GL02 TD#3: Langages réguliers et expressions régulières

(*tr EN: Regular grammar and regular expression*)

1. Grammaires régulières (S → γ ou S → γB, pour γ∈ Σ)*

Cet exercice peut se réaliser aussi bien sur papier que sur ordinateur (en ce cas TextEditor ou n'importe quel bloc note fera l'affaire).

This exercise can be done with paper-pencil as well as on the computer (in this last case any text editor make the job).

Définissez les règles des grammaires régulières permettant de générer les expressions suivantes :

Define the rules of a regular grammar able to generate the following expressions:

(for instance, in order to generate the expressions: ah, aaaah, aaaaaaaaaaaaaaaah ; the following grammar is an efficient solution: S --> aS; S--> h)

a. lol, lololol …

b. ab, abc, abcabcabc …

c. ba, baba, bababiba …

// Les grammaires régulières étant équivalente à des automates à états finis, vous pouvez aborder le problème sous cet angle avant de décrire les règles de vos grammaires.

// Regular grammar being equivalent to finite-state machine, you can view the problem with this lens as a starting point.

2. Regex marathon

Les expressions régulières sont une notation compacte permettant notamment de décrire des grammaires régulières pour reconnaître des motifs de caractères dans des fichiers. Nous allons mettre en application l'utilisation des expressions régulières en travaillant sur le texte de la RFC5234 (Request for Comment) définissant le format ABNF (en vue de notre prochain cours par ailleurs).

Regular expressions are a compact notation for describing regular langages. They are especially useful at pattern matching from unstructured text data retrieved from a file or stream. Here we apply regular expression to analyze the text of the document RFC5234 (Request for Comment) that defines the ABNF format (besides we are going to discuss about this format next week).

Environnement de travail (Work environment)

Télécharger sur le bureau au format texte le document : https://tools.ietf.org/rfc/rfc5234.txt (n'oubliez pas de supprimer le fichier à la fin du TD)

Download the RFC file in text format on the desktop : https://tools.ietf.org/rfc/rfc5234.txt

(Don't forget to delete the file at the end of the class)

Ouvrez l'invite de commande/terminal et retrouvez le fichier sur votre bureau.

Open the terminal (command-line) and find the file on your desktop.

Si en lançant le terminal la dernière ligne est [Opération terminée] et que vous le parvenez pas à écrire, alors une petite opération s'impose :

If when you launch the terminal the last line is [Opération terminée] and you cannot type inside the window, then you will need to perform a supplementary operation:

- dans le menu sélectionnez "Lancer une nouvelle commande", *(in the menu select "Launch a new command")*

- dans la fenêtre de dialogue entrez : bash (*in the popup, type* bash)

- une nouvelle fenêtre de terminal cette fois-ci fonctionnelle et initialisée avec votre compte utilisateur est désormais lancée. *(result: a new working terminal is ready for work)*

Deux commandes vous seront utiles pour naviguer dans l'arborescence des fichiers et dossiers :

Two commands will be useful to navigate the files and folders structure:

- **ls :** vous permet de lister le contenu d'un dossier *(list a folder content)*

- **cd :** vous permet de vous positionner dans un dossier (ex : cd monDossier). Par ailleurs un lien symbolique vous permet de remonter dans le dossier parent : cd ..

  Help you to enter inside a filder (e.g. cd MyFolder). By the way, a symbolic link helps you to go upward in a folder hierarchy: cd ..

- Bon à savoir *(Tips)* :

  - L'invite de commande inclus un mécanisme d'autocomplétion des noms de fichier et de dossier. En saisissant les premières lettres d'un fichier/dossier vous pouvez demander à la compléter en appuyant sur Tab.

    The bash shell has an autocomplete mechanism. You just have to type the first letters of a file/folder and then to push the tab button to complete the existing name you were typing.

  - Les touches flèche haut et bas vous permettent de vous déplacer dans un historique des commandes que vous avez saisie jusqu'ici.

    The up and down arrow button, let you move in the history of the command you have typed so far.

Avec l'éditeur de texte en ligne de commande 'less', consultez le fichier pour avoir une idée de son contenu : less rfc5234.txt

Enter the command "less <myFile>" and read the file to get an idea of its content: less rfc5234.txt

Créez un nouveau document avec TextEditor (regardez dans le dossier application du dock) pour le TD, vous y enregistrerez l'ensemble des réponses aux différentes questions.

Create a new document with your favorite text editor in order to take notes along the exercize.

Utilisation de grep (Using grep)

A l'aide de la commande 'egrep' nous allons réaliser un ensemble de recherche de motifs afin de mieux comprendre les expressions régulière.

With the 'egrep' command, we are going to search for different character patterns in order to understand regular expression. (egrep is the extended version of grep – the general regular expression processor – and its syntax looks more like the one used by Javascript)

(egrep est la version étendue de grep, la syntaxe des expressions régulière est plus proche de ce que vous trouverez en javascript).

