



Présentation Stage Recherche Sémantique

Nathan Cerisara

03 Juin - 30 Août 2024

Plan Global

- I. Introduction du sujet
- II. Explication du fonctionnement / Architecture du projet
- III. Démon (WebApp puis Bot dans Sandbox Rainbow)
- IV. Conclusion
- V. Questions?
- VI. Annexes (NER, découpe de conversation)

I. Introduction du sujet

- Recherche Syntaxique dans Rainbow
- Limitations: Fautes de frappes, pas les bons mots
 - Ex:
 - “La réunion de 10h” → “La réu de 10h”
 - “J’ai du mal à travailler” → “J’arrive pas à bosser”
- Solution
 - -> Recherche sémantique

II. Fonctionnement / Architecture - Plan

- 1) Formalisation d'une instance Rainbow
- 2) Structure de moteurs / Exemple d'un moteur de recherche
- 3) Optimisations / Stockage de cache de données
- 4) Intégration à Rainbow / Bot de recherche
- 5) Mentions supplémentaires
 - a) Benchmarks
 - b) Optimisation des hyper-paramètres des configurations

1) Formalisation d'une instance Rainbow - Éléments de base

Bubble

- id
- name
- members_ids
- messages_ids

User

- id
- name
- bubbles_ids
- messages_ids

Message

- id
- content
- author_id
- author_name
- date
- bubble_id
- answered_msg_id

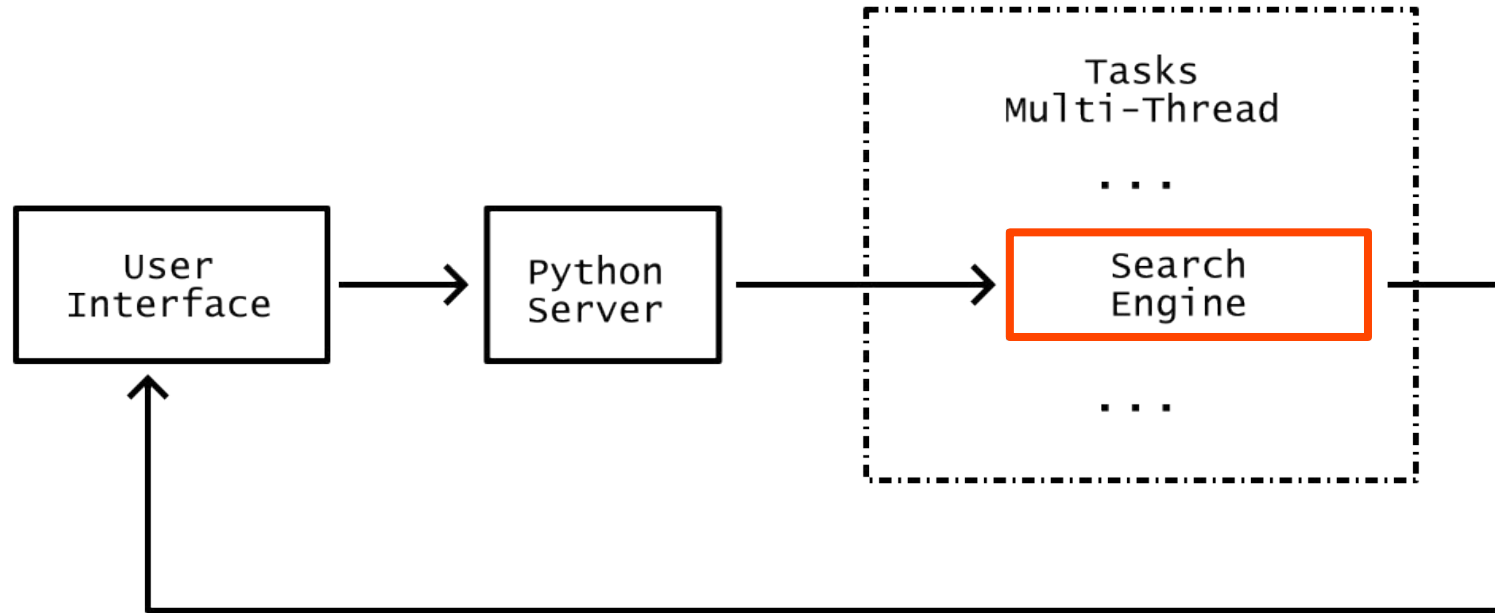
1) Formalisation d'une instance Rainbow - Définition RBI

