# 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**

Rappel : les supports de TD sont disponibles à l'adresse suivante:
**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 »

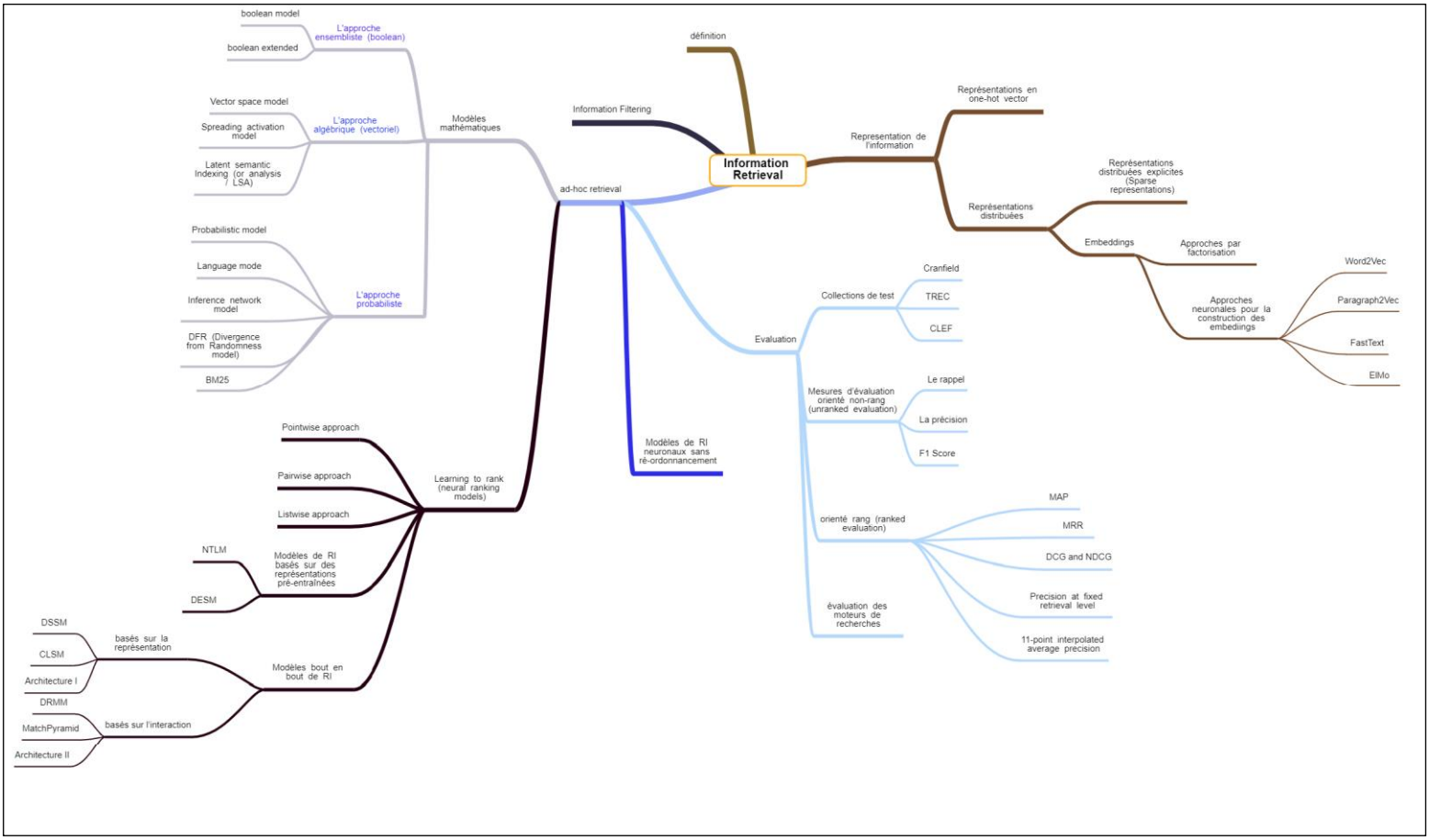
## Les mots clés :

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

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, ainsi que le fichier en extension « .bib » généré par Zoteron norme : IEEE


