Κανονικές Εκφράσεις

Τις Κανονικές εκφράσεις τις συναντάμε σε πολλές μορφές. Η σύνταξη που χρησιμοποιούμε στο μάθημα είναι πιο απλή απ' αυτή, που χρησιμοποιείται, για παράδειγμα, στη **grep** εντολή του UNIX, ωστόσο και οι δύο συντάξεις έχουν την ίδια περιγραφική δύναμη. (Το περιβάλλον **grep**, παρεμπιπτόντως, χρησιμοποιείται ακόμη και στο North Pole.)

Εδώ βρίσκεται μία ενδιαφέρουσα περιγραφή των κανονικών εκφράσεων UNIX, και ειδικότερα στο **regexp**, και εδώ βρίσκεται μια πιο λεπτομερής εξήγηση.

Εδώ βρίσκονται οι UNIX **man** σελίδες (εγχειρίδιο που δημιουργείται με την εντολή **man**) του **regexp**, καθώς και των **regexp grep, egrep**, και **fgrep** (συνδεσμος1, συνδεσμος2, συνδεσμος3). Οι κανονικές εκφράσεις χρησιμοποιούνται στην διαδικασία αναζήτησης σε αρκετούς συντάκτες (editors) όπως οι **ed,** vi, και **emacs**. Ακόμη χρησιμοποιήθηκαν σε γλώσσες προγραμματισμού όπως η **Tcl** και η **Perl.** Εδώ υπάρχει ένα άρθρο, που περιγράφει εκφράσεις αναζήτησης στην **Perl** και εδώ άλλο άρθρο που περιγράφει τη χρησιμοποίηση της Perl για αναζήτηση στο Διαδίκτυο. Εδώ βρίσκονται επίσης κάποιες άλλες αναφορές για την χρησιμοποίηση κανονικών εκφράσεων στην γλώσσα **Perl**.

Στοιχειώδης Κανονικές Εκφράσεις

Μία κανονική έκφραση χρησιμοποιείται για να περιγράψει μία κανονική γλώσσα. Η κανονική έκφραση αναπαριστά ένα "μοντέλο": συμβολοσειρές που ταιριάζουν σ' αυτό το μοντέλο ανήκουν στην γλώσσα, που αυτό περιγράφει, όσες δεν ταιριάζουν, δεν ανήκουν στην γλώσσα αυτή.

Ως συνήθως, οι συμβολοσειρές αναφέρονται σε κάποιο αλφάβητο Σ.

Τονίζεται: Τα παρακάτω αποτελούν **στοιχειώδης κανονικές εκφράσεις***:*

- $\mathbf{x}$, για κάθε $\mathbf{x} \in \Sigma$,
- $\boldsymbol{\varepsilon}$, η κενή συμβολοσειρά, και
- $\varnothing$, παριστάνει την καμία συμβολοσειρά.

Έτσι, αν $|\Sigma| = \mathbf{n}$, τότε υπάρχουν $\mathbf{n+2}$ στοιχειώδης κανονικές εκφράσεις που ανήκουν σε ένα αλφάβητο Σ.

Παρακάτω περιγράφουμε τις γλώσσες, που ορίζονται από τις στοιχειώδεις κανονικές εκφράσεις:

- Για κάθε $\mathbf{x} \in \Sigma$, η στοιχειώδης κανονική έκφραση x περιγράφει την γλώσσα **{x}**. Δηλαδή την γλώσσα που έχει μια μοναδική συμβολοσειρά **"x",** που περιέχει ένα μοναδικό σύμβολο, το **'x'**.
- Η στοιχειώδης κανονική έκφραση $\boldsymbol{\varepsilon}$ περιγράφει την γλώσσα $\{\boldsymbol{\varepsilon}\}$**.** Η μοναδική συμβολοσειρά που υπάρχει στη γλώσσα αυτή είναι η κενή συμβολοσειρά.
- Η στοιχειώδης κανονική έκφραση $\varnothing$ περιγράφει την γλώσσα $\{\}$. Κενή γλώσσα, δεν υπάρχουν συμβολοσειρές στην γλώσσα αυτή.

Κανονικές Εκφράσεις

Κάθε στοιχειώδης κανονική έκφραση είναι μία κανονική έκφραση.

