

Exercice Scientifique et technique

La mécanique de l'Information : Du texte brut à la recherche vectorielle.

Ainouche Rayan

Sommaire

1. Contexte
2. État de l'art
3. Préparer la donnée
4. Vectoriser & Interroger
5. Expérimentation & Résultats
6. Bilan

Contexte

Bull SAS — Service R&D Support

Dizaines de manuels PDF techniques (hardware & software), plusieurs milliers de pages.

Un technicien cherche une information pour résoudre un ticket.

La recherche classique cherche les mots exacts.

Le problème

Requête du technicien :

“moniteur éteint”

→ **0 résultat**

Dans les manuels :

“absence de signal vidéo”

Aucun mot en commun.

Question scientifique :

La recherche vectorielle permet-elle de retrouver de l'information sémantiquement proche
même sans terme commun ?

État de l'art - Deux familles d'approches

Étape	Approche classique	État de l'art retenu	Référence	Lien ESIEE
01 - Extraction	OCR / Regex → Perte de structure	Document Layout Analysis (Vision par ordinateur)	Pfitzmann et al. DocLayNet, IBM (2022)	Analyse syntaxique
02 - Représentation	TF-IDF / BM25 → Fréquentiel strict	Embeddings denses (Réseau siamois)	Reimers & Gurevych Sentence-BERT (2019)	Stats & Probabilités
03 - Recherche	K-NN exact / SQL → $O(N)$, lent	ANN via graphes HNSW (Approx. $O(\log N)$)	Johnson et al. FAISS, Meta AI (2017)	Algo & Graphes

Préparer la donnée - Extraction, tokenisation & découpage

Avant de vectoriser, le texte brut doit être découpé en unités sémantiques exploitables.

1. Extraction - Docling (IBM)

Récupère la structure du PDF : titres, tableaux, paragraphes.

Résultat : texte structuré, sans perte de hiérarchie.

2. Tokenisation & découpage

Régex → découpage aux ponctuations fortes (frontières lexicales).

Isolation des tableaux et listes.

Lien cours : Analyse Syntaxique - lexèmes, tokens, ANTLR.

3. Fusion sémantique

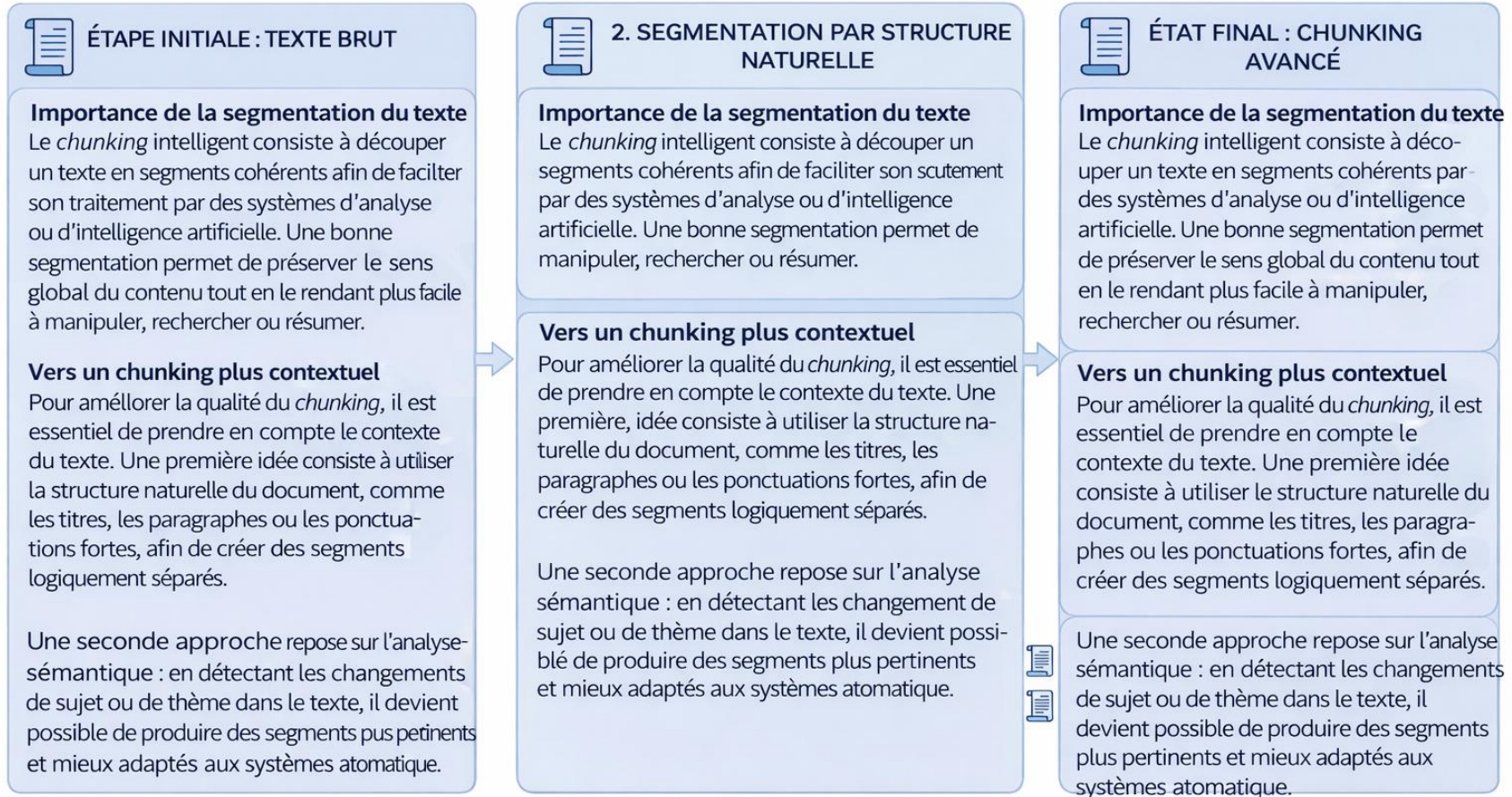
Calcul de similarité cosinus entre blocs adjacents.

