Algorithmes et structures de données 2
Laboratoire n°5 : Correcteur orthographique

19.12.2018

Introduction

Dans ce laboratoire, notre intention est de vous persuader que l'utilisation de structures de données adaptées est d'une importance fondamentale lorsque les données à stocker, traiter et rechercher sont nombreuses.

Objectifs

Le but de ce projet est d'implémenter un correcteur orthographique en anglais qui permet de trouver les fautes d'orthographe dans les mots composant un texte donné (allant du mot tout seul au livre complet) à l'aide d'un dictionnaire. Ce laboratoire s'inspire d'un ancien cours de programmation Ada donné l'EPFL[1].

Un mot est considéré comme correctement orthographié, s'il se trouve dans le dictionnaire de référence. Si un mot n'est pas dans le dictionnaire, votre programme devra proposer un ensemble de corrections possibles et les valider à l'aide du dictionnaire.

Durée

- 3 semaines.
  A rendre sur la page *CyberLearn* du cours au plus tard le **dimanche 20.01.2019** à **23h55**.
  Attention : un seul rendu est autorisé et considéré comme définitif.
  Séance de présentation de votre travail durant la dernière semaine du semestre.

Donnée

- Vous trouverez les structures et exemples fournis sur la page *CyberLearn* du cours.

[1] Jörg Kienzle, *Programming course: spellchecker project*, EPFL, 2002

- Contrairement aux laboratoires précédents, vous serez cette fois-ci beaucoup plus libre pour déterminer quelles seront les structures à utiliser ainsi que comment organiser votre code. Mais le langage à utiliser reste bien entendu le C++11.

- Votre programme devra lire 2 formats de fichiers différents :
  - Le **dictionnaire** : fichier texte, encodé en UTF-8 sans caractères accentués, comporte un mot par ligne.
  - Un **document texte** : fichier texte, encodé en UTF-8, comporte plusieurs lignes de plusieurs mots. Vous devrez être en mesure d'accéder aux mots un par un.

  Dans les deux cas, vous convertirez toutes les lettres en minuscules et supprimerez tous les caractères qui ne sont pas une lettre (a-z) ou une apostrophe (') en milieu de mot.

- Votre solution devra comporter au minimum 2 implémentations différentes, il doit être aisé de passer de l'une à l'autre (constante, switch, booléen à vous de voir mais cela doit être clairement expliqué) :
  - Une solution utilisant une/plusieurs structure/s STL. Vous devrez justifier votre choix dans les commentaires.
  - Une solution utilisant un *Ternary Search Trie*, que vous devrez implémenter vous-même. De la documentation est disponible dans le dossier du laboratoire sur *CyberLearn*. Pour des raisons de performance, vous équilibrerez votre *Ternary Search Trie*.

  Vous pouvez bien entendu réutiliser du code des précédents laboratoires. Mais vous devrez dans tous les cas être en mesure de comprendre et de pouvoir expliquer le code rendu, en détail. Celui-ci devra être suffisamment commenté et les éventuelles sources clairement référencées. Votre code devra bien entendu pouvoir être compilé sur l'environnement de développement fixé en début de semestre. L'échange de code entre groupes n'est pas autorisé. Le non-respect de ces consignes sera lourdement sanctionné.

- Si un mot du texte n'est pas présent dans le dictionnaire, il sera considéré comme mal orthographié. Le logiciel générera 4 ensembles de propositions de corrections sur la base des 4 hypothèses suivantes :

  1. L'utilisateur a tapé une lettre supplémentaire : acqueux → aqueux (il y a un c en trop) ;
  2. L'utilisateur a oublié de taper une lettre : aqeux → aqueux (il manque la lettre u) ;
  3. L'utilisateur a mal tapé une lettre : awueux → aqueux (il y a un w à la place du q) ;
  4. L'utilisateur a échangé 2 lettres consécutive : auqeux → aqueux (u et q intervertis).

  Ces propositions seront vérifiées avec le dictionnaire et seules celles qui s'y trouvent seront proposées à l'utilisateur.

- Pour chaque document pour lequel vous vérifierez l'orthographe, vous générerez un fichier texte comportant les mots mal orthographiés (préfixés d'une étoile) suivi immédiatement des propositions de correction vérifiées (préfixées du numéro de l'hypothèse). Les mots doivent rester dans l'ordre dans lequel ils apparaissent dans le texte, en cas de répétition d'un mot les propositions de correction devront à nouveau être présentées. Voir un exemple de la sortie attendue sur la Figure 1.
  Nous utiliserons un script pour vérifier vos résultats, vous serez donc pénalisés si vous ne respectez pas le format demandé.

- Vous afficherez dans la console le temps chargement du dictionnaire ainsi que le temps de correction du texte.

- Si le temps de chargement du dictionnaire en mémoire est supérieur à 1 minute (prétraitement compris) alors nous vous suggérons d'utiliser une version prétraitée du dictionnaire. Dans ce cas, vous devrez également nous remettre la version prétraitée du dictionnaire.

```
*lates
1:late
2:plates
2:latest
3:fates
3:gates
4:altes
…
*motsuivant
…
…
```

Figure 1 - Exemple de fichier de sortie

Rendu/Evaluation

Il n'y a pas de rapport à rendre pour ce laboratoire. Vous devrez par contre apporter une attention particulière aux commentaires dans votre code, ceux-ci devront permettre d'identifier les étapes principales de vos implémentations, les choix que vous avez effectués, ainsi que les justifications correspondantes.

Une courte séance de présentation aura lieu, groupe par groupe, lors de la dernière semaine du semestre. Un horaire de passage vous sera annoncé après que les groupes soient définitivement formés. Durant cette séance, vous présenterez votre travail, justifierez votre choix de la structure de données utilisée pour la première partie et discuterez à propos des résultats et performances des 2 méthodes utilisées. Il n'est pas nécessaire de préparer une présentation *PowerPoint*.
Le temps à disposition pour les présentations sera relativement court (max 10 min.), soyez prêts !

Bonne chance !