

RÉDUCTION DE LA MASSE DOCUMENTAIRE PAR DÉTECTION DE QUASI-DOUBLONS

Malgré les bonnes pratiques de travail et toute la meilleure volonté du monde, il est difficile de maintenir une seule version d'un document donné. Copie de sauvegarde, échange par courriel, brouillon conservé « au cas où », etc., tout ceci crée des artefacts de travail pouvant s'infiltrer dans votre masse documentaire alourdissant ainsi les diverses tâches liées à la gestion de vos documents.

Les experts en traitement du langage naturel de l'équipe DETI utilise une méthode basée sur la similarité de Jaccard qui permet d'identifier dans un ensemble de documents, ceux partageant une base commune et dont le contenu a été légèrement modifié.

L'INTÉRÊT DE CETTE MÉTHODE EST :

- 1 Elle fait complètement abstraction des modifications relatives à la mise en page ou à l'ordre de présentation du contenu.
- 2 Ne requiert aucune utilisation préalable de système de « versioning », et est directement applicable à vos répertoires de fichiers.

Cette méthode a permis d'obtenir des **réductions de masses documentaires de 37 %** lors d'un mandat précédent¹.

¹ LAVALLÉE, Jean-François, BARRIÈRE, Caroline, *Analyse textuelle dans le domaine du recrutement*, IC2011, volume 2, numéro 1, automne 2011 (À paraître).



Regrouper vos documents fortement similaires (FIGURE 1) de façon à conserver l'instance la plus pertinente et/ou d'appliquer les mêmes traitements à l'ensemble de ces documents.

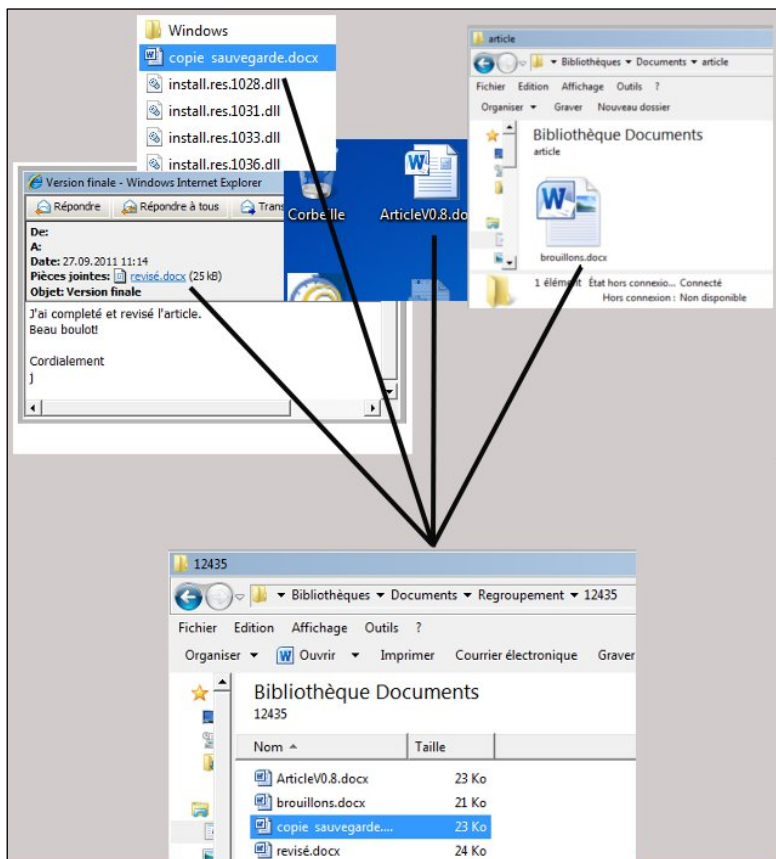


FIGURE 1 REGROUPEMENT DE QUASI-DOUBLONS.

Renseignez-vous à propos des possibilités pour intégrer à votre plateforme de gestion documentaire cette technologie de détection de quasi-doublons développée par l'équipe Développement et technologies Internet (DETI) du CRIM.

RECHERCHE ET DÉVELOPPEMENT :

Jean-François Lavallée, M. Sc., agent de recherche, équipe DETI, CRIM

INFORMATION :

Sacha Leprêtre, directeur de l'équipe DETI, CRIM

Tél. : 514 840-1238 ou 1 877 840-2746 - sacha.lepretre@crim.ca