Fusion si $\cos(\theta) > 0.90$ & leur taille reste faible.

Résultat : unités prêtes pour la vectorisation.

Préparer la donnée - Schéma du pipeline de traitement

Texte brut → Tokenisation (Regex / Lexer) → Chunks sémantiques



Vectoriser & Interroger - Du texte à l'espace mathématique

Chaque chunk devient un point dans un espace mathématique à 1024 dimensions.

Vectorisation — multilingual-e5-large

- Chunk → vecteur dans \mathbb{R}^{1024} (Wang et al., 2022)
- 100 % local — FR / EN
- Même sens → vecteurs géométriquement proches
- → $\cos(\theta)$ = corrélation normalisée
- Lien cours : Stats & Probabilités

Interroger — FAISS + HNSW

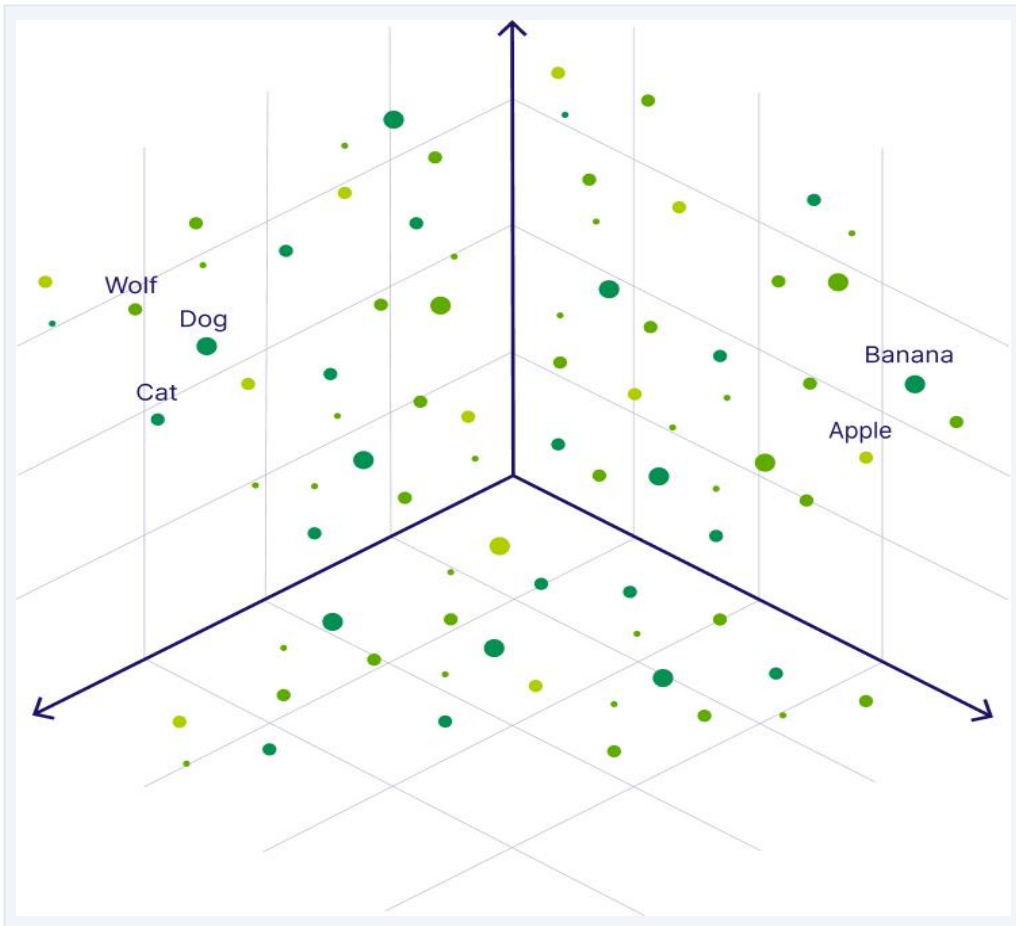
- Vectoriser la demande d'information (question technicien)
- Requête vectorisée → K=5 voisins proches
- FAISS HNSW : complexité $O(\log N)$
- Lien cours : Algorithmique & Graphes → HNSW

