ACL-rlg: Un dataset pour la génération de listes de lecture

Julien Aubert-Béduchaud[1] Florian Boudin[1, 2]
Béatrice Daille[1] Richard Dufour[1]
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
(2) JFLI, Tokyo, Japan
julien.aubert-beduchaud@univ-nantes.fr

RÉSUMÉ

Se familiariser avec un nouveau domaine scientifique et sa littérature associée peut s'avérer complexe en raison du nombre considérable d'articles disponibles. Les listes de références académiques compilées par des experts, également appelées listes de lecture, offrent un moyen structuré et efficace d'acquérir une vue d'ensemble approfondie d'un domaine scientifique. Dans cet article, nous présentons ACL-rlg, le plus grand ensemble de données ouvertes rassemblant des listes de lecture annotées par des experts. Nous proposons également plusieurs bases de référence pour évaluer la génération de listes de lecture, que nous formalisons comme une tâche de récupération d'information. Notre étude qualitative met en évidence les performances limitées des moteurs de recherche académiques traditionnels et des méthodes d'indexation dans ce contexte, tandis que GPT-4o, bien que produisant de meilleurs résultats, présente des signes potentiels de contamination des données.

ABSTRACT

ACL-rlg : A Dataset for Reading List Generation

Familiarizing oneself with a new scientific field and its existing literature can be daunting due to the large amount of available articles. Curated lists of academic references, or reading lists, compiled by experts, offer a structured way to gain a comprehensive overview of a domain or a specific scientific challenge. In this work, we introduce ACL-rlg, the largest open expert-annotated reading list dataset. We also provide multiple baselines for evaluating reading list generation and formally define it as a retrieval task. Our qualitative study highlights the fact that traditional scholarly search engines and indexing methods perform poorly on this task, and GPT-4o, despite showing better results, exhibits signs of potential data contamination.

MOTS-CLÉS : listes de lecture, jeu de données, recherche d'information, recommandation d'articles, contamination des données.

KEYWORDS: reading lists, dataset, information retrieval, article recommandation, data contamination.

ARTICLE : **Accepté à** 31st International Conference on Computational Linguistics (COLING 2025) (https://aclanthology.org/2025.coling-main.327).