La fonction principale de egrep est à partir d'un motif (expression régulière) et d'un flux d'entrée (tel un fichier) définis en paramètres de retourner chaque ligne contenant le motif. La commande prend une expression régulière et un fichier en paramètres. Des options peuvent être ajoutées à la suite du nom de la commande (par exemple pour compter le nombre de correspondance).

The main feature of egrep is to return each line of an input stream (be it a file or not) that match a pattern defined by a regular expression. Thus, there are two mandatory parameters: the regular expression and the input stream. Options can be append to change the program behavior (for instance counting the lines instead of return them)

egrep [options] <expression_regulière> <fichier>

(La commande 'man egrep' vous permet à tout moment de consulter la documentation détaillée).

(the command 'man egrep' allows you to read the program documentation at any time)

Questions :

1/ Rechercher l'ensemble des lignes contenant le mot : "define" (les quotes/guillemets sont nécessaires pour que l'expression soit considérée comme un motif).

Search for all the lines that contains the word: "define" (the quotes are needed so that egrep sees it as a regular expression)

2/ Combien de fois apparait ce mot (vous pouvez utiliser le paramètre -c de la commande ou compter à la main :)

Count the number of occurrences of the word 'define' in the text (you can use the option -c or read all the RFC document :)

Note : Nous allons au cours des 6 prochains items découvrir les éléments fondamentaux du langage des expressions régulières

The six following items will help you to discover the basics of regular expression.

3/ Les **classes de caractères**, définies entre [], qui permettent de chercher une correspondance pour un ensemble de caractères, par exemple de [a-zA-Z]. Plusieurs classes de caractères sont prédéfinies ou bénéficient de raccourci ([:alpha:], [:alnum:], [:space:], \w). Par ailleurs le caractère "." correspond à n'importe quel caractère (espaces et ponctuations comprises). On peut définir une classe de caractère par son complémentaire grâce au caractère ^, c'est à dire par la négative : [^a-z], les caractères qui ne sont pas entre 'a' et 'z'.

*A **character class** is defined between []. It allows to search for a set or a range of chars, for instance [a-zA-Z] (ie, all the lower and upper case letter in latin alphabet). Several characters classes are predefined or benefits from shortcuts ([:alpha:], [:alnum:], [:space:], [:digit:] \w). The "." means any char (space and punctuation marks included). A character class can be defined by its complement: [^a-z] (all the chars that are different from 'a' to 'z').*

Recherchez l'ensemble des lignes contenant une date au format YYYY (il y en a 20 dans le document).

Search for each lines that contains a date following the YYYY format (there are 20 in the document).

4/ Les **quantificateurs** qui permettent définir le nombre de répétitions de tout ou partie d'un motif : * (0 ou n fois), + (au moins 1 fois), ? (0 ou une fois), {n,m} (un nombre précis d'occurrences entre n et m).

***Quantifiers** allow to define the number of occurrences of characters classes or a pattern to look for: * (0 or n times), + (at least one occurrences), ? (0 or one time), {n, m} (precise min and max boundaries).*

Recherchez les lignes contenant des entiers ayant plus de deux chiffres (il y en a 30).

Search for the lines that contain integer that compound of more than two numbers (there are 30 occurrences)

Recherchez l'ensemble des lignes contenant au moins un mot de 15 lettres (il y en a 11).

Search for the lines that contains at least one 15 letters long word (there are 11 occurrences).

5/ Des **marqueurs de début et fin** permettent d'indiquer si le motif doit se trouver en début, '^' ou en fin de ligne '$'. '\<' et '\>' pour marquer respectivement le début ou la fin d'un mot (en tant que succession de caractères séparé par un caractère d'espacement.

***Context marks** allows to tell egrep whether the pattern has to be found in the beginning of the line (with '^'), end of a line '$', as well as, '\<' and '\>' for the beginning and end of a word (as a character sequence separated by spaces).*

Recherchez les lignes se terminant par un chiffre (il y en a 61).

Search for the lines that end with a number (there are 61 occurrences).

6/ Les **alternatives** avec le caractère | permet d'exprimer une succession de motifs alternatifs foo| bar, correspond à soit la chaîne "bar", soit à la chaîne "foo".

Alternatives are expressed with the '|' symbol (ie, foo|bar will look for either the string "bar", either the string "foo").

Recherchez les lignes qui contiennent "foo" ou "bar" exactement dans le document (il y en a 12)

Search for the lines that contain precisely either "foo", either "bar" in the document (there are 12 occurrences).

7/ Les **regroupements** qui avec les () permettent de définir des sous-motifs qui permettent de segmenté le motif (exemple : aaa(b|c), c'est à dire 'aaab' ou 'aaac'). Il est possible par ailleurs de faire référence aux sous-motifs dans leur ordre d'apparition dans l'expression régulière (\1, \2, ...). Par exemple, le motif "([a-zA-Z]{3,}) \1" peut être utilisé pour détecter de potentielles erreurs de répétitions (le même mot répété deux fois à la suite dans la même ligne – il n'y en a pas dans la rfc5234).