Rainbow Instance(RBI)

- bubbles
- users
- messages

2) Structure de moteurs / Exemple d'un moteur de recherche

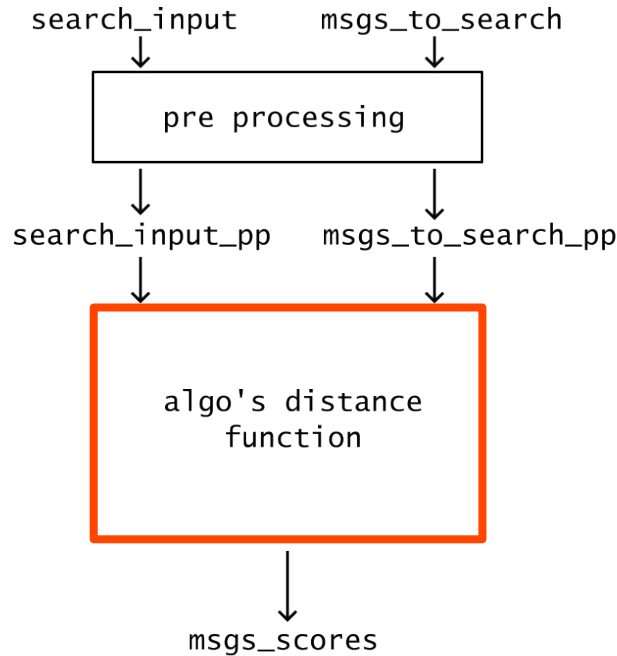
Architecture Globale Simplifiée



2) Structure de moteurs / Exemple d'un moteur de recherche

Structure d'un Algorithme de Recherche

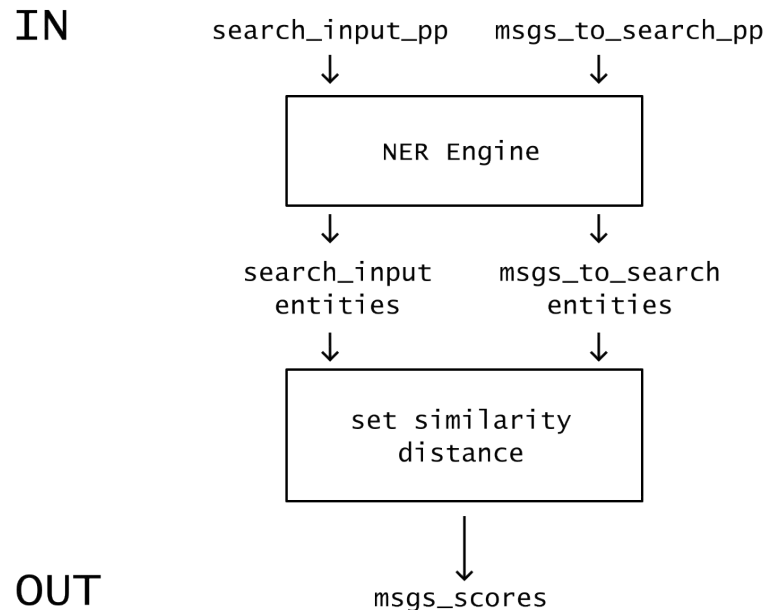
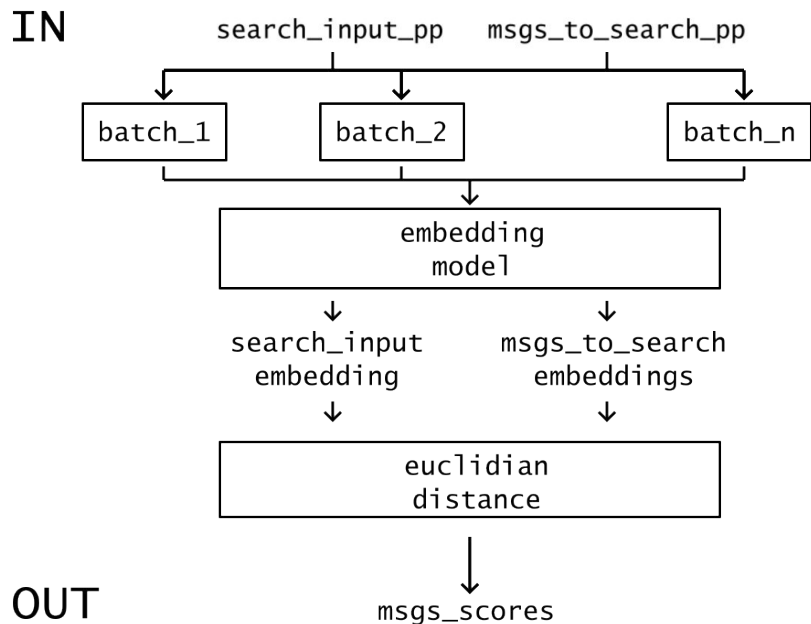
IN



