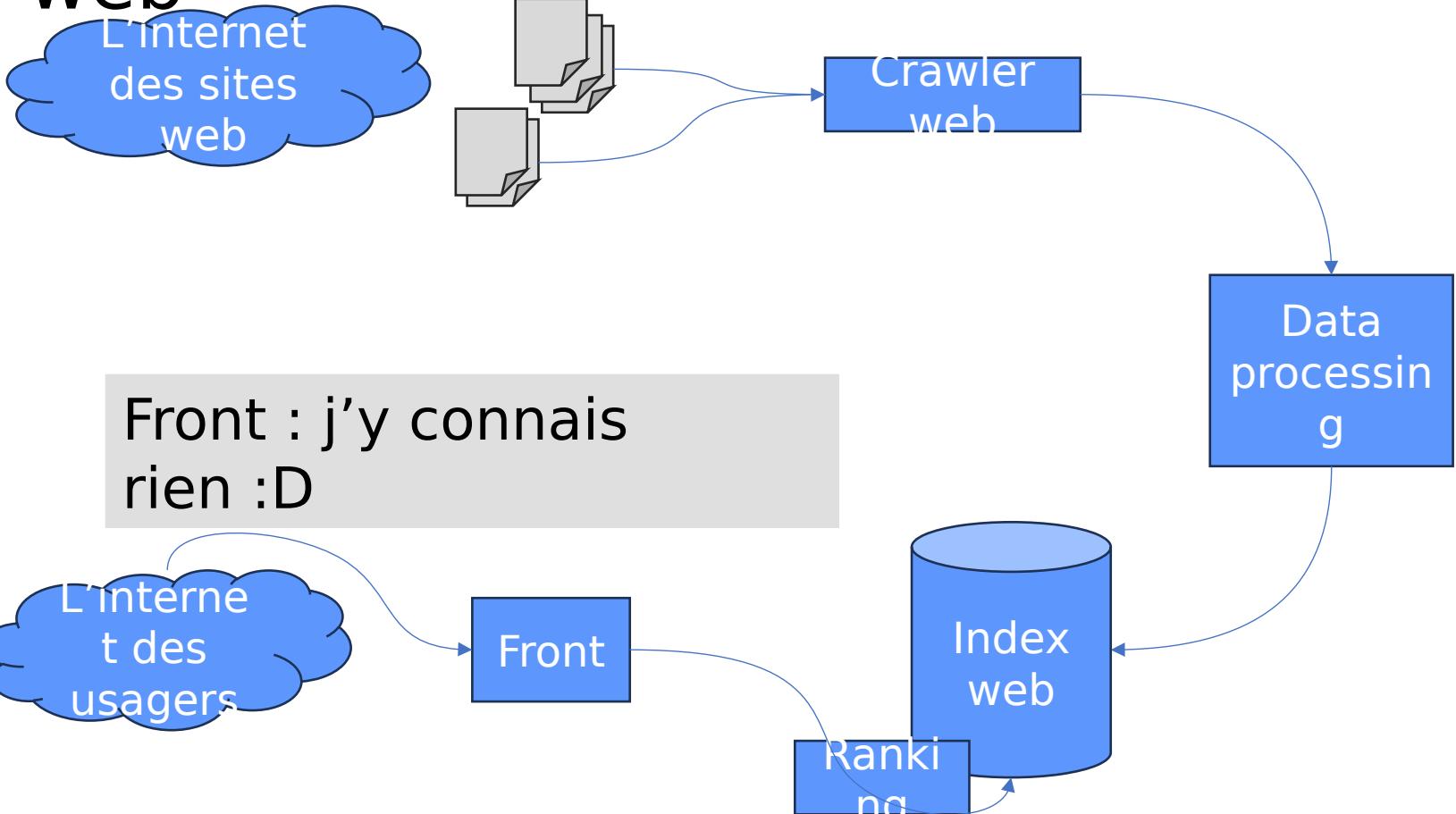
Ranking : « scoring » des pages face à une query

- Prédiction des documents pertinents pour une query
 - Valorisation des signaux textuels, synchrones et asynchrones
 - Apprentissage de modèles sur ces signaux
- Modèles de ranking : ni classif, ni régression