[2]Z. Lu et H. Li, « A Deep Architecture for Matching Short Texts », p. 9.
[3]J. Guo, Y. Fan, Q. Ai, et W. B. Croft, « A Deep Relevance Matching Model for Ad-hoc Retrieval », Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16, p. 55–64, 2016, doi: 10.1145/2983323.2983769.
[4]Y. Yue, T. Finley, F. Radlinski, et T. Joachims, « A support vector method for optimizing average precision », in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07, Amsterdam, The Netherlands, 2007, p. 271, doi: 10.1145/1277741.1277790.
[5]Q. Wu, C. J. C. Burges, K. M. Svore, et J. Gao, « Adapting boosting for information retrieval measures », Information Retrieval, vol. 13, n° 3, p. 254–270, juin 2010, doi: 10.1007/s10791-009-9112-1.
[6]J. Xu et H. Li, « AdaRank: A Boosting Algorithm for Information Retrieval », p. 8, 2007.
[7]D. L. T. Rohde, L. M. Gonnerman, et D. C. Plaut, « An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence », p. 33.
[8]A. Joulin, E. Grave, P. Bojanowski, et T. Mikolov, « Bag of Tricks for Efficient Text Classification », arXiv:1607.01759 [cs], août 2016.
[9]B. Hu, Z. Lu, H. Li, et Q. Chen, « Convolutional Neural Network Architectures for Matching Natural Language Sentences », arXiv:1503.03244 [cs], mars 2015.
[10]M. E. Peters et al., « Deep contextualized word representations », arXiv:1802.05365 [cs], mars 2018.
[11]E. Cosijn et P. Ingwersen, « Dimensions of relevance », Information Processing & Management, vol. 36, n° 4, p. 533–550, juill. 2000, doi: 10.1016/S0306-4573(99)00072-2.
[12]Q. Le et T. Mikolov, « Distributed Representations of Sentences and Documents », in Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196, p. 9.
[13]Z. S. Harris, « Distributional Structure », WORD, vol. 10, n° 2, p. 146–162, août 1954, doi: 10.1080/00437956.1954.11659520.
[14]M. Baroni, G. Dinu, et G. Kruszewski, « Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors », in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, 2014, p. 238–247, doi: 10.3115/v1/P14-1023.
[15]T. Mikolov, K. Chen, G. Corrado, et J. Dean, « Efficient Estimation of Word Representations in Vector Space », In : Advances in neural information processing systems, sept. 2013.
[16]Y. Nakamoto, « FOREWORD », IEICE Transactions on Information and Systems, vol. E94-D, n° 1, p. 1–2, 2011, doi: 10.1587/transinf.E94.D.1.

[17]P. D. Turney et P. Pantel, « From Frequency to Meaning: Vector Space Models of Semantics », Journal of Artificial Intelligence Research, vol. 37, p. 141–188, févr. 2010, doi: 10.1613/jair.2934.
[18]J. Pennington, R. Socher, et C. Manning, « Glove: Global Vectors for Word Representation », in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, p. 1532–1543, doi: 10.3115/v1/D14-1162.
[19]S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman, « Indexing by latent semantic analysis », Journal of the American Society for Information Science, vol. 41, n° 6, p. 391–407, sept. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
[20]A. Berger, « Information Retrieval as Statistical Translation », ACM SIGIR Forum, vol. 51, n° 2, p. 8, 2017.
[21]G. Zuccon, B. Koopman, P. Bruza, et L. Azzopardi, « Integrating and Evaluating Neural Word Embeddings in Information Retrieval », in Proceedings of the 20th Australasian Document Computing Symposium on ZZZ - ADCS '15, Parramatta, NSW, Australia, 2015, p. 1–8, doi: 10.1145/2838931.2838936.
[22]T. Francesiaz, R. Graille, et B. Metahri, « Introduction aux modèles probabilistes utilisés en Fouille de Données », p. 27.
[23]G. Salton et M. J. McGill, Introduction to modern information retrieval. New York: McGraw-Hill, 1983.
[24]D. M. Blei, « Latent Dirichlet Allocation », p. 30.
[25]P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, et L. Heck, « Learning deep structured semantic models for web search using clickthrough data », in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13, San Francisco, California, USA, 2013, p. 2333–2338, doi: 10.1145/2505515.2505665.
[26]Y. Shen, X. He, J. Gao, L. Deng, et G. Mesnil, « Learning semantic representations using convolutional neural networks for web search », in Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion, Seoul, Korea, 2014, p. 373–374, doi: 10.1145/2567948.2577348.
[27]T.-Y. Liu, Learning to Rank for Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
[28]C. Burges et al., « Learning to rank using gradient descent », in Proceedings of the 22nd international conference on Machine learning - ICML '05, Bonn, Germany, 2005, p. 89–96, doi: 10.1145/1102351.1102363.
[29]C. J. Burges, R. Ragno, et Q. V. Le, « Learning to Rank with Nonsmooth Cost Functions », in Advances in Neural Information Processing Systems 19, B. Schölkopf, J. C. Platt, et T. Hoffman, Éd. MIT Press, 2007, p. 193–200.
[30]Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, et H. Li, « Learning to rank: from pairwise approach to listwise approach », in Proceedings of the 24th international conference on Machine learning - ICML '07, Corvalis, Oregon, 2007, p. 129–136, doi: 10.1145/1273496.1273513.
[31]O. Levy et Y. Goldberg, « Linguistic Regularities in Sparse and Explicit Word Representations », in Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Ann Arbor, Michigan, 2014, p. 171–180, doi: 10.3115/v1/W14-1618.
[32]A. P. Dempster, N. M. Laird, et D. B. Rubin, « Maximum Likelihood from Incomplete Data Via the EM Algorithm », Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, n° 1, p. 1–22, sept. 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.
[33]P. Li, Q. Wu, et C. J. Burges, « McRank: Learning to Rank Using Multiple Classification and Gradient Boosting », p. 8.
[34]S. Galeshchuk et B. Chaves, « Modélisation thématique à l'aide des plongements lexicaux issus de Word2Vec », p. 5.
[35]B. Mitra et N. Craswell, « Neural Models for Information Retrieval », arXiv:1705.01509 [cs], mai 2017.
[36]T. Hofmann, « Probabilistic Latent Semantic Indexing », ACM SIGIR Forum, vol. 51, n° 2, p. 8, 2017.
[37]K. Lund et C. Burgess, « Producing high-dimensional semantic spaces from lexical co-occurrence », Behavior Research Methods, Instruments, & Computers, vol. 28, n° 2, p. 203–208, juin 1996, doi: 10.3758/BF03204766.
[38]A. Shashua et A. Levin, « Ranking with Large Margin Principle: Two Approaches », in Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun, et K. Obermayer, Éd. MIT Press, 2003, p. 961–968.
[39]V. Lavrenko et W. B. Croft, « Relevance-Based Language Models », p. 8.
[40]G. Cao, J.-Y. Nie, J. Gao, et S. Robertson, « Selecting good expansion terms for pseudo-relevance feedback », in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08, Singapore, Singapore, 2008, p. 243, doi: 10.1145/1390334.1390377.
[41]G. H. Golub et C. Reinsch, « Singular value decomposition and least squares solutions », p. 18.
[42]S. E. Robertson et S. Walker, « Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval », in SIGIR '94, B. W. Croft et C. J. van Rijsbergen, Éd. London: Springer London, 1994, p. 232–241.
[43]D. Cossock et T. Zhang, « Subset Ranking Using Regression », in Learning Theory, Berlin, Heidelberg, 2006, p. 605–619, doi: 10.1007/11776420_44.
[44]N. Abdul-Jaleel et al., « UMass at TREC 2004: Novelty and HARD »:, Defense Technical Information Center, Fort Belvoir, VA, janv. 2004.
[45]J. Rocchio, « XXIII. RELEVANCE FEEDBACK IN INFORMATION RETRIEVAL », p. 18.

