TP1 : Traitement basique d'un texte : Expression régulière et mesure d'édition

Exercice 1 :

On veut récupérer **les adresses émail**, **les codes postaux** et **les numéros de téléphone** qui se figurent dans les pages "contactez-nous" des sites web de quelques instituts universitaires tunisiens.

Le problème est que ces pages représentent ces informations de plusieurs façons. Pour récupérer et unifier la forme de ces informations, on va utiliser les expressions régulières.

1. Informations recherchées

Ici, on va décrire quelques variations existantes des informations qui existent dans les pages. Ce n'est pas une liste complète ; donc, il faut ouvrir les pages où le système a échoué afin de localiser les formes non reconnues.

1.2. Téléphones

Il existe plusieurs formes des numéros de téléphone. Par exemple :

- (216) 73 683 100 ;
- (+216) 73 683 100 ;
- (216) 73683100 ;
- 73 683100 ;
- (+216) 73 68 31 00 ,
- 73 68 31 00,
- etc.

Cette liste n'est pas complète. On doit examiner les fichiers pour détecter toutes les formes possibles. La forme voulue est : "(+216) XX XX XX XX" où X est un chiffre.

1.2. Adresses mail

Il n'y a pas de conditions sur les adresses mail.

1.3. Les code postaux

Les codes postaux se trouvent généralement dans les adresses des instituts. Ils se composent généralement de 4 chiffres (exemple : 3021).

2. Travail à faire :

Vous trouvez dans votre espace de classrooms un code non complet. Vous devez compléter ce code par l'ensemble d'expressions régulières qui permettent de détecter les numéros de téléphone, les codes postaux et les adresses émail. À la fin du travail, vous devez donner les valeurs de l'évaluation de votre travail en terme de rappel, précision et F-mesure.

Exercice 2 :

On veut créer un correcteur orthographique simple pour la langue anglaise qui se base sur un lexique pour corriger et proposer les corrections correspondantes. Le principe de fonctionnement de ce correcteur est assez simple :

- Il faut parcourir un texte en faisant des comparaisons avec les mots du lexique ;
- Les mots qui appartiennent au lexique seront considérés comme corrects ;
- Pour les mots non reconnus, on calcule la distance d'édition "distance Levenshtein" avec les mots de lexique. Le mot avec la plus petite distance sera considérée comme la forme orthographique correcte.

  Pour simplifier la tâche, on va comparer les mots incorrects de longueur *l1* avec les mots de longueur *l2*, où *l2<=l1+2*.

À la fin, vous devez comparer le texte corrigé avec le texte de référence et donnez les valeurs de rappel et précision.