OUT

2) Structure de moteurs / Exemple d'un moteur de recherche

Exemples: Algorithme d'embedding & Algorithme avec NER Engine



2) Structure de moteurs / Exemple d'un moteur de recherche

Détails sur le modèle d'embedding utilisé

Modèle d'embedding:

- Architecture Transformers
- AllMiniLM-l6-v2, Hugging Face, licence Apache 2.0
 - 22 millions de paramètres
 - Architecture de type Bert
 - Tourne en local sur ce Laptop
- Code Modulable -> Simple de changer de modèle si souhaité (config)

Plus de détails dans les questions pour ceux qui veulent.

3) Optimisations / Stockage de cache de données

- Compromis Vitesse - Stockage (Cache)
- Traduction (car modèles meilleurs en anglais)
- Stratégie actuelle: tout pré-calculer une seule fois à l'initialisation du serveur + mises à jours rapides à la réception de messages, tout côté serveur.
- Il faudra quand même bien étudier la mise à l'échelle dans l'architecture finale

4) Intégration à Rainbow / Bot de recherche

- Sandbox Rainbow
- Utilisation SDK C#
- Connexion socket entre script C# et serveur Python
- Limitations de la SDK
 - On n'a donc qu'un petit bot qui fonctionne par messages
 - Si intégré à Rainbow, ce sera directement dans l'interface de la recherche

5) Mentions supplémentaires

a) Benchmarks

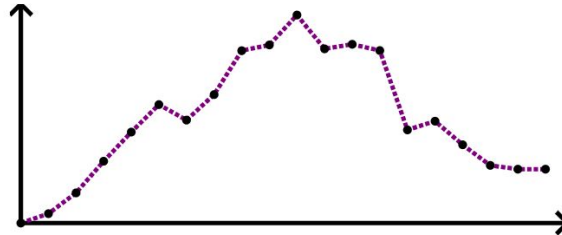
Machine Support : CPU: AMD Ryzen 5 PRO 4650U with Radeon Graphics - GPU: ▼

	Moyenne	Evaluation Sémantique Basique en Anglais		Evaluation Sémantique basique en Français		Evaluation Sémantique simple en Anglais		Evaluation Sémantique simple en Français		Test Sémantique en Français qui se concentre sur la NER		Test d'usage qui va tester des recherches dans une réelle bulle extraite de Rainbow	
- nom	v score	- score	- vitesse	- score	- vitesse	- score	- vitesse	- score	- vitesse	- score	- vitesse	- score	- vitesse
Embeddings all-MiniLM-L6-v2 with NER replacement and translation	0.8995	1	1.9837 sec	1	1.5499 sec	0.9167	0.3058 sec	0.9167	0.3649 sec	0.87	3.7941 sec	0.6938	138.7033 sec
Embeddings all-MiniLM-L6-v2 with NER Engine and translation	0.8862	1	1.0069 sec	1	1.0615 sec	0.9167	0.1974 sec	0.9167	0.2059 sec	0.7146	2.4419 sec	0.7694	120.4834 sec
Embeddings all-MiniLM-L6-v2 with NER replacement	0.7662	1	0.8477 sec	0.7456	1.0317 sec	0.9167	0.2141 sec	0.5387	0.202 sec	0.7751	2.5621 sec	0.6211	89.1231 sec
Embeddings all-MiniLM-L6-v2	0.7204	1	0.9578 sec	0.7456	1.1697 sec	0.9167	0.309 sec	0.5387	0.3342 sec	0.5003	4.0981 sec	0.6211	116.0886 sec
Embeddings all-MiniLM-L6-v2 with NER Engine	0.7053	1	1.0966 sec	0.7456	0.8938 sec	0.9167	0.2013 sec	0.5387	0.2044 sec	0.4677	2.3278 sec	0.5632	103.5798 sec
Simple Syntactic	0.3083	0.4852	0.003 sec	0.2767	0.002 sec	0.125	0 sec	0.0313	0 sec	0.5905	0.0081 sec	0.3408	0.5071 sec

