Linguistique Informatique

20 mai 2016

—Kata Gábor—

Examen Linguistique Informatique

Expressions régulières

1. Indiquez les chaînes de caractères reconnues par les expressions.

   (a) `/\([^)]+\)/`

   i. ()
   ii. (((((()
   iii. (Laurent Charnois)
   iv. (Nicolas (L.) Dubois
   v. (N.D.) (L.L.)
   vi. (Pierre Maurel PhD.)

   (b) `/[^a-z]+[a-z]*$/`

   i. 8a33uau
   ii. ,= 6a2h
   iii. korte
   iv. FORMAL11
   v. 003300 alma
   vi. 778+7-19k

   (c) `/\*[^prt]\s*[0-9]/`

   i. *a1
   ii. *prt88
   iii. p 1 1
   iv. *v 7
   v. pop pop 1
   vi. *o111

   (d) `/[a-z]+[\.\?!]\s[a-z]/`

   i. a?o
   ii. c! d
   iii. e f
   iv. g. H
   v. i? J
   vi. k L

2. Donnez une expression régulière qui accepte les langages suivants définis sur l'alphabet V = 0,1 :

   (a) Toutes les chaînes qui terminent par un nombre pair de zéros,

   (b) Toutes les chaînes dont chaque 0 est suivi d'au moins trois 1s,

   (c) Toutes les chaînes dans lesquelles le nombre de 1s consécutifs ne dépasse pas deux,

   (d) Toutes les chaînes ne contenant pas plus de deux 1s au total.

3. Donnez l'expression régulière qui désigne

   (a) une chaîne qui ne contient que des mots (suites de caractères alphabétiques anglais) séparés par des espaces,

   (b) une chaîne qui ne contient qu'un nombre impair de mots (suites de caractères alphabétiques anglais) séparés par des espaces,

   (c) les mots de 10 lettres maximum qui commencent et terminent par une consonne et qui ne contiennent pas plus de deux voyelles (sur l'alphabet anglais),

   (d) les adresses physiques aux formats suivants (les numéros, ainsi que les noms des rues, villes et pays étant variables) :
   22, rue Crimée
   33, rue de Charonne 75011
   115, rue Alexandre Dumas, 75011 Paris FRANCE
   9bis place Voltaire Paris

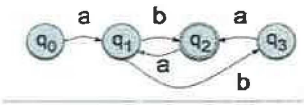
4. Construisez des commandes pour remplacer :

   (a) l'extension des noms de fichiers .jpg en .png,

   (b) les suites de deux mots commençant par une majuscule en 'NOM PROPRE'

   (c) le troisième et le sixième caractère de n'importe quelle chaîne en '3' et '6' respectivement.

5. Construisez l'expression régulière qui reconnaît les mêmes chaînes que l'automate suivant.



Automates & Morphologie

6. Donnez un exemple, en français, des phénomènes morphologiques suivants. Indiquer la forme du suffixe, un exemple de radical et de mot dérivé correspondant. (Evitez les exemples qui figurent sur les polys distribués.)
   (a) Dérivation verbe → nom
   (b) Dérivation nom → nom
   (c) Dérivation adjectif → nom
   (d) Dérivation adjectif → verbe

7. Considérez le tableau de conjugaison allemande du verbe 'singen' (*chanter*) avec les formes en indicatif présent, passé et subjonctif présent.
   (a) Construisez un automate *minimal* qui reconnaît toutes ces formes.
   (b) Pour chaque temps et mode verbal listés, construisez un transducteur individuel qui reconnaît les formes et associe à chaque forme l(es) analyse(s) correspondantes(s).

Personne	Indicatif présent	Indicatif passé	Subjonctif présent
S1	singe	sang	sänge
S2	singst	sangst	sängest
S3	singt	sang	sänge
PL1	singen	sangen	sängen
PL2	singt	sangt	sänget
PL3	singen	sangen	sängen

8. Considérez le tableau de déclinaison de l'adjectif hongrois *nagy* (grand).
   (a) Construisez un automate *minimal* qui reconnaît toutes les formes.

   (b) Construisez un transducteur *minimal* qui reconnaît les formes et associe à chaque forme l(es) analyse(s) correspondantes(s).

Cas	Base Singulier	Base Pluriel	Comparatif Singulier	Comparatif Pluriel
Nominatif	nagy	nagyok	nagyobb	nagyobbak
Accusatif	nagyot	nagyokat	nagyobbat	nagyobbakat
Datif	nagynak	nagyoknak	nagyobbnak	nagyobbaknak
Ablatif	nagytól	nagyoktól	nagyobbtól	nagyobbaktól

Probabilités et modèles de langue

9. Nous avons un moteur de recherche qui retourne des pages sur une base de 1.000.000 de documents indexées. A la suite d'une requête d'utilisateur, le moteur propose 551 documents pertinents.
   En réalité, il y avait 900 documents pertinents par rapport à la requête.
   Sur les documents retournés, 480 sont pertinentes.

   (a) Calculez le rappel et la précision du moteur partant de cet exemple.
   (b) Nous avons un deuxième moteur de recherche, avec une base de 200.000 pages web indexées. A la suite d'une requête d'utilisateur, le moteur propose 29 pages pertinentes.
   Chacune des 29 pages est pertinente.
   En réalité, il y avait 27.000 pages pertinentes.
   Calculez la F-mesure des deux moteurs. Lequel des deux est plus performant ?

Considérez le mini-corpus suivant (163 mots) :

Les insectes font-ils partie de notre avenir nutritif ?
Ils font certainement partie de l'avenir nutritif d'une partie de l'humanité. Notamment des populations rurales des pays en développement qui en consomment déjà. En Afrique centrale, au Cameroun, au Mozambique, dans les deux Congos ou encore en Zambie, les habitants mangent des milliers de tonnes de chenilles chaque année. Cette pratique de l'entomophagie se généralise aux nouvelles populations urbaines paupérisées. D'autre part, dans les pays occidentaux, les insectes, s'ils ne se retrouvent pas directement dans nos assiettes, pourraient concourir à l'alimentation du bétail, notamment dans les farines animales. Les recherches actuelles concernent des larves de mouches et de coléoptères. Si tous les humains, à l'horizon démographique de 9 milliards d'habitants sur la planète d'ici à 2050, consomment de la viande

de boeuf ou de la volaille, nous allons dans le mur. Car il existe, de fait, une concurrence entre l'alimentation des élevages et celle des humains. Les insectes peuvent contribuer à diminuer la pression.

10. Estimez les probabilités suivantes (**<S> : début de phrase**) :
    (a) P(des)
    (b) P(les)
    (c) P(populations)
    (d) P(habitants)
    (e) P(la| de)
    (f) P(Les| <S>)
    (g) P(l'| de)
11. Donnez l'estimation de la probabilité des séquences suivantes **selon un modèle unigram** et selon un modèle **bigram**.
    (a) (de, fait)
    (b) (et, de)
    (c) (le, de, le)
    (d) (Car, il, existe)
    (e) (Les, insectes)
    (f) (l'entomophagie, se, généralise, aux, nouvelles, populations)
    (g) (si, tous, les, insectes)
    (h) (nourriture, dans, nos, assiettes)