Examen d'ingénierie linguistique
Master 1 d'informatique
26 avril 2013

Tous documents autorisés
2h

Exercice 1 : Outils et problèmes (2 points)

Pour chacun des problèmes ci-dessous, indiquez le numéro de l'outil le plus adapté pour le résoudre. Justifiez chaque lien en une phrase.

Problèmes :

1) Finir une phrase non terminée
2) Insérer un caractère retour à la ligne à chaque fin de phrase
3) Reconnaître les noms de lieu dans un texte
4) Reconnaître les nombres dans un texte
5) Répartir en sous-dossiers un ensemble de fichiers textes placés dans un dossier

Outils :

a) algorithme des *k* moyennes
b) algorithme des *k* plus proches voisins
c) automate
d) modèle de Markov
e) transducteur

Exercice 2 : Automates finis (3 points)

Donnez une expression rationnelle et un automate fini permettant de reconnaître les dates écrites « à la française » (de type 26/04/2013 ou 26/4/2013) ou « à l'américaine » (de type 04/26/2013 ou 4/26/2013), pour les années allant de 1000 à 9999.

Remarque : Si vous utilisez une même transition pour un ensemble de lettres, vous expliciterez les abréviations utilisées.

Exercice 3 : Classification non supervisée de phrases (5 points)

On souhaite réaliser une classification hiérarchique non supervisée des cinq phrases ci-dessous :

- 1. la souris mange le fromage
- 2. le chat mange la souris
- 3. le chat ne mange pas le fromage
- 4. la souris est devant mon écran
- 5. je souris devant mon écran

a) Calculez les distances de Jaccard d(P1,P2) pour chaque paire de phrases P1 et P2 :
d(P1,P2) = 1-(C(P1,P2))/(M(P1,P2)), où C(P1,P2) est le nombre de mots distincts présents à la fois dans P1 et dans P2, et M(P1,P2) est le nombre de mots distincts présents dans P1, dans P2, ou dans P1 et P2.
Remarque : vous pouvez éviter de détailler les calculs et laisser les résultats sous forme de fractions.

b) Appliquez l'algorithme de classification hiérarchique sur la matrice obtenue au point précédent pour obtenir un arbre de classification de ces 5 phrases : lorsque vous fusionnerez deux classes C1 et C2, vous considérerez que la distance aux autres classes n'est pas la moyenne mais le minimum : pour tout k distinct de 1 et 2, d(C1+C2,Ck) = min(d(C1,Ck),d(C2,Ck)).
Faites apparaître les différentes étapes de calcul.

c) Déduisez de l'arbre de classification obtenu les deux classes qui vous semblent les plus pertinentes pour classer les 5 phrases ci-dessus en justifiant votre réponse.

Exercice 4 : Deviner la lettre suivante (4 points)

On souhaite écrire un algorithme ***DevineLettreSuivante*** qui prend en entrée une chaîne de caractères *texte* sur l'alphabet des 26 lettres a à z, et renvoie en sortie la lettre la plus probable qui suivra la dernière lettre de la chaîne de caractères *texte*, selon le modèle des bigrammes.
Par exemple, ***DevineLettreSuivante***("abababababab") renvoie "a" et ***DevineLettreSuivante***("ababacada") renvoie "b".

a) Écrivez cet algorithme en Python.
b) Quelle est la complexité de cet algorithme en fonction de *n*, le nombre de caractères de la chaîne de caractères *texte* ?
c) Quelle est la complexité de l'algorithme si on passe au modèle des *k*-grammes ?

Exercice 5 : Précision et rappel (3 points)

Supposons que vous devez programmer une méthode automatique pour repérer les noms de personnes dans un texte. L'objectif est que ce texte sera alors transformé en page web, et pour chaque mention de personne identifiée, un lien conduira à de nombreuses informations sur la personne, provenant d'une base de données. Le problème est que l'accès à la base est très cher : à chaque fois que l'on fait une requête pour y vérifier l'existence d'une personne, cela coûte 10 dollars, même si la requête a échoué (si aucune personne n'a ce nom dans la base de données). A l'inverse, ce n'est pas très gênant si la méthode n'identifie pas tous les noms de personnes du texte (par exemple : seulement 50%).

a) Expliquez en une phrase comment calculer la précision et le rappel de la méthode automatique d'identification de noms de personnes, dans ces conditions. Vous pouvez éventuellement fournir une formule en définissant précisément chacune des variables utilisées.
b) Quelle est l'évaluation la plus appropriée pour cette méthode, dans ces conditions ? Le rappel, la précision, la moyenne de précision et rappel, ou bien la F-mesure ?

Exercice 6 : Création d'une page d'actualités (3 points)

Une entreprise souhaite proposer un service de résumé d'actualités. Ils ont déjà développé des programmes qui parcourent le web (sites de journaux, notamment) à la recherche d'articles d'actualités, et stockent, pour chaque article trouvé, son URL et le code HTML correspondant à l'article (en incluant d'éventuels visuels ou photos).

Vous devez désormais utiliser ces éléments pour proposer un affichage d'articles d'actualités, triés par thématique (6 thématiques : international, France, économie, high-tech, divertissement, sport, santé), avec 6 articles portant sur des sujets différents pour chaque thématique.

a) Quelle approche choisissez-vous ? Décrivez (longueur attendue : une demi-page environ) les différentes étapes en faisant référence à des algorithmes vus en cours ou en décrivant brièvement leur fonctionnement (ce qu'ils prennent en entrée et renvoient en sortie). N'hésitez pas à faire également un schéma récapitulatif de votre chaîne de traitements.

b) Comment pourrez-vous évaluer-vous la qualité de l'approche choisie, quand elle sera programmée ? Éventuellement, vous pouvez décrire l'évaluation de certaines étapes de l'approche décrite à la question a.