Millian Poquet
Maître de conférences
Univ. Toulouse, IRIT, équipe Sepia
IRIT2-469, 120 cours Rose Dieng-Kuntz, 31400 Toulouse
millian.poquet@utoulouse.fr

Hekzam : fuzzy matching de numéros identifiants

Contexte

Hekzam est un ensemble de logiciels en cours de conception qui a pour objectif de permettre l'écriture d'examens papiers qui peuvent être corrigés automatiquement ou pseudo-automatiquement. Voici la manière typique dont Hekzam est utilisé pour faire et corriger un examen.

1. L'équipe enseignante crée un formulaire de sujet d'examen grâce aux logiciels fournis de Hekzam
2. Les apprenant·e·s passent leur examen en complétant le formulaire
3. L'équipe enseignante scanne les formulaires complétés
4. L'équipe enseignante utilise les logiciels fournis de Hekzam pour obtenir
   - Des notes (et analyses détaillées) pour les questions à correction complètement automatisée
   - Les saisies des apprenant·e·s (images, booléens, nombres…) pour les autres types de questions

De nombreux sous-problèmes existent dans Hekzam, et ce sujet se concentre sur l'identification des apprenant·e·s qui ont complété chaque formulaire. En examen les apprenant·e·s s'identifient en écrivant un numéro unique (à l'apprenant·e ou au couple examen apprenant·e) dans un champ dédié du formulaire. Les logiciels de Hekzam vont récupérer cette saisie via des méthodes de traitement d'image et de reconnaissance d'écriture manuscrite, en particulier via des réseaux de neurones ou autres algorithmes de classement automatique.

Objectifs du projet

Le but de ce projet est de développer un algorithme robuste, efficace et sobre pour identifier un ensemble de copies du même examen. Ce problème comprend deux entrées : une liste d'émargement (l'ensemble des identifiants des apprenant·e·s ayant participé à l'examen, typiquement signé pendant un examen), et un ensemble d'identifiants reconnus par le système automatique de Hekzam. Il faut calculer une correspondance de chaque identifiant reconnu vers les identifiants possibles. La difficulté est que le système automatique de Hekzam peut faire des erreurs et va en faire, on s'attend typiquement à une précision de reconnaissance des chifres entre 90 et 95 % en situation réelle. Il faut donc réaliser une correspondance la plus vraisemblable possible, mais aussi détecter et faire remonter des erreurs ou points de vigilances à l'équipe enseignate (apprenant·e inconnu, duplication d'un apprenant·e, doute fort sur l'identification d'un apprenant·e…).

Vous réaliserez un état de l'art (veille bibliographique) sur la question, vous implémenterez plusieurs manières de résoudre le problème, et vous comparerez la robustesse et la performance de ces implémentations sur un jeu de données factice. Vous écrirez ensuite un rapport sur ce qui a été fait sous la forme d'un mini article décrivant le problème, l'état de l'art, les implémentations réalisées et une analyse des résultats. Comme piste de départ il est attend que vous utilisiez des structures de données *avancées* comme des BK-trees, et que vous étudiez l'impact de différentes métriques de distance (Levenshtein, Hamming…). Une veille technologique sur les outils de fuzzy matching (fzf…) sera également pertinente à réaliser.

Millian Poquet
Maître de conférences
Univ. Toulouse, IRIT, équipe Sepia
IRIT2-469, 120 cours Rose Dieng-Kuntz, 31400 Toulouse
millian.poquet@utoulouse.fr

Université de Toulouse

Licences, droit d'auteur, propriété intellectuelle

Les implémentations réalisées seront faites sous licence libre ; en Apache-2.0 si possible ou dans une autre licence libre si le projet d'origine en impose déjà une. Les données issues de l'expérience, les documentations, rapports et articles seront écrits sous licence libre CC-BY. Un document de cession de droit d'auteur à l'UT sera signé pour pérenniser la liberté d'accès et de modification de vos productions.

Candidature

Pour maximiser vos chances de candidature, contactez-moi par mail avec les informations suivantes.

- Très courte motivation par rapport au sujet (2 phrases)
- Bulletins de notes (master et licence)
- CV court (parcours d'études, expériences professionnelles, compétences)

Si vous candidatez en groupe à un projet, merci de ne m'envoyer qu'un seul mail pour tout le groupe.