FST
FACULTÉ DES SCIENCES ET TECHNOLOGIES
Nancy

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Corector

Acquisition semi-automatique de cas de corrections de phrases en français

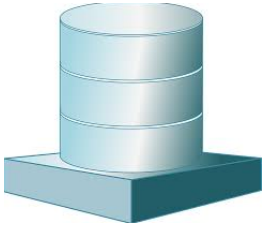
Licence informatique L3
Université de Lorraine

UNIVERSITÉ DE LORRAINE

Corrector : logiciel de correction de phrases

Projet divisé en 3 :
- Moteur de Recherche à partir de cas
- Interface Web et Base de données
- Acquisition semi-automatique de cas de correction de phrases en français

MOTEUR



Base de cas minimale

- Recherche des erreurs courantes
- Création d’un fichier csv

Exemple :

J'aime pas les pommes.	phrase fausse
Je n'aime pas les pommes.	phrase corrigé
True	statut
j'aime	groupe de mots contenant l'erreur
je n'aime	groupe de mots corrigé
1	indice de l'endroit ou se trouve la première différence
2	indice utilisateur
fr	langue

WiCoPaCo

- Historique des modifications de pages sous Wikipédia
- WiCoPaCo est réalisé sous licence GNU Free Documentation License (GFDL) qui impose l'inclusion de la citation suivante :

  Aurélien Max and Guillaume Wisniewski, Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History, LREC 2010.

Exemple

<modif id="27" wp_page_id="9" wp_before_rev_id="33563" wp_after_rev_id="110030" wp_user_id="182" wp_user_num_modif="255" wp_comment="francisation de la phrase">

<before>Lalgèbre abstraite est une branche des mathématiques , qui <m num_words="1">concerne</m> principalement des structures et des fonctions entre elles .</bbefore>

<after>Lalgèbre abstraite est une branche des mathématiques , qui <m num_words="1">porte</m> principalement sur l' étude des structures et des fonctions entre elles .</aafter>

</modif>

Extraction de cas

- Étude de la forme du fichier XML
- Extraction des couples phrase fausse/phrase corrigée
- Mise en place d’un filtre afin d’avoir un maximum de cas pertinent
- Écriture automatique des cas sélectionnés dans un fichier csv

Exemple de cas à supprimer

- <modif id="18" wp_page_id="7" wp_before_rev_id="3549368" wp_after_rev_id="3977925" wp_user_id="0" wp_user_num_modif="1096911" wp_comment="">

  <before>On nomme <m num_words="2">algèbre linéaire</m> la branche des mathématiques qui se penche sur l' étude des vecteurs (ensembles ordonnés de scalaires) , des espaces vectoriels (ou espaces linéaires) , des transformations linéaires et des systèmes d' équations linéaires (théorie des matrices) .</bbefore>

  <after>On nomme <m num_words="1">elbi</m> la branche des mathématiques qui se penche sur l' étude des vecteurs (ensembles ordonnés de scalaires) , des espaces vectoriels (ou espaces linéaires) , des transformations linéaires et des systèmes d' équations linéaires (théorie des matrices) .</aafter>

  </modif>

  <modif id="19" wp_page_id="7" wp_before_rev_id="3977925" wp_after_rev_id="3977943" wp_user_id="0" wp_user_num_modif="1096911" wp_comment="">

  <before>On nomme <m num_words="1">elbi</m> la branche des mathématiques qui se penche sur l' étude des vecteurs (ensembles ordonnés de scalaires) , des espaces vectoriels (ou espaces linéaires) , des transformations linéaires et des systèmes d' équations linéaires (théorie des matrices) .</bbefore>

  <after>On nomme <m num_words="2">Algebre lineaire</m> la branche des mathématiques qui se penche sur l' étude des vecteurs (ensembles ordonnés de scalaires) , des espaces vectoriels (ou espaces linéaires) , des transformations linéaires et des systèmes d' équations linéaires (théorie des matrices) .</aafter>

  </modif>

Statistiques

- Nombre d’éléments sans filtre : 408816
- Nombre de couples supprimés grâce au filtre présenté : 54177
- Nombre d’éléments restants après l’utilisation du filtre : 300462
- Pourcentage d’éléments supprimés : ~26,5 %

Choix du langage : python 2

- Version identique à celui utilisé pour le moteur
- Bibliothèque : xml.etree.ElementTree

Problèmes rencontrés

- La comparaison des couples est très longues
- Filtrer de manière pertinente afin d’avoir un maximum de cas utiles dans la base

Amélioration possible

- Ajout de filtres : détection de modification de chiffres / dates