Τονίζεται: Χρησιμοποιώντας τους παρακάτω κανόνες, πεπερασμένου αριθμού φορές, μπορούμε να συνθέσουμε νέες κανονικές εκφράσεις:

- Αν το $\mathbf{r}$ είναι μία κανονική έκφραση, τότε είναι και η $\mathbf{(r)}$**.**
- Αν το $\mathbf{r}$ είναι μία κανονική έκφραση, τότε είναι και η $\mathbf{r^*}$.
- Αν τα $\mathbf{r_1}$ και $\mathbf{r_2}$ είναι κανονικές εκφράσεις, τότε είναι και η $\mathbf{r_1 r_2}$.
- Αν τα $\mathbf{r_1}$ και $\mathbf{r_2}$ είναι κανονικές εκφράσεις, τότε είναι και η $\mathbf{r_1 + r_2}$.

Οι παραπάνω προτάσεις σημαίνουν τα εξής:

- Οι παρενθέσεις χρησιμεύουν μόνο για ομαδοποίηση.
- Το αστέρι συμβολίζει καμία ή περισσότερες επαναλήψεις της εκάστοτε κανονικής έκφρασης που προηγείται. Έτσι, αν $\mathbf{x} \in \mathbf{\Sigma}$, τότε η κανονική έκφραση $\mathbf{x^*}$ συμβολίζει την γλώσσα $\mathbf{\{\varepsilon, x, xx, xxx, ...\}}$.
- Η παράθεση των $\mathbf{r_1}$ και $\mathbf{r_2}$ συμβολίζει τις συμβολοσειρές που περιγράφονται από το $\mathbf{r_1}$ και ενώνονται με αυτές που περιγράφονται από το $\mathbf{r_2}$. Για παράδειγμα, αν $\mathbf{x, y} \in \mathbf{\Sigma}$, τότε η κανονική έκφραση $\mathbf{xy}$ περιγράφει την γλώσσα $\mathbf{\{xy\}}$.
- Το σύμβολο της πρόσθεσης, διαβάζεται ως διαζευκτικό "ή", χρησιμοποιείται στην περιγραφή των γλωσσών που περιγράφονται από κάποιο υποσύνολο της μιας ή της άλλης κανονικής έκφρασης. Για παράδειγμα, αν $\mathbf{x, y} \in \mathbf{\Sigma}$, τότε η κανονική έκφραση $\mathbf{x+y}$ περιγράφει την γλώσσα $\mathbf{\{x, y\}}$.

Προτεραιότητα: * το αστέρι είναι πρώτο, έχει , δηλαδή, τη μεγαλύτερη προτεραιότητα, ακολουθεί η παράθεση, και τέλος το +. Για παράδειγμα , το $\mathbf{a+bc^*}$ περιγράφει την γλώσσα $\mathbf{\{a, b, bc, bcc, bccc, bcccc, ...\}}$.

Γλώσσες που περιγράφονται από Κανονικές Εκφράσεις

Υπάρχει μία απλή αντιστοιχία στις κανονικές εκφράσεις και τις γλώσσες που περιγράφονται απ' αυτές:

Κανονική έκφραση	**L(κανονική έκφραση)**
$\mathbf{x}$, **για κάθε** $\mathbf{x} \in \mathbf{\Sigma}$	$\mathbf{\{x\}}$
$\boldsymbol{\varepsilon}$	$\mathbf{\{\varepsilon\}}$
$\varnothing$	$\mathbf{\{\ \}}$

(r_1)	$L(r_1)$
r_1^*	$(L(r_1))^*$
$r_1 r_2$	$L(r_1) L(r_2)$
$r_1 + r_2$	$L(r_1) \cup L(r_2)$

Δημιουργώντας Κανονικές Εκφράσεις

Παρακάτω υπάρχουν κάποιες οδηγίες για την δημιουργία κανονικών εκφράσεων. Θα θεωρήσουμε ότι **Σ= {a, b, c}.**

Κανένα ή περισσότερα.

> Το **a*** σημαίνει ***"κανένα ή περισσότερα a."*** Αν πούμε ***"κανένα ή περισσότερα ab",*** αυτό σημαίνει, **{ε, ab, abab, ababab, ...},** και συμβολίζεται **(ab)*.** Το **ab*** δεν είναι σωστό επειδή περιγράφει την γλώσσα **a(b)*, δηλαδή την {a, ab, abb, abbb, abbbb, ...}.**

Ένα ή περισσότερα.

> Εφόσον το **a*** σημαίνει ***"κανένα ή περισσότερα a",*** μπορούμε να χρησιμοποιήσουμε το **aa*** (ή ισοδύναμα, **a*a**) για να γράψουμε ***"ένα ή περισσότερα a."*** Ομοίως, για να περιγράψουμε ***"ένα ή περισσότερα ab",*** δηλαδή, **{ab, abab, ababab, ...},** γράφουμε **ab(ab)*.**

Κανένα ή ένα.

> Μπορούμε να περιγράψουμε ένα ή κανένα **a** ως **(a+ε).**

Οποιαδήποτε συμβολοσειρά.

