

Projet RITAL - Reproduction de papiers

L'objectif de ce projet est multiple :

- Découvrir différentes applications de la recherche d'information (et du traitement automatique du langage)
- Comprendre des papiers scientifiques
- Reproduire les expérimentations de papiers scientifiques
- Aller plus loin dans les expérimentations et les analyses

Etapes du projet

- Choisir un article par binôme. Se concerter sur l'ensemble de la promo, un article ne peut pas être choisi par plusieurs groupes.
- Lire l'article choisi, comprendre la contribution, les modèles, le protocole expérimental et identifier les analyses réalisées
- Collecter les données - si les données sont non disponibles, vous pouvez exploiter des données en libre accès (e.g. MS Marco, ou d'autres jeux de données disponibles sous PyTerrier)
- Implémenter le modèles
- Réaliser les expérimentations
- Prendre du recul sur les résultats/expérimentations. Qu'est-il possible d'améliorer
- Proposer et évaluer des (petites) améliorations si vous le souhaitez
- Mettre en place des évaluations utilisateurs quand le papier le permet. Demandez à un binôme de jouer les utilisateurs. Analyser les résultats

Documents attendus : évaluation

L'évaluation sera sous forme de soutenance (date à déterminer, mais probablement après les vacances d'avril).

Dans les slides, devront figurer les informations suivantes :

- Sujet du papier et quelles sont ses contributions (idées générales)
- Le positionnement par rapport à l'état de l'art
- Description de la contribution
- Forces et faibles de la contribution
- Protocole d'évaluation
- Résultats/analyses
- Avez-vous trouvé le papier facilement reproductible ? Quels ajustements ont été nécessaires ? Pour quelles raisons ?

Liste des papiers possibles

Suggestions de papiers possibles (vous pouvez également identifier des papiers dans les conférences SIGIR et ECIR - à valider ensemble pour évaluer la difficulté) :

- Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators (<https://arxiv.org/pdf/2111.13057.pdf>)
- Weakly Supervised Label Smoothing (<https://arxiv.org/pdf/2012.08575.pdf>)
- Searching and Stopping : An Analysis of Stopping Rules and Strategies (<http://eprints.gla.ac.uk/112731/1/112731.pdf>)
- Ad Hoc Table Retrieval using Semantic Similarity (<https://arxiv.org/pdf/1802.06159.pdf>)
- Local and Global Query Expansion for Hierarchical Complex Topics (<https://maroo.cs.umass.edu/pub/web/getpdf.php?id=1346>)

- WikiPassageQA : A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval (<https://arxiv.org/pdf/1805.03797.pdf>)
- User Intent Prediction in Information-seeking Conversations (<https://arxiv.org/pdf/1901.03489.pdf>)
- On the Benefit of Incorporating External Features in a Neural Architecture for Answer Sentence Selection (http://www.marksanderson.org/publications/my_papers/sigir2017d.pdf)
- A Language Modeling Framework for Selective Query Expansion (<http://ciir.cs.umass.edu/pubfiles/ir-338.pdf>)
- Combining global and local semantic contexts for improving biomedical information retrieval (<https://hal.archives-ouvertes.fr/hal-00588336/document>) - papier sur le médical mais peut être utilisé avec Ms Marco et Wordnet
- Off the Beaten Path : Let's Replace Term-Based Retrieval with k-NN Search (<https://arxiv.org/abs/1610.10001>)
- Retrieving Passages and Finding Answers (<https://ciir-publications.cs.umass.edu/pub/web/getpdf.php?id=1155>)
- Explicit Search Result Diversification through Sub-Queries (http://terrierteam.dcs.gla.ac.uk/publications/ecir2010_rodrygo_div.pdf)
- Query Performance Prediction Focused on Summarized Letor Features (<https://dl.acm.org/doi/abs/10.1145/3209978.3210121>)
- Topic-Sensitive PageRank (<http://www-cs-students.stanford.edu/~taherh/papers/topic-sensitive-pdf>)
- LDA-Based Document Models for Ad-hoc Retrieval (<http://ciir.cs.umass.edu/pubfiles/ir-464.pdf>)
- Accelerating Learned Sparse Indexes Via Term Impact Decomposition (<https://aclanthology.org/2022.findings-emnlp.205/>)
- Injecting the BM25 Score as Text Improves BERT-Based Re-rankers (<https://arxiv.org/abs/2301.09728>)
- Zero-shot Query Contextualization for Conversational Search (<https://arxiv.org/abs/2204.10613>)
- Faster Learned Sparse Retrieval with Guided Traversal (<https://dl.acm.org/doi/abs/10.1145/3477495.3531774>)
- Offline Retrieval Evaluation Without Evaluation Metrics (<https://arxiv.org/abs/2204.11400>)

Attention Certains papiers comportent des modèles neuronaux parfois complexes. Ne considérer que les modèles que vous savez implémenter ! Choisissez des papiers qui ne sont pas en lien avec PLDAC !