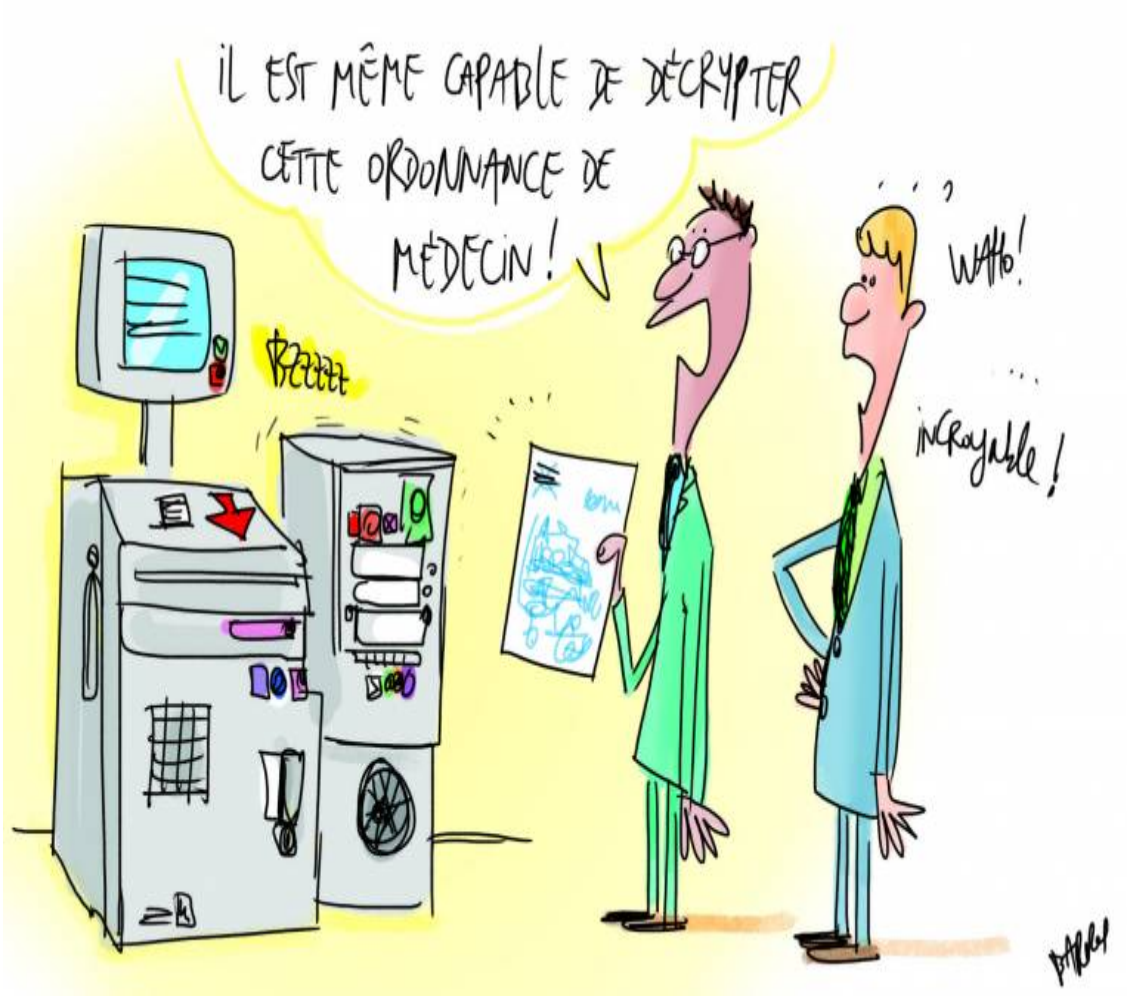
OCR, une technologie qui a de plus en plus de caractère

Le 01/12/2011 Anonyme (non vérifié)



Barros

La reconnaissance optique de caractères affiche des taux de résolution toujours plus performants. Utilisée pour les collections patrimoniales, elle est de plus sollicitée par les entreprises et se décline désormais en mode Saas.

Les 24 et 25 octobre derniers, la vénérable British Library de Londres accueillait la conférence du projet Impact. Ce programme européen, lancé en 2008, vise un objectif : améliorer les techniques de numérisation à grande échelle. L'un des points abordés par les conférenciers concernait la reconnaissance optique de caractères (OCR), étape indispensable – et parfois sous-estimée - de tout chantier de numérisation : « La reconnaissance optique de caractères, l'indexation et l'extraction de connaissances dans les textes sont aussi nécessaires pour déverrouiller les collections physiques et les transformer en trésors numériques », estime Richard Boulderstone, directeur de l'e-stratégie et des systèmes d'information au sein de la British Library.

Le projet Impact compte 26 partenaires dont plusieurs institutions culturelles prestigieuses telles que les Bibliothèques nationales de France et des Pays-Bas. Il compte également des acteurs privés comme la société Abbyy qui travaille spécifiquement sur les problématiques de reconnaissance optique de caractères. Les développeurs d'Abbyy ont fourni au projet Impact les versions les plus récentes de leurs outils et services.