> Για να περιγράψουμε οποιαδήποτε συμβολοσειρά (με **Σ = {a, b, c}**), χρησιμοποιούμε την έκφραση **(a+b+c)*.**

Οποιαδήποτε μη κενή συμβολοσειρά.

Αυτό μπορεί να γραφεί σαν οποιοδήποτε σύμβολο του **Σ** ακολουθούμενο από οποιαδήποτε συμβολοσειρά: **(a+b+c)(a+b+c)*.**

Οποιαδήποτε συμβολοσειρά δεν περιέχει....

Για να περιγράψουμε οποιαδήποτε συμβολοσειρά δεν περιέχει το a (με **Σ** = **{a, b, c}**), χρησιμοποιούμε την έκφραση **(b+c)*.**

Οποιαδήποτε συμβολοσειρά περιέχει ακριβώς ένα...

Για να περιγράψουμε οποιαδήποτε συμβολοσειρά περιέχει ακριβώς ένα a, τοποθετούμε "οποιαδήποτε συμβολοσειρά δεν περιέχει ένα a", σε κάθε μεριά του **a**, με αυτόν τον τρόπο: **(b+c)*a(b+c)*.**

Παραδείγματα Κανονικών Εκφράσεων

Τα παρακάτω είναι από τις ασκήσεις του δευτέρου εργαστηρίου.

Δημιουργήστε τις κανονικές εκφράσεις για τις παρακάτω γλώσσες στο αλφάβητο **Σ** = {a, b, c}.

Όλες οι συμβολοσειρές που περιέχουν ακριβώς ένα a.

(b+c)*a(b+c)*

Όλες οι συμβολοσειρές που δεν περιέχουν περισσότερα από τρία a.

Μπορούμε να περιγράψουμε την συμβολοσειρά που περιέχει μηδέν, ένα, δύο ή τρία a (και τίποτα άλλο) ως εξής

(ε+a)(ε +a)(ε +a)

Τώρα θέλουμε να συμπεριλάβουμε συμβολοσειρές που δεν περιέχουν a όπου βρίσκονται τα X:

X(ε+a)X(ε+a)X(ε+a)X

έτσι τοποθετούμε όπου X το (b+c)* :

(b+c)*(ε +a)(b+c)*(ε +a)(b+c)*(ε +a)(b+c)*

Όλες τις συμβολοσειρές που περιέχουν τουλάχιστον μία φορά κάθε σύμβολο του Σ .

Το πρόβλημα σ' αυτή την έκφραση βρίσκεται στην σειρά με την οποία θα τοποθετήσουμε τα σύμβολα.Δεν μπορούμε να τα τοποθετήσουμε μόνο σε μία σειρά για αυτό και θα εξετάσουμε όλες τις περιπτώσεις:

abc+acb+bac+bca+cab+cba

Για να δούμε πιο εύκολα την διαδικασία ας βάλουμε το X σε κάθε σημείο που θέλουμε να τοποθετήσουμε μία διαφορετική συμβολοσειρά:

XaXbXcX + XaXcXbX + XbXaXcX + XbXcXaX + XcXaXbX + XcXbXaX

Τελικά, αντικαθιστώντας τα X με (a+b+c)* έχουμε την λύση:

(a+b+c)*a(a+b+c)*b(a+b+c)*c(a+b+c)* +
(a+b+c)*a(a+b+c)*c(a+b+c)*b(a+b+c)* +
(a+b+c)*b(a+b+c)*a(a+b+c)*c(a+b+c)* +
(a+b+c)*b(a+b+c)*c(a+b+c)*a(a+b+c)* +
(a+b+c)*c(a+b+c)*a(a+b+c)*b(a+b+c)* +
(a+b+c)*c(a+b+c)*b(a+b+c)*a(a+b+c)*

Όλες τις συμβολοσειρές που δεν περιέχουν το a πάνω από δύο φορές συνεχόμενα.

Μπορούμε πολύ εύκολα να δημιουργήσουμε μια έκφραση, που να περιέχει κανένα a, ένα a, ή ένα aa:

(b+c)*(ε +a+aa)(b+c)*

αλλά αν θέλουμε να το επαναλάβουμε, πρέπει να είμαστε σίγουροι ότι έχουμε τουλάχιστον ένα μη-a σύμβολο μεταξύ των επαναλήψεων:

(b+c)*(ε+a+aa)(b+c)*((b+c)(b+c)*(ε+a+aa)(b+c)*)*

Όλες τις συμβολοσειρές που περιέχουν το a εις τριπλούν.

(aaa+b+c)*