La recherche sémantique à l'ère de l'IA, mirage ou miracle ?

WEB DAYS CONVENTION

- Guillaume Loulier / @Guikingone
- Expert Technique @SensioLabs (détruit des CI's et des quotas Github) - *On recrute*
- Grand amoureux de PHP, Rust et d'automobile
- Publie une *veille* hebdomadaire parlant de cloud, machine learning, PHP et bien plus via Substack

01

Planifions

01 - Chercher n’est pas aisé, trouver est un problème

02 - L’IA, l’éléphant dans la bouteille

03 - *Semantiquement*, c’est compliqué

04 - Demain, tout changera (ou pas)

- Jusqu'à récemment, les *expériences* de recherche étaient basées sur les mots(-clefs)
- Ce type de recherche se calque sur la répétition, la **cohérence** et *un peu de chance*
- Quid des situations où les mots sont présents plusieurs fois ? Quid du contexte ?

- Le cerveau humain fonctionne par motifs, faits, expérimentations et habitudes, un ordinateur se *limite* aux binaires et mathématiques

- Un ordinateur *ne “sait” pas* chercher de façon *logique* / *sensée*

- Tout est question d’entropie, de temps et surtout, d’*énergie*

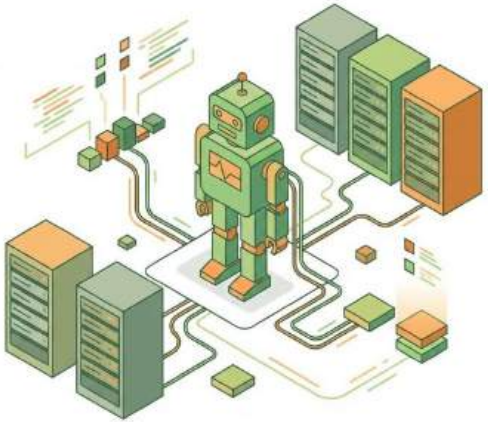
- La majorité des utilisateurs ne savent pas ce qu’ils cherchent / veulent
- La quantité de données filtrables est sans limites et croît continuellement
- Quid de l'essor de l’IA ?

30 Novembre 2022

- Première version de ChatGPT
- Aucun accès internet, *agent* conversationnel
- 1ère itération principalement pensée pour les recherches en anglais


Comment puis-je vous aider ?
Poser une question
Joindre
Rechercher
Étudier
Créer l'image
Voix

- L’IA révolutionne la recherche et le web et l’utilisation des données
- L’IA peut interagir avec le langage, la voix et/ou l’image (*multimodal*)
- L’IA peut (re)chercher sur le web comme sur vos données, quel que soit la langue



- Les révolutions récentes sont centrées autour des LLMs / transformers / MoE, des *sous-types* de “réseau neuronaux”

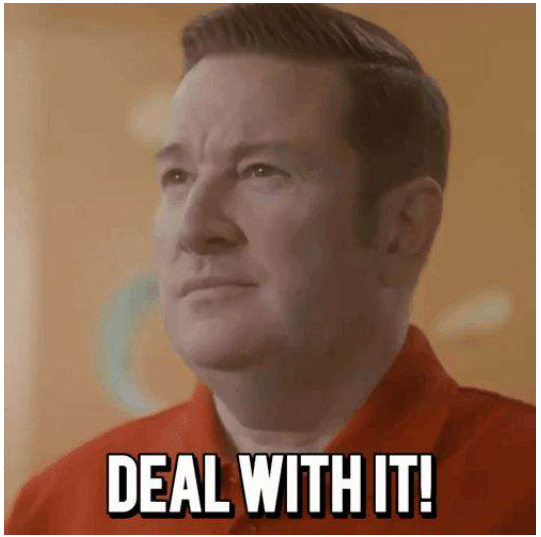
- Tout repose sur l’idée de prédire un jeton, un motif et “extrapoler”

- Les modèles sont tous **biaisés**

Quid du contexte ?! Quid de mes / nos / vos données ? Quid de la recherche sémantique ?

