Modélisation linguistique pour l'analyse automatique de textes
Les expressions régulières (RegEx)

Aline ETIENNE
acm.etienne@gmail.com

Université Paris Nanterre
17 octobre 2019

Définition et utilité

Les expressions régulières ou RegEx

- ≃ rechercher (/remplacer) puissant
- système rapide et puissant de recherche de motifs dans des chaînes de caractères
- suivent une syntaxe particulière
- rechercher un motif pour l'extraire, le remplacer, découper une chaîne de caractères etc.

Exemples

Quelques exemples simples

- t[iao]c ➔ tic, tac, toc
- [-_a-zA-Z0-9.]+@[-_a-zA-Z0-9.]+ ➔ des adresses mails (ex. acm.etienne@gmail.com)
- \d{4} ➔ une suite composée de 4 chiffres exactement (une date, par exemple)
- \W+ ➔ une suite d'au moins un caractère non-mot (ex. espace, retours à la lignes etc.)

Mémoriser un motif

Mémorisation et réutilisation de motifs

- Utile pour extraire un motif, le réutiliser etc.
- parenthèses pour mémoriser : « le (\d+) octobre (\d{4}) »
- $ pour appeler : $1 → (\d+) ; $2 → (\d{4})

Ex. : le 3 octobre 2019 ➔ $1 = 3 et $2 = 2019
donc « $1/10/$2 » donne « 3/10/2019 »

Divers utiles

Groupes de caractères

- utilisation des [] : définir un ensemble de caractères (ex. [aeiouy])
- utilisation du tiret - : 0-9 ➔ de 0 à 9 ; a-z ➔ de a à z ; A-G ➔ de A à G etc.
- \d : chiffres de 0 à 9
- \D : opposé du précédent
- \w : alphanumériques et underscore [a-zA-Z0-9_]
- \W : opposé du précédent
- \s : tous les espaces [\t\n\r\f\v]
- \S : opposé du précédent

Caractères spéciaux

- \t : tabulation ; \n : retour à la ligne (parfois avec \r pour Windows)
- ^ : début de chaîne ; $: fin de chaîne

Divers utiles

Opérateurs de répétition

- ? : exactement 0 ou 1 fois
- + : au moins 1 fois
- * : de 0 à autant qu'on veut
- {4 } : exactement 4 fois
- {2,4} : de 2 à 4 fois

Ressources

Liens utiles sur les RegEx

- Les expressions régulières dans plusieurs langages de programmation :
  https://johndcook.com/blog/regex-per
- Des exemples d'expressions régulières déjà faites :
  http://regexlib.com/Search.aspx?AspxAutoDetectCookieSupport=1
- Des exercices avec les expressions régulières :
  https://elizia.net/regex/