5) Mentions supplémentaires

b) Optimisation des hyper-paramètres des configurations

avg conversation_test_1.json conversation_test_chatgpt_en_1.json conversation_test_chatgpt_en_2.json conversation_test_chatgpt_fr_1.json conversation_test_cl



Nom de la configuration : Clustering Séquentiel
Calcul en cours : 1 / 20 points
Score max : 0.7188320992230998 pour la valeur 1.4

Exporter la courbe

Revenir en arrière

III. Démo - WebApp



on peut faire une prime sur la performance

Marc Leclerc

Résultats pour la recherche `on peut faire une prime sur la performance` avec l'utilisateur `Marc Leclerc` :

1.0131885051727294 - managers_2 - Laura Dupuis - 2023/00/16-09:33 | msg id : 161
Parfait, et que pensez-vous d'une prime de performance trimestrielle pour encourager le dépassement des objectifs ?

1.221188163757324 - managers_2 - Olivier Morel - 2023/00/16-09:35 | msg id : 162
Les primes de performance sont intéressantes mais attention aux KPIs choisis. Ils doivent être SMART (Spécifiques, Mesurables, Atteignables, Réalistes, Temporellement définis).

1.2592915534973144 - managers_2 - Olivier Morel - 2023/00/23-09:26 | msg id : 189
Ça me semble raisonnable. Sophie, sur les feedbacks anonymes des primes de performance, as-tu des solutions pour garantir l'objectivité ?

1.3281070232391357 - managers_2 - Marc Leclerc - 2023/00/25-10:05 | msg id : 204
Bonjour Laura, c'est une bonne nouvelle. Concernant les primes de performance, quelles sont les prochaines étapes ?

1.3618499279022216 - managers_2 - Laura Dupuis - 2023/00/19-10:05 | msg id : 173
Oui, Marc. La direction est favorable aux primes de performance et à la flexibilité horaire. En revanche, ils veulent approfondir les aspects du télétravail et des initiatives RSE.

1.3768389701843262 - managers_2 - Laura Dupuis - 2023/00/30-09:23 | msg id : 233
Nous allons surveiller l'atteinte des objectifs individuels, les feedbacks d'équipe, l'engagement sur les outils de visio et les niveaux de satisfaction via des

III. Démo - Bot dans Sandbox Rainbow

17

The screenshot displays a chat application interface. On the left, a sidebar contains a list of recent conversations: 'Bubble Test 2' (Search result for "How can I fix t..."), 'Bubble Test 1' (Bot is connected !!!), and 'Bot Search' (Bot is connected !!!). The main chat area is titled 'Bubble Test 2' and shows a search query: '/search Are you free for a business lunch tomorrow?'. Below the query, the search results are displayed as a list of messages from Nathan Cerisara to Bubble Test 2. The messages include a confirmation of a meeting at 12h30, a confirmation of a meeting at 10h, a modification of the meeting time to 15h, and a reminder to join the meeting at 14h. The interface also includes a search bar at the top, a list of recent conversations on the left, and a chat input area at the bottom.

People, bubbles...

RECENT CONVERSATIONS

- Bubble Test 2
Search result for "How can I fix t..."
- Bubble Test 1
Bot is connected !!!
- Bot Search
Bot is connected !!!

