**M1 Informatique –UE Projet**

**Carnet de bord : les coulisses de la recherche documentaire**

*Les éléments que vous indiquez dans ce carnet donneront lieu à une notation*

**Noms, prénoms et spécialité :**

|  |
| --- |
| **Amine Djeghri** |
| **Idles Mamou** |
| **Master 1 DAC** |
|  |

**Sujet :**

|  |
| --- |
| **Sparse Embeddings pour la Recherche d’Information** |

**Consigne :**

1. **Introduction (5- 10 lignes max) :** Décrivez rapidement votre sujet de recherche, ses différents aspects et enjeux, ainsi que l’angle sous lequel vous avez décidé de le traiter.
2. **Les mots clés retenus (5- 10 lignes max) :** Listez les mots clés que vous avez utilisés pour votre recherche bibliographique. Organisez-les sous forme de carte heuristique.
3. **Descriptif de la recherche documentaire (10-15 lignes) :** Décrivez votre utilisation des différents outils de recherche (moteurs de recherche, base de donnée, catalogues, recherche par rebond etc.) et comparez les outils entre eux ? A quelles sources vous ont-ils permis d’accéder ? Quelles sont leurs spécificités ? Leur niveau de spécialisation ?
4. **Bibliographie produite dans le cadre du projet :** Utilisez la norme ACM ou IEEE.
5. **Evaluation des sources (5 lignes minimum par sources)** : Choisissez 3 sources parmi votre bibliographie, décrivez la manière dont vous les avez trouvées et faites-en une évaluation critique en utilisant les critères vus en TD.

Votre carnet de bord doit être remis en mains propres au formateur LE JOUR DU TUTORAT. Une copie numérique devra être envoyée à l’adresse suivante : [**Adrien.Demilly@scd.upmc.fr**](mailto:Adrien.Demilly@scd.upmc.fr)

Rappel : les supports de TD sont disponibles à l’adresse suivante:

[**http://www.pearltrees.com/formationbsu/master-info/id23514400**](http://www.pearltrees.com/formationbsu/master-info/id23514400)

**Introduction :**

La recherche d’information est le domaine qui étudie la manière de retrouver des informations dans un corpus. Notre projet consiste à implémenter un modèle de ranking de recherche d’information, qui consiste à retrouver les documents tout en les triant pour l’utilisateur dans un ordre du plus pertinent au moins pertinent sur ce qu’il recherche.

Dans notre projet, nous aborderons les modèles de recherche d’information à base de réseaux de neurones, ainsi que des modèles de représentation de textes. Une des manières de représenter du texte dans une machine dans le domaine du Traitement Automatique de la Langue, c’est, d’utiliser une représentation continue, appelée « embedding »

**Les mots clés :**

Information Retrieval, ad-hoc retrieval, Information Filtering, boolean model, Vector space model, Latent semantic Indexing, Probabilistic model, Language model, Inference network model, BM25, Learning to rank, neural ranking models, Pointwise approach, Pairwise approach, Listwise approach, NTLM, DESM, DSSM, CLSM, Architecture I,