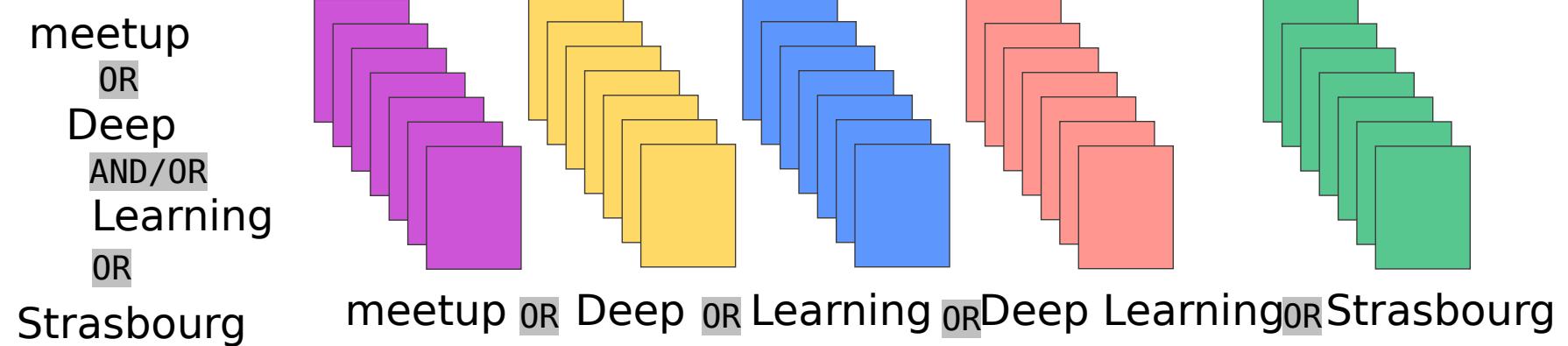
Architecture d'un moteur de recherche web



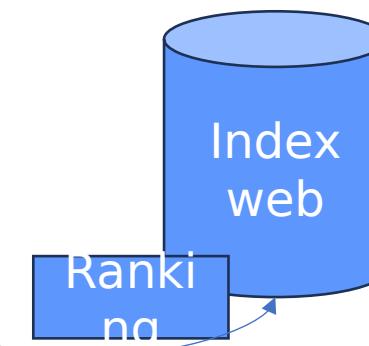


Architecture d'un moteur de recherche

Un index qui n'est pas une simple DB SQL



Query « **meetup DL Strasbourg** »



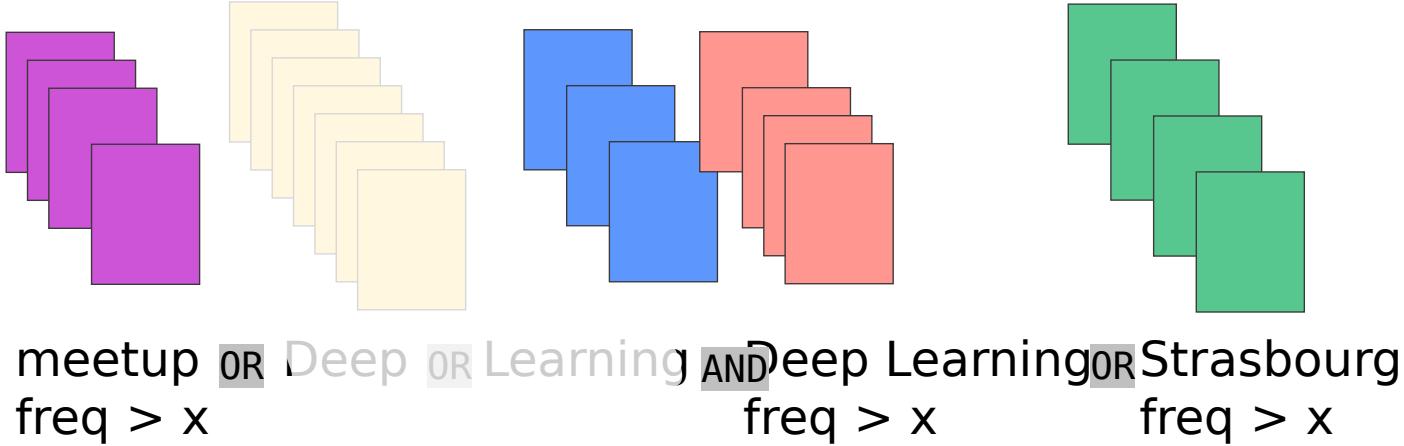


Architecture d'un moteur de recherche

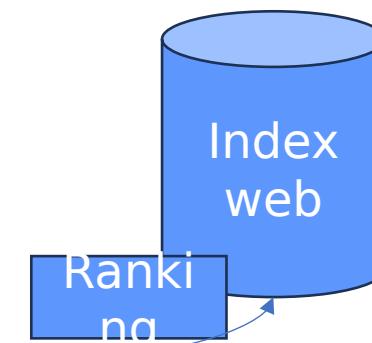
Un index web n'est pas une simple DB SQL

Trop de documents divers pour un simple SELECT : filtrage progressif !

meetup
OR
Deep
AND
Learning
OR
Strasbourg



Query « **meetup DL Strasbourg** »