Mesure de similarité :

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

Vectoriser & Interroger - Visualisations

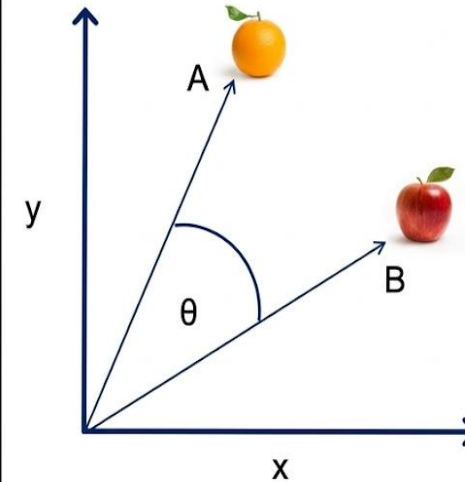
Représentation graphique de l'espace mathématique



Similarité Cosinus - distance angulaire

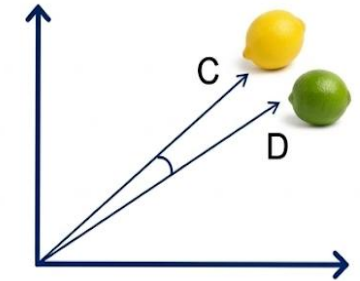
COSINE SIMILARITY EXAMPLES

Cosine Similarity

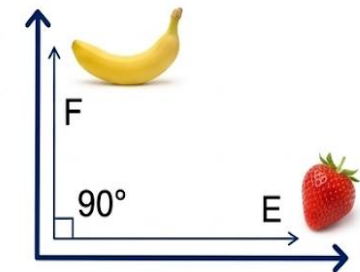


Comparison of Orange (A)
& Apple (B)
Intermediate Angle θ

High Similarity (Small Angle)



Low Similarity (Large Angle / Orthogonal)



Expérimentation - Protocole de mesure

L'expérience s'appuie sur 14 requêtes métiers : 7 directes (Type A) + 7 à vocabulaire divergent (Type B).

Document PDF

Extrait d'une RFC contenant 60 segments (environ 20 pages)

Requêtes

Type A - vocabulaire partagé entre requête et document.
Type B - aucun mot en commun.

Métriques

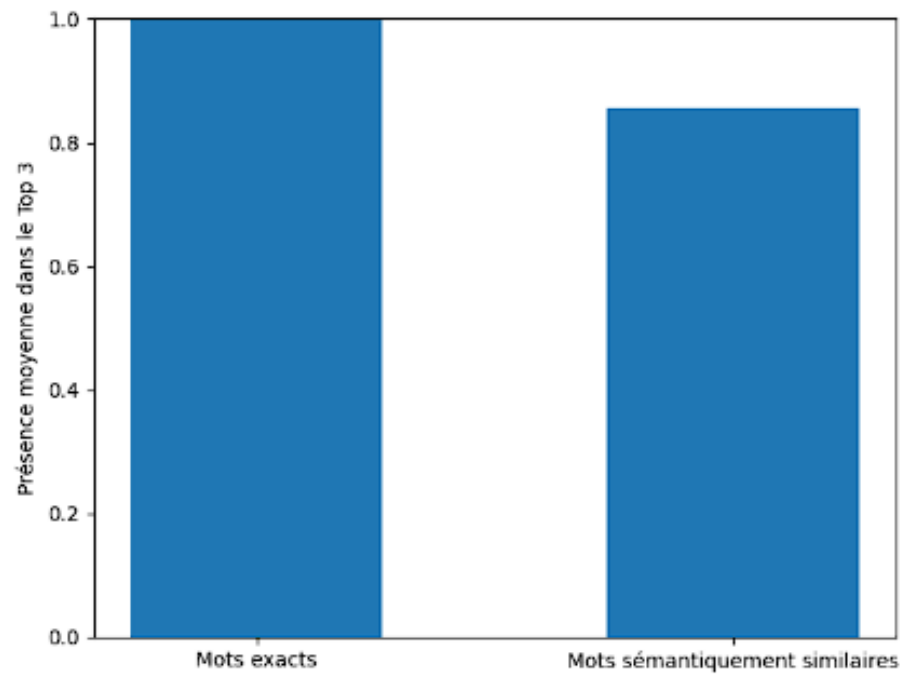
Précision@3 : parmi les 3 premiers, combien sont pertinents ?
Score cosinus moyen.
Évaluation manuelle.