## Evaluation des sources

### Source 1 :

**[15]T. Mikolov, K. Chen, G. Corrado, et J. Dean, « Efficient Estimation of Word Representations in Vector Space », in Advances in neural information processing systems, sept. 2013.**

Cet article est l'une des sources principales fournies par notre encadrant. Il a été publié en septembre 2016 dans « Neural Information Processing Systems » par T. Mikolov, K. Chen, G. Corrado, et J. Dean qui sont des chercheurs chez Google et a été cité plus de 18000 fois (par google scholar). Les auteurs débutent leur article avec une revue de l'état de l'art en citant chaque source, ils proposent par la suite leur modèle et extensions, les résultats de leurs tests et à la fin une comparaison avec des résultats publiés par d'autres auteurs du modèle.
Le but de ce papier est de présenter plusieurs extensions qui améliorent à la fois la qualité des vecteurs et la vitesse de l'apprentissage.

### Source 2 :

**[12]Q. Le et T. Mikolov, « Distributed Representations of Sentences and Documents », in Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196, 2014**

Cet article écrit par Q. Le et T. Mikolov qui sont des chercheurs chez google, a été publié dans « Proceedings of the 31st International Conference on Machine Learning » en 2014.
Nous avons trouvé l'article par une simple recherche sur arxiv, et nous avons également constaté qu'il a été cité plus de 5000 fois (par Google Scholar).
Les auteurs débutent leur article avec une revue de l'état de l'art en citant chaque source, puis par une présentation des algorithmes dont ils se sont inspirés, ensuite leur framework avec les expérimentations, et enfin, ils donnent les résultats de leurs tests et comparent avec des résultats publiés par d'autres auteurs.
Le but de ce papier est de propose « Paragraph Vector », un framework non supervisé qui apprend des vecteurs de représentations continues et distribuées de morceaux de textes.

### Source 3 :

**[21]G. Zuccon, B. Koopman, P. Bruza, et L. Azzopardi, « Integrating and Evaluating Neural Word Embeddings in Information Retrieval », in Proceedings of the 20th Australasian Document Computing Symposium on ZZZ - ADCS '15, Parramatta, NSW, Australia, 2015,**
Cet article est l'une des sources trouvées comme référence l'un des autres articles l. Il a été publié en septembre 2015 dans « Proceedings of the 20th Australasian Document Computing Symposium on ZZZ - ADCS '15, Parramatta, NSW, **Australia** » par G. Zuccon, B. Koopman, P. Bruza, et L. Azzopard. Les trois premiers sont des chercheurs à Queensland University of Technology, le deuxième particulièrement est également chercheur à Australian e-Health Research, enfin, le dernier est chercheur à University of Glasgow.
Les auteurs débutent leur article avec une revue de l'état de l'art tout en essayant de montrer les améliorations qu'ils peuvent apporter, puis ils présentent le « language modelling framework » et quelques autres modèles, Enfin ils décrivent les des expérimentations et tests qu'ils ont faits, et donnent les résultats obtenus.
Le but de ce papier est de déterminer comment les 'words embeddings' peuvent être utilisées dans un modèle de recherche d'information et quels avantage pourraient-ils apporter.