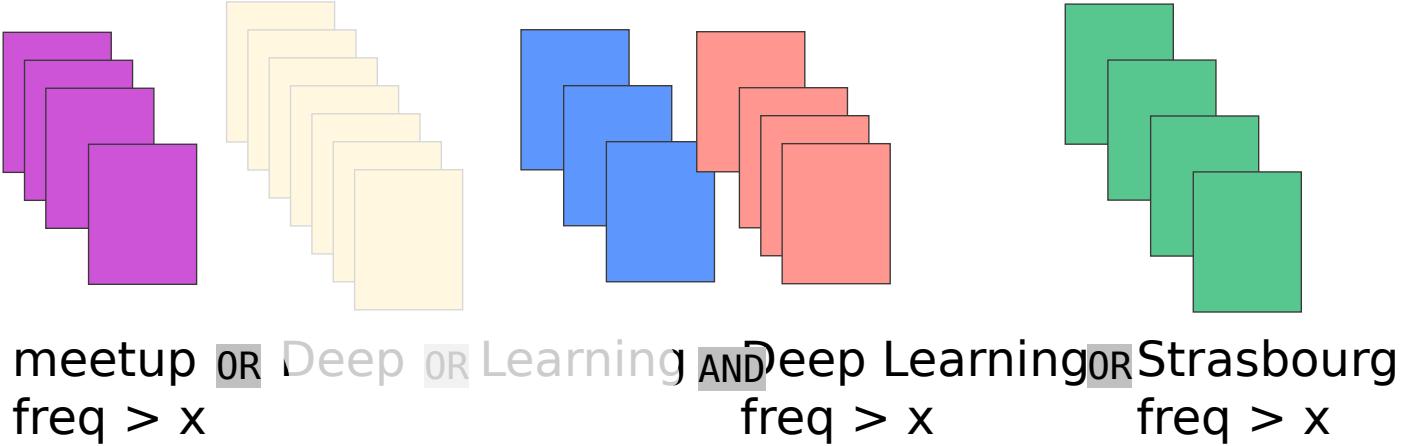
Architecture d'un moteur de recherche

Un index **web** n'est pas une simple DB SQL

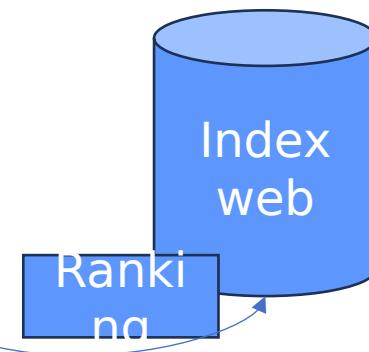
Trop de documents divers pour un simple SELECT : filtrage progressif !

meetup
OR
Deep
AND
Learning
OR
Strasbourg

freq > x



Query « **meetup DL Strasbourg** »



À ce stade, de l'ordre de 100k documents peuvent encore matcher
Comment ne rendre que le top20 à l'utilisateur ?



Architecture d'un moteur de recherche web

meetup
Deep
Learning
Strasbourg

Query « **meetup DL Strasbourg** »

Strasbourg est une commune française située dans la collectivité européenne d'Alsace dont elle est le chef-lieu. Elle est la préfecture du Bas-Rhin et de la région Grand Est.

Meetup.com est une plateforme de réseautage social créée par Scott Heiferman, Matt Meeker et Peter Kamali en 2002. **Meetup** permet aux membres de rencontrer des groupes unis par un intérêt commun ainsi que par zones géographiques.

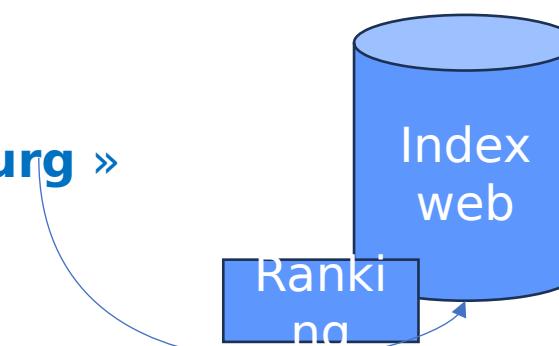
L'apprentissage automatique, (en anglais : machine **learning**, ML) est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données.