Mission : accélérer les processus de reconnaissance de caractère dans les textes anciens et supporter des briques logicielles d'outils linguistiques.

nettoyage d'image

De l'avis de nombreux observateurs, les progrès accomplis par les solutions d'OCR ont été considérables. « Aujourd'hui, les technologies de reconnaissance optique de caractères sont matures, constate Christophe Rebeccchi, directeur général de ReadSoft France ; les taux de reconnaissance approchent les 100 % pour les documents structurés et environ 85 % pour les documents semi-structurés au premier passage ». La technologie OCR a en effet profité des progrès réalisés en matière de nettoyage d'image. Cette opération permet de corriger les effets générés par les fonds grisés ou les tampons apposés sur les documents. Ces facteurs de « bruit » ont longtemps nui à l'efficacité des solutions de reconnaissance, mais ils sont aujourd'hui en passe d'être résolus.

achat de logiciels vs mode Saas

Les secteurs économiques les plus variés font appel à l'OCR, mais, en France, ce sont les acteurs industriels et financiers qui la plébiscitent : Areva, Bouygues Construction, Schneider figurent parmi les grands comptes qui utilisent les solutions de ReadSoft. En Grande-Bretagne, ce sont les banques.

Présente dans 70 pays, la société constate également que les entreprises continuent de privilégier l'achat de logiciel d'OCR plutôt que le mode Saas. Selon Christiophe Rebecchi, « le mode achat est encore le modèle majoritaire parmi nos clients, mais le mode Saas intéresse les TPE-PME séduites par cette solution qui ne nécessite ni la présence d'informaticiens, ni maintenance interne ». Certains grands comptes transnationaux peuvent eux aussi recourir au mode Saas pour leurs filiales situées à l'étranger.

Autre éditeur spécialisé dans la reconnaissance automatique de caractères, A2iA propose une série de logiciels dédiés aussi bien au tri du courrier postal qu'à la lecture automatique de chèques ou à la dématérialisation de documents entrants. Ces documents plus ou moins structurés selon l'organisme producteur (administration, banque, entreprise...) sont composés d'une multitude de champs et d'objets documentaires qui sont automatiquement reconnus : signature, dates, cases à cocher, précasés, raturages...

L'éditeur est également présent dans le traitement de contenus historiques et patrimoniaux avec sa solution A2iA Document Reader : dans un premier temps, le logiciel analyse l'image des documents scannés en se basant sur leur géométrie et sur leur contenu ; dans un deuxième temps, il réalise une transcription littérale des zones manuscrites ou dactylographiées puis en extrait les mots ou expressions clés.

OCR pour les professionnels et OCR pour le grand-public

Du côté de l'éditeur Iris, des solutions OCR sont proposées aux professionnels, bien sûr, mais également au grand-public. Le logiciel IrisDocument Server 9, réservé aux entreprises, permet de convertir les images en fichiers modifiables et indexés, le tout à partir de 130 langues différentes (langues latines, alphabets grec et cyrillique, japonais, chinois, coréen...). Un « rapport de précision de l'OCR » donne à l'utilisateur la possibilité de comparer la qualité du processus d'OCR avec des configurations différentes. En cas de résultat insuffisant, il peut optimiser le niveau de reconnaissance des caractères. Différents formats de sortie sont proposés : PDF, suites bureautiques propriétaire et libres, RTF...

Quant au logiciel ReadIris 12, destiné au grand-public, il permet de convertir un document en texte modifiable et d'en extraire l'intégralité avec une reconnaissance de 20 langues différentes. Une version renforcée (ReadIris 12 Corporate) est disponible pour les petites entreprises.

quand l'humain fait mieux que l'OCR...

Tout le monde est tombé, un jour ou l'autre, sur un captcha, ce petit programme qui permet de différencier les êtres humains des robots à spam. L'internaute est invité à recopier un mot composé de lettres tordues et illisibles pour prouver qu'il n'est pas une machine... Selon une étude publiée dans la revue

britannique *Sciences*, les captchas auraient une autre fonction : ils serviraient à pallier les limites des logiciels de reconnaissance optique de caractères pour aider à la numérisation des livres. Les scientifiques de l'université de Pittsburgh (Etats-Unis) estiment en effet que le taux de reconnaissance de caractères approche les 99 % chez les êtres humains contre environ 80 % pour un logiciel OCR. Une performance qui n'a pas échappé à Google... Le géant de Mountain View s'est empressé de racheter la société reCaptcha dont les programmes de captcha sont installés sur plus de 100 000 sites dont quelques poids lourds comme Facebook. Objectif : contribuer à la numérisation de livres et des anciennes éditions du *New York Times* en complément d'une informatique parfois impuissante. Google estime en effet que « *le problème de l'OCR, c'est que ça n'est pas parfait* »... Concrètement, les mots qui ne sont pas reconnus par les logiciels sont présentés aux internautes qui, en les résolvant, font avancer la numérisation de Google.

Chaque jour, près de 200 millions de captchas sont résolus gratuitement par les internautes. Cette opération de reconnaissance prend en moyenne 10 secondes. A l'échelle de la planète, cela représente 150 000 heures par jour dédiées à la reconnaissance de caractères. Une aubaine pour le projet de bibliothèque numérique de Google !

Les derniers articles Archimag

Tic tac tic tac...
Abonné

Entrée difficile du bulletin de paie dans le coffre-fort électronique

Trésors d'archives
Abonné

Archives audiovisuelles : des plateformes pour les professionnels et le grand public

Droit
Abonné

Diffuser la jurisprudence dans le mouvement de l'open data

Geek

Barack Obama archiviste compulsif des réseaux sociaux

Trésor caché
Abonné

Le tweet-documentaire : un format inédit qui a fait le buzz

Portrait

Clotilde Vaissaire-Agard, apôtre de la formation

Tribune

La numérisation est morte ? Vive la numérisation !

Patrimoine

Réouverture des bibliothèques du site Richelieu : acte I

Toc, toc, toc !

Open data : les données en temps réel de la RATP sont enfin ouvertes !

C'est dans la boîte !

3 applications gratuites pour transformer son smartphone en scanner de poche