Fonctionnalités des outils de détection de similarités

PMU : Le plagiat de la négligence à la fraude

Lausanne, le 16 janvier 2014

Isabelle de Kaenel Isabelle.de-Kaenel@chuv.ch
Vincent Demaurex Vincent.Demaurex@chuv.ch
Jolanda Elmers Hains Jolanda.Elmers@chuv.ch

Objectifs

- Rappel: la diversité des informations électroniques
- Fonctionnalités des logiciels
- Résultats de quelques tests
- Quel rôle pour les bibliothèques?

Les ressources électroniques

- Les publications sous licences
  - Périodiques : revues W E L N S S
  - Livres : éditeurs plus diversifiés
  - Thèses : sites spécialisés
  - Synthèses : éditeurs spécialisés
  - Vidéos : nouveau marché

- Les ressources en accès ouvert sur des sites
  - Publics : rapports administratifs, thèses uni…
  - Partagés : production collective, échanges, scans…
  - Personnels : blogs, forums…

Les fonctionnalités principales

- Développement d'un index ou base de connaissance
  - Accords avec les sites non publics : éditeurs privés…
  - Robots moissonneurs pour le web public
  - Les textes soumis à vérification par les clients

- Un algorithme
  - Comparaison de chaînes de caractères
  - Périphrases, similarités de sens, de structures
  - Traductions, inversions, alphabets différents
  - Paramétrage possible : éviter les bibliographies, etc.

Les autres fonctionnalités

- Environnement informatique
  - Installation, connexion, intégration applications locales
- Chargement des fichiers
  - Copier/coller : texte brut
  - Chargement de fichiers un par un ou groupés
  - Formats des fichiers, longueur
- Affichage des similarités
  - Temps de réponse
  - Calcul et affichage du taux de similarité, liens
  - Mise en évidence des zones de texte: couleurs, fenêtre…
  - Affichage graphique d'une vue d'ensemble

Définition des corpus de textes

- Corpus de textes libres de droits à l'Université de Weimar
  - PAN-PC-11 : 27 000 documents
  - PAN-PC-09 : 41 000 documents : 50% plagiés
    - 50%: 1-10 pages
    - 35% : 10-100 pages
    - 15% : 100-1000 pages
- Université technique de Berlin
  - En 2013 un corpus de 35 fichiers, 2/3 non libres de droits
    - Copier/coller, mélanges, déguisements, traductions
    - Textes en anglais, allemand, hébreux
    - Formats pdf, doc, txt, zip

Tableaux et tests compartifs

- Etude pour le JISC en 2007
  - Technical review of plagiarism detection software report
  - 5 produits analysés selon 26 critères
  - Hitparade: Turnitin, Coypcatch, Eve2
- Comparatif Université de Pau : 2010, complété en 2012
  - Tableau comparatif des fonctions : 8 logiciels, 10 critères
- Université technique de Berlin : 6 tests 2004-2013
  - Un corpus de 35 fichiers, 15 systèmes testés , 27 critères
  - Tableau comparatif avec évaluation selon des points
  - http://plagiat.htw-berlin.de/software/

Les logiciels testés pour le symposium

- Logiciels payants
  - Coût selon le nombre d'étudiants ou de documents
  - Société iParadigms, partenaire de CrossCheck
    - Turnitin : http://turnitin.com/fr/
    - iThenticate : https://app.ithenticate.com/en_us/login

- Logiciel en libre accès
  - Copyscape (avec publicité)
    - http://www.copyscape.com/
  - Copyscape Premium
  - Utilisation de l'index des moteurs de recherche
  - Indigo Stream Technologies

Le corpus la bibliothèque de médecine

- Extraits d'articles scientifiques de janvier 2014
- Extraits d'articles scientifiques anciens sous licence
- Extraits d'articles scientifiques anciens en libre accès
- Rapid responses : lettres à l'éditeur sur site éditeur
- Extraits d'articles de revues suisses : en français, en allemand
- Extrait d'une revue Cochrane : en anglais, en français
- Traductions de l'anglais en français par Google
- Rapports : Raisons de santé
- Thèse électronique et thèse imprimée scannée + OCR
- Chapitre de livre en français ou en anglais
- Chapitre de livre traduit de l'anglais
- Mélanges: shake and paste, slicing

Résultats des tests

- Forces
  - Les articles des éditeurs internationaux sont repérés
    - sauf les articles de la semaine
  - Vitesse correcte, mais grand nombre de textes courts
  - Pas de différences entre Turnitin et iThenticate
  - De bonnes surprises avec Coypscape
- Faiblesses
  - Faux négatifs, positifs liés aux mauvaises sources
  - Les documents récents ne sont pas détectés
  - Les livres ne sont pas toujours pas détectés
  - Traductions pas détectées

Le rôle des bibliothèques

- Support à la publication: un accompagnement complet
  - Stratégies documentaires
  - Fourniture des documents
  - Formation aux logiciels bibliographiques
  - Recherche de similarités dans les publications
  - Observation du marché de l'édition scientifique
  - Inventaire des publications des chercheurs
  - Archivage des fichiers des chercheurs
- Place des bibliothèques
  - Participer à un projet institutionnel
  - « *Informer, accompagner, contrôler*… »

© Question Mark Graffiti Bilal Kamoon
http://www.flickr.com/photos/55255903@N07/6835060992