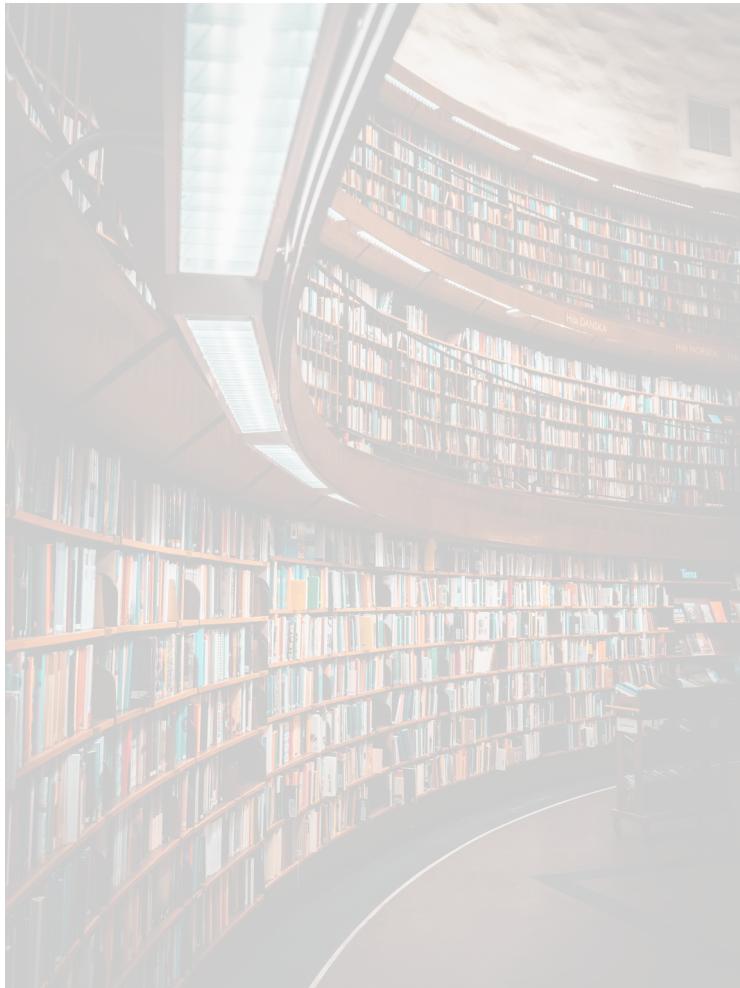
Strasbourg Deep Learning Meetup – If you are interested in **deep learning** and want to learn or share your knowledge, join our community. I hope that we will do this together and build a strong **deep learning** community in **Strasbourg**.

L'apprentissage profond (en anglais : **deep learning**) est un sous-domaine de l'IA qui utilise des réseaux neuronaux pour résoudre des tâches complexes grâce à des architectures articulées de différentes transformations non linéaires

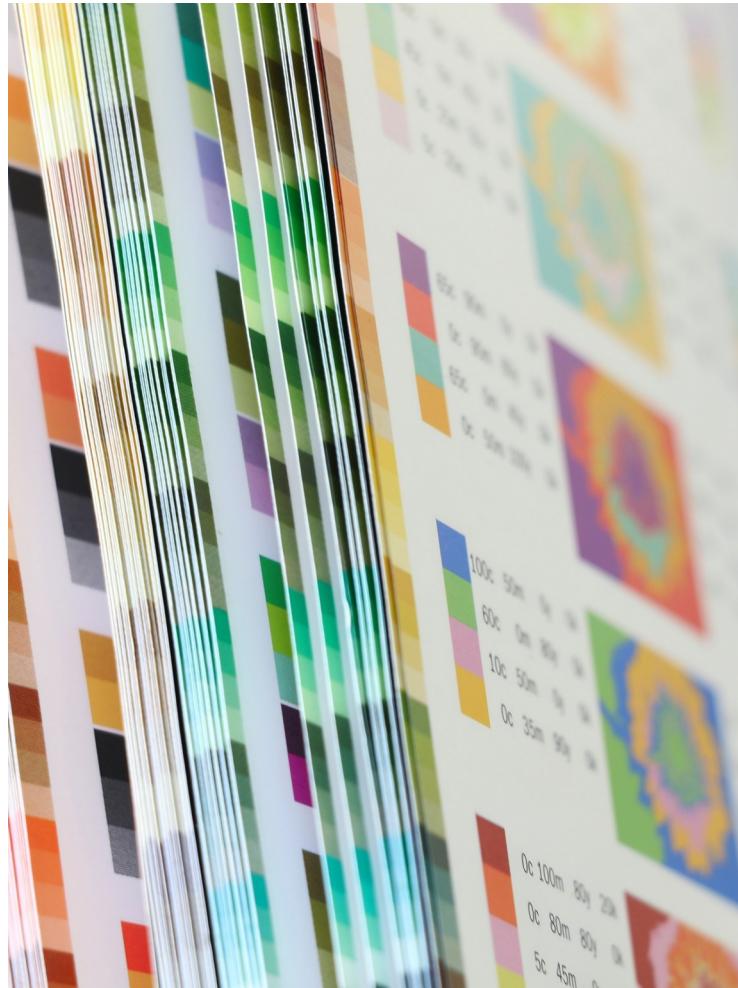
Académie de la Bière **Strasbourg** Cathédrale. Avis: They have a wide selection of regional beers and serve many of them in glassware specific to the label. Staffs are helpful. Fees are reasonable If you love beer and want to **learn** flam try it.

Unistra Université de **Strasbourg** – **Deep Learning** 1590€ 3j (21 heures en présentiel). Points forts de la formation : découverte des techniques au cœur de la révolution IA, utilisation de libraries et d'outils récents





Architecture d'un moteur de recherche web



Nécessité
d'ordonnancement
optimal : le learning to
rank



L'usage du NLP dans les moteurs modernes

Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Meetup Deep Learning Strasbourg

Contenu: Strasbourg Deep Learning Meetup

- If you are interested in deep learning and want to learn [...] and build a strong deep learning community in Strasbourg.



Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Recours à l'ancêtre du NLP : le **Term-Frequency (TF)**
Meetup Deep Learning Strasbourg

Content: Strasbourg Deep Learning Meetup

- If you are interested in deep learning and want to learn [...] and build a strong deep learning community in Strasbourg.

- . Déterminer un vocabulaire des mots intéressants $\mathcal{V} := (\nu_i)_{i \leq V}$
- . Représenter chaque doc par ses fréquences de mots du vocabulaire :

	chat	fromage	mange	mangée	souris
Le chat mange la souris	1/3	0	1/3	0	1/3
la souris est mangée par le chat	1/3	0	0	1/3	1/3
chat mange la souris qui mange le fromage	1/5	1/5	2/5	0	1/5



Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Recours à l'ancêtre du NLP : le **Term-Frequency (TF)** et
Meetup Deep Learning Strasbourg Content: Strasbourg Deep Learning Meetup
Inverse Doc Frequency

- If you are interested in deep learning and want to learn [...] and build a strong deep learning community in Strasbourg.

Déterminer un vocabulaire des mots intéressants $\mathcal{V} := (\nu_i)_{i \leq V}$

Représenter chaque doc par ses fréquences de mots du vocabulaire

Normalise par la log-fréquence d'apparition de chaque mot dans le corpus

	chat	fromage	mange	mangée	souris
Le chat mange la souris	0.5	0	0.6	0	0.5
la souris est mangée par le chat	2	7	2		
chat mange la souris qui mange le fromage	0.4 5	0 0	0 0.7	0.7 6	0.4 5



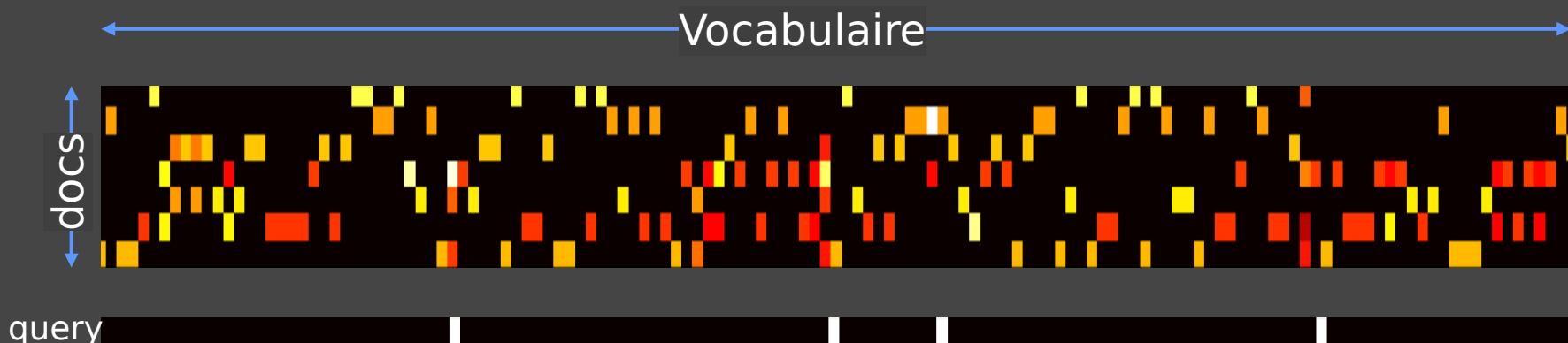
Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Recours à l'ancêtre du NLP : le **Term-Frequency (TF)** et
Inverse Doc Frequency

Content: Strasbourg Deep Learning Meetup

- If you are interested in deep learning and want to learn [...] and build a strong deep learning community in Strasbourg.



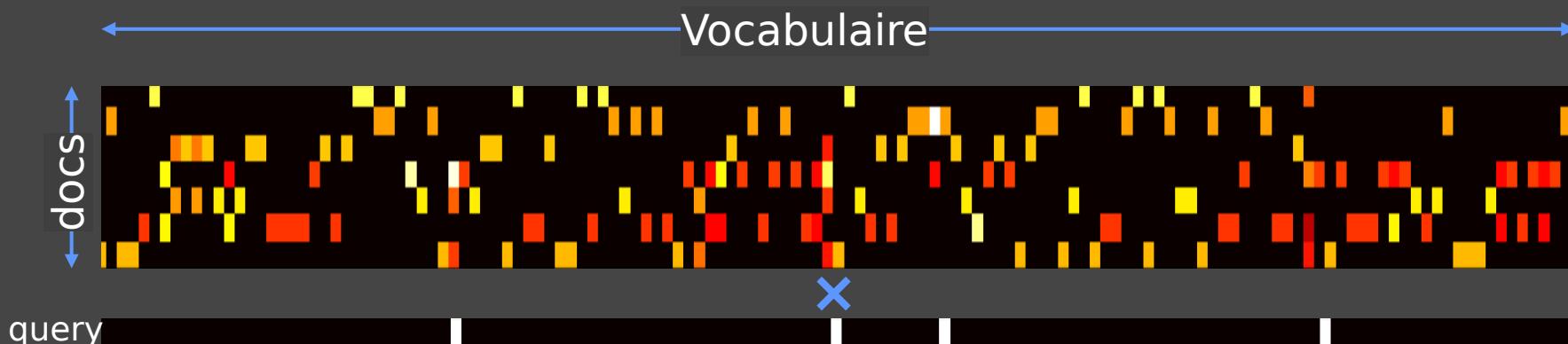
Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Recours à l'ancêtre du NLP : le **Term-Frequency (TF)** et
Inverse Doc Frequency