Houston, nous avons une tartine de problèmes

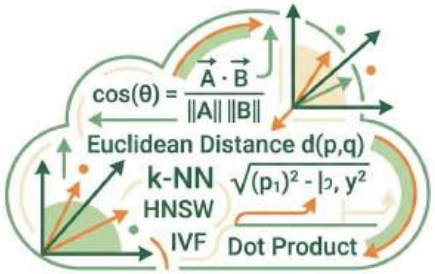
- De par sa structure probabiliste, l'IA pense en prédictions et non en contexte
- Les utilisateurs pensent plus vite qu'ils n' écrivent et/ou demandent *souvent* sans réfléchir
- Les recherches par mot-clef sont de facto, invalides



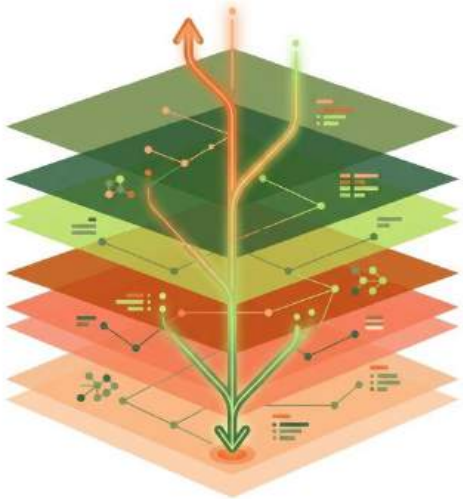
- La recherche sémantique (ou recherche *vectorielle*) est centrée autour de la proximité et de la similarité

- Plus les matrices sont proches, plus elles semblent *similaires*, plus la pertinence semble être *correcte*

- Quid de l'intention ? Quid du sens ?



- La “profondeur de recherche” n’a que *peu* d’impact sur les performances
- Plus les matrices sont denses et “contextualisées”, plus le résultat sera cohérent
- Envie de chercher dans une image ? Une vidéo ? *Vectorize the world* !



Similarité ne veut pas dire pertinence ou intention

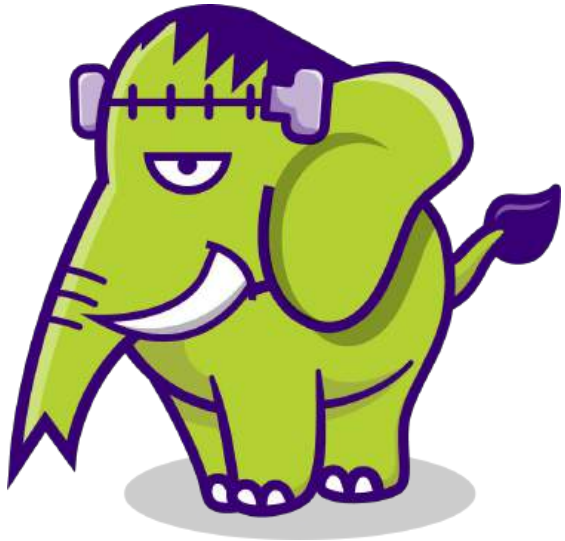
- **0.1** est proche de **0.2** mais aussi de **0**, que cache **0** ?
- La *similarité* ne veut pas dire que la signification est bonne
- Selon le contexte, la similarité peut introduire un **biais**
- Le contexte est la clef de voûte, tout le reste n'est que du **bruit**
- Plus le contexte est détaillé, plus les résultats seront cohérents

"Bien le bonsoir" est proche de "bonsoir" mais "bonjour" est plus pertinent selon l'intention

FrankenPHP est un serveur applicatif plus efficient que PHP-FPM

PHP-FPM est moins efficient que FrankenPHP

Quel serveur applicatif est le plus efficient ?



Quel	serveur	applicatif	est	le	plus	efficient	?
0	-1	0	1	-1	0	1	1
0	2	2	2	0	0	2	-1
3	0	0	2	3	3	0	1
1	1	2	0	1	-1	3	2

Encoder

[-0.88440161, -0.00996133, 0.243678553, …]

* Les valeurs sont donnés à titre d'exemple

[-0.88440161, -0.00996133, 0.243678553, …]

Recherche par similarité (cosine, angular, …)

[[-0.88440161, -0.00996133, 0.243678553, …], [-0.88440171, -0.00996144, 0.243678564, …], …]