Bubble Test 2

/search Are you free for a business lunch tomorrow?

August 14th 4:20 PM

August 14th 4:44 PM

Bot Search

Me

/search Are you free for a business lunch tomorrow?

Search result for "Are you free for a business lunch tomorrow?" (3,4842467308044434 sec):

- (0) msg id : @web_060ce78e-4afa-4a3b-9490-38c4bb7d6b7036 (0,7662493944168091)
> 'Seriez-vous disponible pour un déjeuner d'affaires demain à 12h30 ?' (from: Nathan Cerisara - in: Bubble Test 2)
- (1) msg id : @web_060ce78e-4afa-4a3b-9490-38c4bb7d6b7058 (1,309795904159546)
> 'Je confirme notre rendez-vous de demain à 10h au siège social.' (from: Nathan Cerisara - in: Bubble Test 2)
- (2) msg id : @web_060ce78e-4afa-4a3b-9490-38c4bb7d6b7038 (1,3233459949493407)
> 'Voici le compte-rendu de la réunion de ce matin. N'hésitez pas à ajouter vos commentaires.' (from: Nathan Cerisara - in: Bubble Test 2)
- (3) msg id : @web_060ce78e-4afa-4a3b-9490-38c4bb7d6b7062 (1,3235830307006835)
> 'L'heure de la réunion de demain a été modifiée à 15h. Merci de mettre à jour votre agenda.' (from: Nathan Cerisara - in: Bubble Test 2)
- (4) msg id : @web_060ce78e-4afa-4a3b-9490-38c4bb7d6b7048 (1,3417017459869385)
> 'Petite pause-café à 15h dans la salle de repos pour discuter des nouvelles idées. Rejoignez-nous !' (from: Nathan Cerisara - in: Bubble Test 2)
- (5) msg id : @web_060ce78e-4afa-4a3b-9490-38c4bb7d6b7022 (1,4283482551574707)
> 'N'oubliez pas la réunion d'équipe aujourd'hui à 14h. Merci d'être ponctuels.' (from: Nathan Cerisara - in: Bubble Test 2)

Enter your text here...

IV. Conclusion

- Meilleure Recherche
- Architecture modulable -> flexibilité et améliorations
- Avec meilleurs modèle d'embeddings pour le Français
 - pas besoin de traduction, donc plus de problèmes de cache de traduction
- Il reste quand mêmes quelques limites sur “l’intelligence” du modèle.
- Tout tourne en local sur le laptop Thinkpad
 - + rapide avec meilleurs CPU / GPU
- Il faudra quand même bien étudier la mise à l'échelle, pour voir quelle puissance serveur sera requise, ou bien déléguer certaines tâches côté client?

Questions ?

VI. Annexes

Annexe 1 - Reconnaissance d'Entités Nommées (NER)

Dictionnaires d'entités nommées

Annexe 2- Découpe de conversation

Annexe 1 - Reconnaissance d'Entités Nommées (NER)

Dictionnaires d'entités nommées

21

```
{  
  "EDT": "emplois du temps",  
  "JIRA": "système de suivi de bugs, de gestion des incidents et de gestion de projets",  
  "LinkedIn Learning": "online learning platform that provides video courses taught by industry experts",  
  "Coursera": "entreprise numérique proposant des formations en ligne ouvertes à tous",  
  "KPI": "indicateur clé de performance",  
  "KPIs": "indicateurs clés de performance",  
  "RSE": "responsabilité sociétale des entreprises",  
  "GAN": "réseau adversaire génératif",  
  "GANs": "réseaux adversaires génératifs",  
  "SurveyMonkey": "site de sondage en ligne gratuit avec sondages personnalisables",  
  "ViT": "Vision Transformer",  
  "DETR": "Detection Transformer, Detectron"  
}
```

Annexe 2 - Découpe de conversation

Algorithme séquentiel:

- Calcul d'une matrice de distances entre tous les messages
- On lit les messages dans l'ordre
- Pour chaque message:
 - Si conversation trouvée qui convient
-> On l'y ajoute
 - Sinon, nouvelle conversation