Content: Strasbourg Deep Learning Meetup
Meetup Deep Learning Strasbourg

- If you are interested in deep learning and want to learn [...] and build a strong deep learning community in Strasbourg.



équation = produit scalaire / cosine / autre ... => « meetup.com/strasbourg-DL »

Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Recours à beaucoup d'autres mesures de réussite
Meetup Deep Learning Strasbourg

Length: 4
Language: EN 75%, FR 24%
Topic: Community 40%, Science 50%, Travel 10%
...

Language-cosine: 0.9554
Topic-cosine: 0.9819
Title-TFIDF: 0.95
H1-TFIDF: 0.93
H2-TFIDF: 0.0
Content-TFIDF: 0.81
Title-BM25: 0.85
H1-BM25: 0.83
H2-BM25: 0.0
Content-BM25: 0.94
...

Title: Strasbourg Deep Learning Meetup |
Meetup
Header 1: Strasbourg Deep Learning
Meetup
Header 2: de quoi s'agit-il ?
Contenu: Strasbourg Deep Learning Meetup
- If you are interested in deep learning and
want to learn [...] and build a strong deep
learning community in Strasbourg.

Content-length: 40
Title-length: 5
Languages: EN 99%, FR 1%
Topic: Community 50%, Science 50%
PageRank: 3.14E-7
TrustRank: 2.71E-6
Spamness: 0.001
...



Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Comment exploiter ces quantifications ?
Meetup Deep Learning Strasbourg

Titre: Strasbourg Deep Learning Meetup | Meetup
Header 1: Strasbourg Deep Learning Meetup
Header 2: de quoi s'agit-il ?
Contenu: Strasbourg Deep Learning Meetup
- If you are interested in deep learning and want to learn [...] and build a strong deep learning community in Strasbourg.

DÉMO !



Nécessité d'ordonnancement optimal : le learning to rank

Comment quantifier l'adéquation entre une query et des documents ?

Métriques pur ranking

$$\text{NDCG}@k \quad nDCG_k = \frac{\sum_{i=1}^k \frac{2^{r(s_i)} - 1}{\log(1+i)}}{IDCG_k}$$

$$\text{Reciprocal Rank}@k \quad R_k = \frac{1}{\min\{i/r_i > 0, i < k\}}$$

Métriques type classification

Recall@k

$$R_k = \frac{\sum_{i=1}^k \mathbb{I}_{r_i > 0}}{\sum \mathbb{I}_{r_i > 0}}$$

Precision@k

$$P_k = \frac{\sum_{i=1}^k \mathbb{I}_{r_i > 0}}{k}$$



Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?



Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?

$$? \quad y \in [0; 4\ddot{\circ}]$$



Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?

Référence (relevance)



Prédiction (score / rang)



$NDCG=0.81$



Nécessité d'ordonnancement optimal : le learning to rank

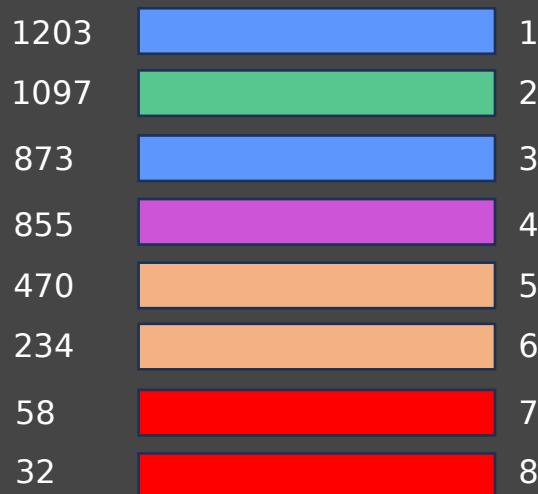
Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?