Exemple de requête Type B :

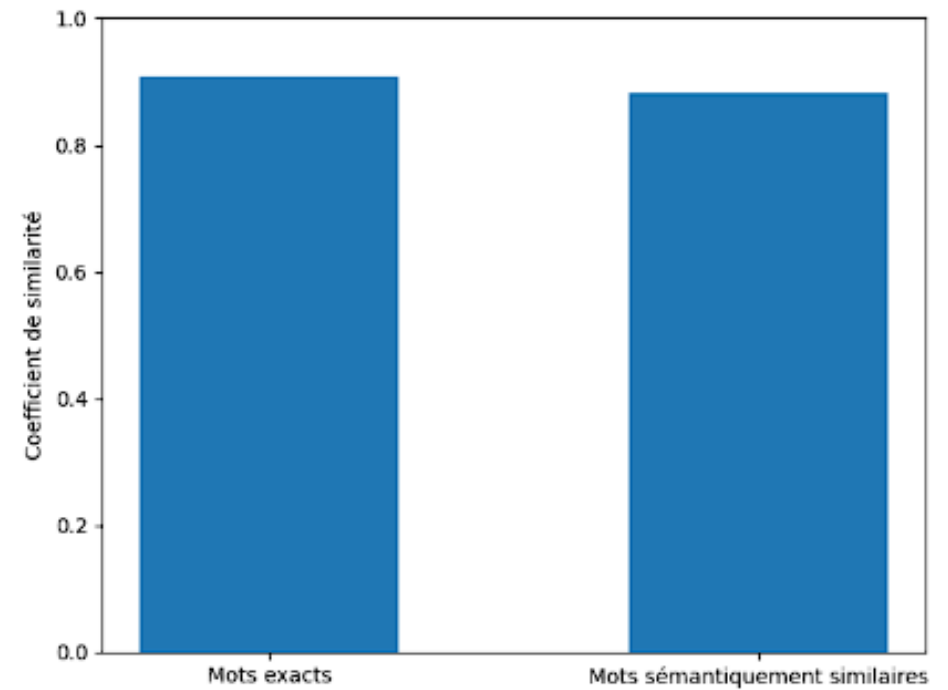
"moniteur qui ne s'allume pas" → les manuels parlent de : "absence de signal vidéo", "vérifier câble HDMI", "cold boot".

Expérimentation - Résultats

Précision@3 par type de requête



Score cosinus moyen



Bilan

La recherche vectorielle permet-elle de retrouver de l'information sémantiquement proche même sans terme commun ?

Oui - le score cosinus moyen reste de 0.88 sur les requêtes sans terme commun, Précision@3 = 0.85.

Ce que l'étude prouve

- La vectorisation encode le sens indépendamment du vocabulaire
- Un score cosinus élevé prédit la pertinence sémantique
- Le pipeline tokenisation → vecteur → FAISS est opérationnel en <100 ms

Limites

- Évaluation manuelle sur 14 requêtes, non généralisable
- Modèle e5-large généraliste et non fine-tuné sur un contexte précis (Documentation d'entreprise...)

Sources & Bibliographie

Papiers scientifiques

- Pfitzmann et al. — DocLayNet, IBM Research (2022)
- Reimers & Gurevych — Sentence-BERT (2019)
- Johnson et al. — FAISS, Meta AI (2017)
- Malkov & Yashunin — HNSW (2016)
- Wang et al. — Multilingual-E5-Large (2022)

Outils & Librairies

- Docling (IBM) — extraction & chunking structurel
- FAISS (Meta AI) — indexation vectorielle
- multilingual-e5-large — embeddings (HuggingFace)

Liens cours ESIEE

- Analyse syntaxique (Regex, Lexer, ANTLR) → Page 4
- Algorithmique & Graphes (HNSW, complexité) → Pages 7/8
- Stats & Probabilités (cosinus, covariance) → Page 7