* Les valeurs sont donnés à titre d'exemple

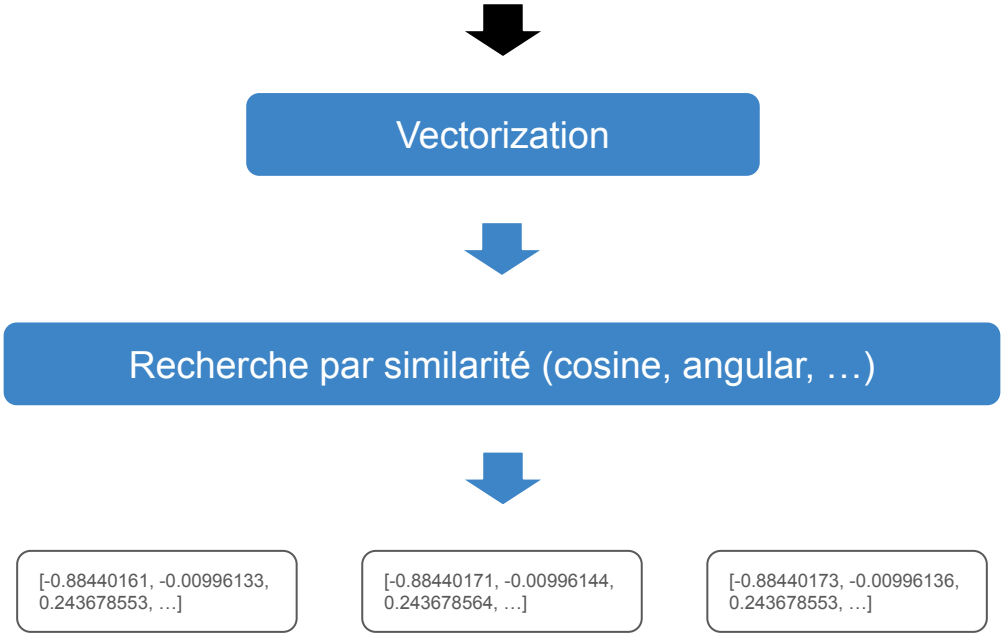
Oui mais moi, je “parle” à l’IA

- Une voix n’est au fond, qu’une fréquence, en somme, des mathématiques

- Une fois *vectorisée*, effectuer une recherche par cosine / autre est triviale

- Idem lors d’une recherche sur du texte, des images, vidéos, fichiers audio, etc

Quel serveur applicatif est le plus efficient ?

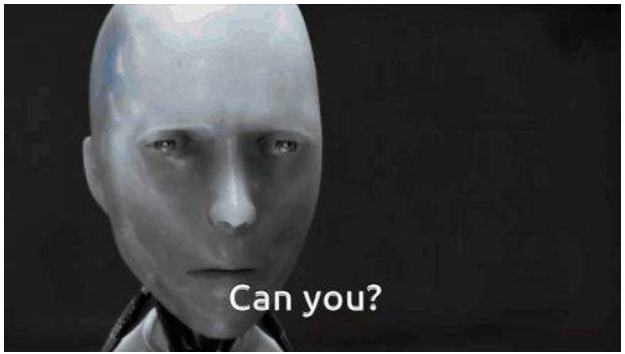


* Les valeurs sont donnés à titre d'exemple

Bravo, vous venez de réinventer la roue carrée

- L’IA ne change *rien* à la recherche, l’IA change l’interface utilisateur

- Notre but est de repenser l’interaction avec les données ainsi que l’expérience finale



- Le **contexte** reste la clef de voûte

- L’IA a *vocation* à nous augmenter, pas de nous remplacer, désolé Mr Altman

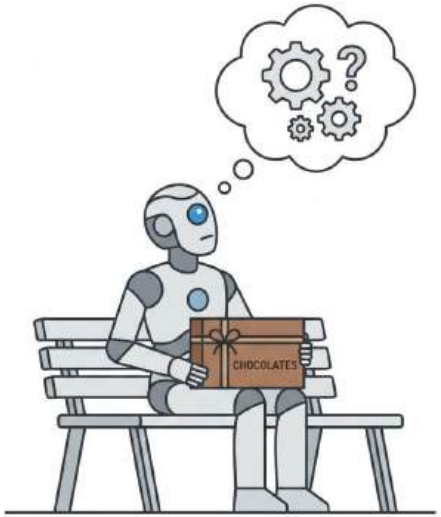
- Les recherches de demain se feront sans humains, *agentic-like*

- Le **contexte** reste / restera la clef de voûte

Comment qu’on fait alors ?

- Structurez vos données selon des besoins, un contexte
- Des résultats incohérents ? Changez de modèle, optimisez vos données et / ou vos vecteurs
- Testez, soyez indulgents et rigoureux

- Demain, vos utilisateurs seront des agents
- Un agent ne fait que retranscrire un besoin, attendez-vous à des erreurs / incohérences
- La recherche sémantique est comme une boîte de chocolat



Merci !