Référence (relevance)



Prédiction (score / rang)



$NDCG=0.81$



Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

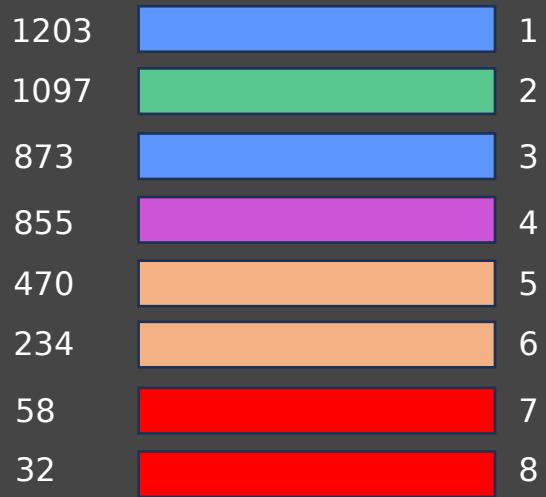
- En optimisant l'ordre ~~ou la distance à la relevance ?~~
- En contournant l'absence de « bon » gradient de loss ?

Optimiser l'ordre = optimiser le nDCG

Référence (relevance)



Prédiction (score / rang)



Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?

Optimiser l'ordre = optimiser le nDCG

Solution historique : rankNet [1], lambda-Rank [2] puis lambda-MART [3]



Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?

Optimiser l'ordre = optimiser le nDCG

Solution historique : rankNet [1], lambda-Rank [2] puis lambda-MART [3]

Fonction de score $f : X \mapsto s$

1 : optimiser un classifieur de paires bien/mal ordonnées [1]

$$\hat{f} \in \operatorname{argmin}_f \mathbb{E}_{i>j} \log(1+e^{s_i - s_j})$$

2 : hacker le gradient précédent en le multipliant par $\Delta NDC G_{i,j}$

$$\lambda_{i,j} = \frac{-1}{1+e^{s_i - s_j}} \vee \Delta NDC G_{i,j} \vee i$$

f(X1)	X1	1
f(X2)	X2	2
...	...	3
:	4	4
:	5	5
:	6	6
:	7	7
:	8	8

Chris, SHAKED, Tal, RENSHAW, Erin, et al. Learning to rank using gradient descent. In : Proceedings of the 22nd ICML. 2005. p. 79-86.

C., Ragno, R., & Le, Q. (2006). Learning to rank with nonsmooth cost functions. Advances in NeurIPS, 19.

Christopher JC. From ranknet to lambdarank to lambdamart: An overview. Learning, 2010, vol. 11, no 23-581, p. 81.

Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?

Solution historique : rankNet [1], lambda-Rank [2] puis lambda-MART [3]

- Solutions historiques à base de NN
- Peu de bonnes implémentations
 - PyTorch-LTR
 - Tensorflow-Ranking
- Solutions SOTA : boosting d'arbres
- LightGBM

$$\lambda_{i,j} = \frac{-1}{1+e^{s_i - s_j}} \vee \Delta NDC G_{i,j} \vee i$$



Nécessité d'ordonnancement optimal : le learning to rank

Comment exploiter ces quantifications ?

- En optimisant l'ordre ou la distance à la relevance ?
- En contournant l'absence de « bon » gradient de loss ?

Ce « gradient truqué » correspond en fait à une loss [4] ! Dans un contexte de relaxation continue des métriques LTR

La super bonne nouvelle est donc :

Vous pouvez désormais optimiser Prec, Recall, F1 ...!