Groups or sub-patterns *can be defined with () (for instance, aaa(b|c), that is 'aaab' or 'aaac'. Backward reference can be made to matched sub-patterns and be refered accordong to their order of appearance inside the complete regular expression (\1, \2, …). For instance, the regular expression "([a-zA-Z]{3,}) \1" can be used to detect potentially erroneous word repetitions in a text (no case in rfc5234).*

Recherchez les lignes contenant deux fois le mot "foo" ou deux fois le mot "blat".

Search for the lines that contain either the word "foo" two times, either the word "blat" two times.

8/ Echappement des caractères : Si l'on souhaite chercher pour un des caractères utilisés par le langage des expressions régulières, il faut alors faire un échappement avec '\'.

Escaping a character*: symbols that are in use in the regular expression langage can be looked for by escaping them with a '\' (for instance, \. match the dot character).*

Recherchez toutes les lignes contenant des expressions entre parenthèses (il y en a 42).

Search for all the lines that contain expressions with parenthesis (there are 42).

Le principal intérêt des expressions régulières est de permettre de rechercher des variations de motifs de caractères (et par extension de mots). Les expressions régulières nous offre un langage riche pour exprimer des motifs. Par exemple :

The main interest of regular expression is to allow to look for different words or characters patterns with an important range of complexity given the use of the operators defined above. For instance:

egrep "^[[:digit:]].*$" file.txt retournera l'ensemble des lignes de file.txt qui commencent par un chiffre (*will output all the lines that begin with a number*).

egrep "^[^0-9].*$" file.txt retournera l'ensemble des lignes de file.txt qui ne commencent pas par un chiffre (*will output all the lines that do not start with a number)* .

Note : Maintenant que vous avez les bases, vous pouvez poursuivre avec les questions suivantes.

Now you have the foundations, you can continue with the following quesitons.

9/ Recherchez toutes les lignes commençant par Internet ou RFC (attention, si vous observez bien le document, la plupart des lignes commencent par des caractères d'espacement). (il y en a 21).

Search for all the lines that begin with Internet or RFC (be careful that most of the lines starts with spacing character in the document) (there are 21 occurrences).

10/ En utilisant le quantificateur +, recherchez tous les mots qui commencent par 'def' (mais ne sont pas exactement la chaîne 'def'). Utilisez le paramètre –o pour ne voir que les caractères qui correspondent à ce que votre expression régulière capture.

Use the quantifier + and search for all the words that start with def but are not exactly the string 'def'. Use the -o parameter in order to see only the matching characters.

11/ Le mot 'def' exactement est-il présent dans le document ? Définissez une expression régulière permettant de le vérifier.

Is the word 'def' present in the document? Define a regular expression that allows to check this.

12/ Utilisez le paramètre -o pour voir uniquement les séquences de caractères correspondant à l'expression régulière définie en 6/. Essayez de n'extraire qu'un seul mot à la fois (et non la ligne entière). Faîtes afficher l'ensemble des variations des mots commençant par 'def' du document.

13/ Les expressions régulière sont-elles sensibles à la casse des caractères ?

14/ Dans rfc5234.txt, l'ensemble des références bibliographiques sont marquées entre '[' et ']'. Définissez un motif permettant d'extraire les référence bibliographique du document quand elles sont cités dans le texte et à la fin du document (et seulement celles-ci).

In rfc5234.txt all the references are written with the use of '[' and ']'. Define a pattern that extract the references of the document (in text and at the end in the references section).

15/ Définissez une expression régulière qui permettent d'extraire toutes les lignes se terminant par un mot écrit en majuscule.

Define a regular expression that allows to get all the lines that end with a word written in upper-case.

16/ Définissez et mettez en application une expression régulière permettant d'extraire les lignes des règles de grammaire au format ABNF du document (et seulement celle-ci)

Define a regular expression that allows to extract the lines of ABNF grammar rules (and only these ones).

17/ Définissez et mettez en application une expression régulière permettant d'extraire toutes les adresses email du document (avec le paramètre -o)

Define and apply a regular expression that allows to extract email addresses (use the -o option from egrep).

18/ Supposons que l'on veuille préparer une version de rfc5234.txt sans pagination (c'est à dire en supprimant les entête et pieds de page), en analysant la forme du document (espacement, indentation…) définissez une expression régulière permettant de le faire avec egrep.

Suppose we want to prepare a version of rfc5234.txt without page headers and footers. Analyse the document structure to define a regular expression to do it with egrep.

3. Pour aller plus loin (Going forward)

Afin de mieux comprendre comment les expressions régulières peuvent être utilisées en javascript, allez sur tryregex.com et complétez le tutoriel interactif suivant :

In order to understand how regular expression can be used in Javascript, go on the following websites and complete the tutorial:

http://tryregex.com/

Notez les variations et possibilités supplémentaires par rapport à grep.

What is different from egrep?

4. Rappel caractères spéciaux sous Mac Os (Special chars on Mac Os)

http://wfr.tcl.tk/1462

Alt + (	{	accolade ouvrante
Alt +)	}	accolade fermante
Alt + ⇧ + (	[	crochet ouvrant
Alt + ⇧ +)	]	crochet fermant
Alt + N	~	tilde
Alt + ⇧ + L	|	pipe
Alt + ⇧ + /	\	antislash