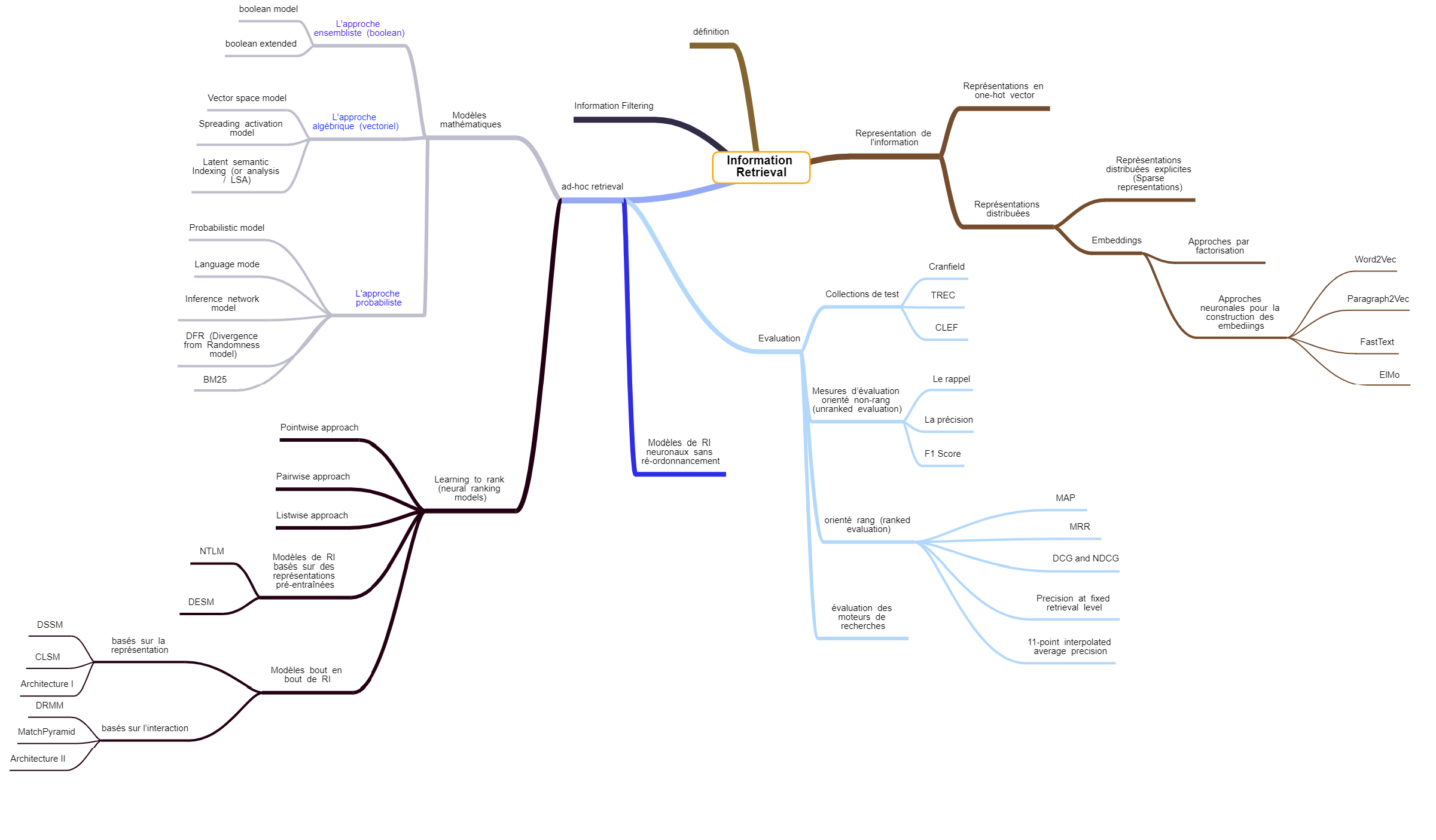
DRMM, MatchPyramid, Architecture II, one-hot vector, Sparse representations, Embeddings, Word2Vec, Paragraph2Vec, FastText, ElMo, Evaluation metrics information retrieval, Collections de test, unranked evaluation, ranked evaluation

(Généralement, nous combinons les mots clés cites au-dessus, avec le mot clé ‘information retrieval’ par exemple :

‘Evaluation information retrieval’ afin d’obtenir les bons résultats qui tournent autour du sujet notre projet) .

**Carte heuristique :**

Pour dessiner la carte heuristique, nous avons utilisé Framindmap.org



**Descriptif de la recherche documentaire**

**Descriptif de la recherche documentaire**

En partant des articles cités dans les références du sujet de notre projet et des sources primaires, nous avons déterminé quelques mots clés qui nous aideront à débuter notre recherche documentaire.

D’abord, après avoir recherché les mots clés sur Google, ce dernier nous affiche généralement en premier lieu wikipédia, que nous utilisons uniquement pour cadrer le sujet et avoir le bon vocabulaire. Après avoir construit le début de la carte heuristique, cette dernière va nous permettre de se concentrer sur les branches une par une et de s’approfondir grâce à Arxiv, Google Scholar et Web of science pour retrouver les articles en lien avec nos recherches.

**Google Scholar** : contient des millions de résultats et propose peu de filtres, par conséquent, nous l’avons utilisé que pour une recherche précise d’un article.

**Portail documentaire Sorbonne / web of science :** Pour la recherche de livres, d’articles et de reviews.

**Arxiv** : Principalement pour la recherche d’articles, et spécialement dans le domaine de l’informatique, cependant, il contient beaucoup de pré-publications, et d’articles qui n’ont pas été publiés.

**Bibliographie produite dans le cadre du projet**

Et enfin, pour ce qui est des sources, grâce à Zotero et son extension sur le navigateur, ce logiciel nous a permis d’insérer et de citer facilement les références et les bibliographies des différents articles que nous avons utilisés pour la rédaction de notre rapport.

Ci-dessous, quelques sources parmi les 60 sources que nous avons citées

[17] Richard Socher JEFFREY PENNINGTON et Christopher D MANNING. « Glove: Global vectors for word representation ». In : (2014).

[18] Qingyao Ai JIAFENG GUO Yixing Fan et W. Bruce CROFT. « A Deep Relevance Mat-ching Model for Ad-hoc Retrieval ». In : (2017).

[19] J. Guo L. PANGY. Lan et J. XU. « Text matching as image recognition ». In : (2016).

[20] J. LAFFER TY et C. ZHAI. « Document language models, query models, and risk minimization for information retrieval ». In :The 24th annual international ACM SIGIR conference on Research and development in information retrieval(2001), p. 111-119.

[21] Quoc V LE et Tomas MIKOLOV. « Distributed Representations of Sentences and Documents ». In : (2014).

**Evaluation des sources**