□ Voir la lib JAX et son extension RAX

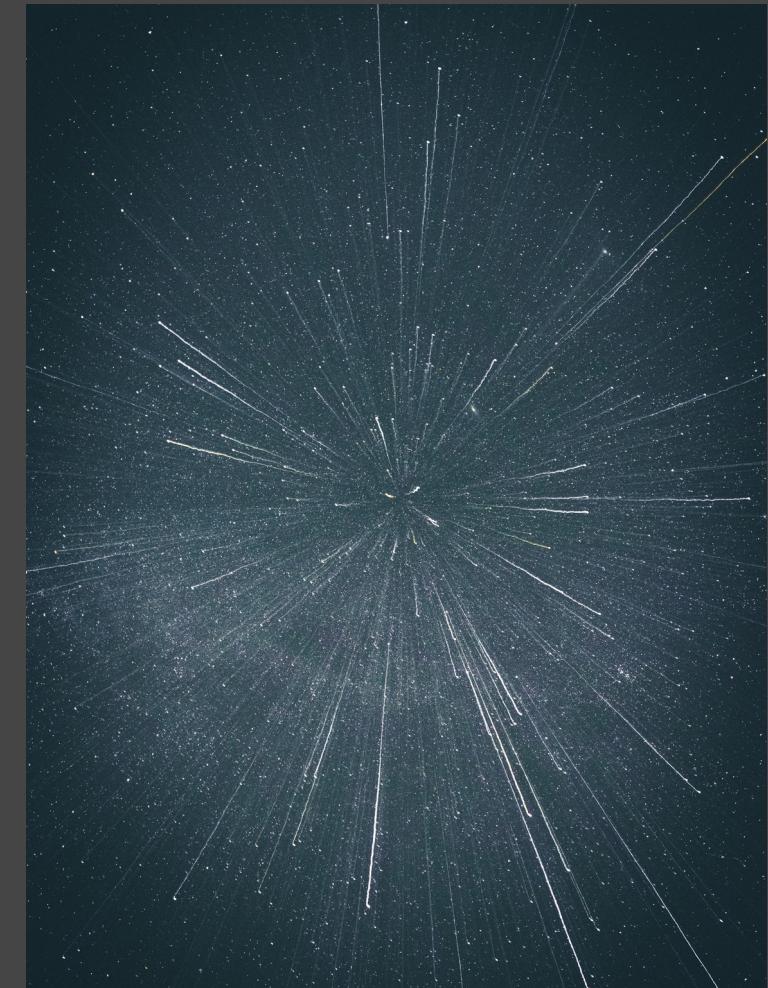
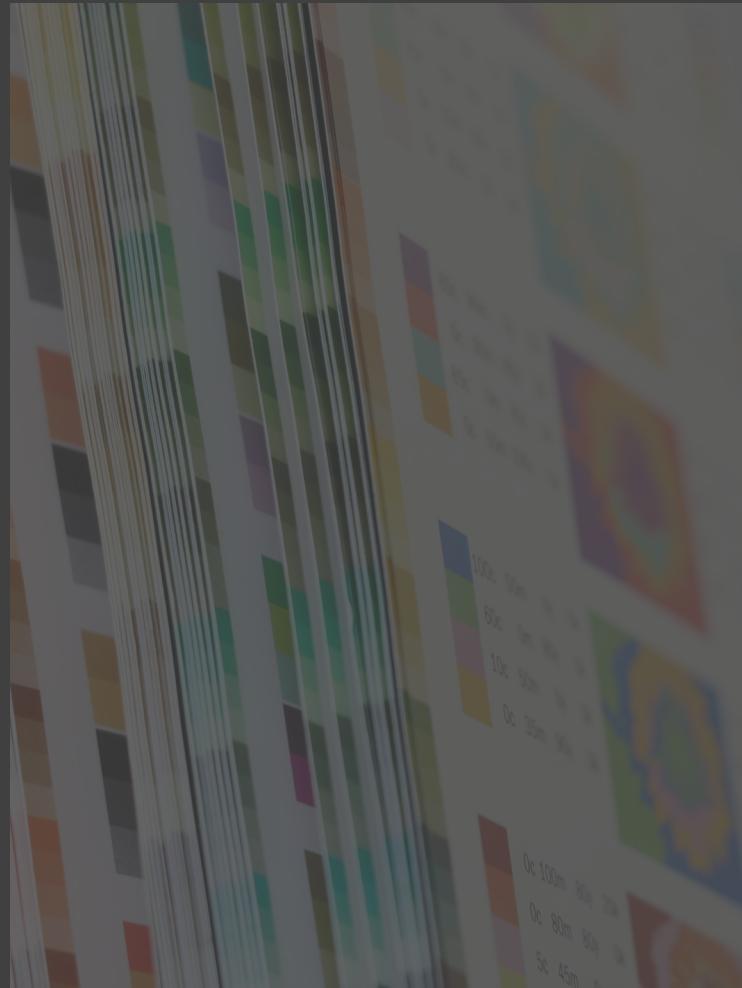
$$\lambda_{i,j} = \frac{-1}{1+e^{s_i-s_j}} \vee \Delta NDC G_{i,j} \vee i$$





Architecture d'un moteur de recherche web

Nécessité d'ordonnancement optimal : le learning to rank



L'usage du NLP dans les moteurs modernes



L'usage du NLP dans les moteurs modernes

NLP moderne :

- Embedding : compression des représentations sparses type TFIDF
- Traduction d'un document en un vecteur qui conserve la sémantique
- Traduction d'une query dans un même espace

Embeddings utilisés

- Pour faire de la recherche de plus proches voisins (lib FAISS)
- En complément des signaux historiques (hybrid search)

Remplacement du LTR par des embeddings toujours mieux taillés ?



L'usage du NLP dans les moteurs modernes

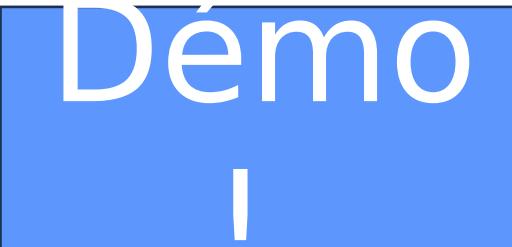
NLP moderne :

- Embedding : compression des représentations sparses type TFIDF
- Traduction d'un document en un vecteur qui conserve la sémantique
- Traduction d'une phrase en une autre dans un espace

Embeddings utilisés

- Pour faire de la recherche de plus proches voisins (lib FAISS)
- En complément des signaux historiques (hybrid search)

Remplacement du LTR par des embeddings toujours mieux taillés ?



Merci !

Vos questions ?