6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Comparaison de méthodes d'extraction de mots-clés non supervisées pour les disciplines des Sciences Humaines et Sociales

Alaric TABARIES
Master Information et Communication, Université de Toulon, France
Encadré par David REYMOND, IMSIC, 70 Avenue Roger Devoucoux, 83000 Toulon
alaric-tabaries@etud.univ-tln.fr

Accéléré par l'émergence de la voie verte, la quantité d'information scientifique disponible en ligne augmente à un rythme sans précédent. Ce phénomène rend le processus de veille documentaire, essentiel à la recherche scientifique, tant complexe que chronophage. C'est dans ce contexte que l'extraction d'information se pose en tant que service support au prétraitement de la sélection documentaire. En effet, les mots-clés, qui représentent les sujets principaux traités dans un document, sont particulièrement utiles pour distinguer les ressources intéressantes dans un ensemble de documents important. Cependant, très peu en sont pourvus. L'extraction automatique de mots-clés permet de remédier à ce problème et montre d'ores et déjà des résultats satisfaisants sur des corpus de référence. Il a cependant été établi que certaines méthodes d'extraction performent mieux que d'autres pour les productions dans les disciplines des Sciences Humaines et Sociales.

Nous proposons donc de mettre au point une expérimentation sur des jeux de données réels issus de publications identifiées sur la plateforme HAL en comparant les résultats selon les disciplines des publications afin d'identifier les méthodes d'extraction non supervisées qui performent le mieux pour servir un outil veille répondant au problème de surcharge informationnelle. Cette expérimentation consiste donc à comparer des mots-clés extraits de résumés de publications HAL à l'aide de méthodes non supervisées à des mots-clés préalablement annotés par des étudiants de Master Langues et Sociétés dans le but d'établir des mesures de pertinence ainsi qu'un classement de performance des différentes méthodes d'extraction selon diverses disciplines des Sciences Humaines et Sociales.

Le tableau 1 présente une partie des résultats obtenus.

	Sciences de l'éducation	Langues	Médias et communication	Histoire	Gestion	Psychologie
TfIdf	5.385	6	5.333	4.125	4	3.714
KPMiner	6.462	6.5	5.444	3.875	3.375	6.571
YAKE	3.462	5	5.778	2.875	**2.25**	5.286
TextRank	3.769	4	3.111	4.375	4.25	2.857
SingleRank	5	4.333	3.111	4.75	3.25	**2.714**
TopicRank	6.769	7.083	5.667	5	6.125	4.286
TopicalPageRank	**3.154**	**2.083**	**2**	**2.375**	2.875	4.143
PositionRank	3.462	3	3.222	3.625	3	3.714
MultipartiteRank	6.846	5.833	5.222	4.875	5.25	4.857

TABLE 1 – Rang moyen par méthode d'extraction pour des disciplines des SHS en anglais

Communications des apprenti-e-s chercheur-euse-s 2020, pages 12–12.