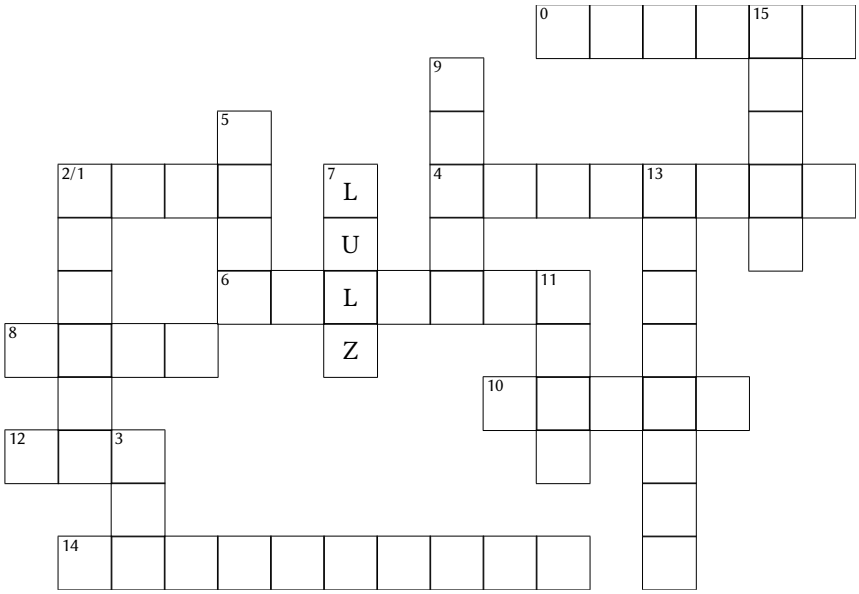
Αν. Καθηγητής Π. Λουρίδας

Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας

Οικονομικό Πανεπιστήμιο Αθηνών

Κανονικά Σταυρόλεξα



L(OL)+	(HO)+	K(EK)*E	(HAR)+)+
H(EH)+	ROT?FL	TE(HE+)+	LAW*L
MWA(HA)+	HE(HE)+	LO+L	HAHA*
(AH)+A+	HA+	(JA)+	LULZ

Στην εργασία αυτή θα φτιάξετε ένα πρόγραμμα που θα λύνει σταυρόλεξα, όπως το παραπάνω σταυρόλεξο του Alex Bellos. Οι ορισμοί των λέξεων που πρέπει να βρείτε και να τοποθετήσετε σε αυτό δεν θα είναι όπως στα συνηθισμένα σταυρόλεξα· θα δίνονται με *κανονικές εκφράσεις*.

Κανονικές Εκφράσεις

Οι κανονικές εκφράσεις (regular expressions) είναι μια ακολουθία χαρακτήρων η οποία περιγράφει συμβολοσειρές. Χρησιμοποιούμε κανονικές εκφράσεις για να περιγράψουμε μια συμβολοσειρά που δεν την γνωρίζουμε μεν ακριβώς, αλλά γνωρίζουμε πιο μοτίβο ακολουθεί.

Για παράδειγμα, αν σας ζητήσει κάποιος να αναζητήσετε σε ένα αρχείο ένα συγκε-

κριμένο αριθμό τηλεφώνου, θα προσπαθήσετε να ταιριάξετε τον αριθμό που αναζητάτε στα περιεχόμενα του αρχείου. Αλλά αν σας ζητήσει κάποιος να βρείτε όλους τους αριθμούς τηλεφώνου, τι θα κάνετε; Ομοίως, αν σας ζητήσει κάποιος να βρείτε μια συγκεκριμένη διεύθυνση ηλεκτρονικού ταχυδρομείου, μπορείτε απλώς να την αναζητήσετε στα περιεχόμενα του αρχείου. Αν όμως θέλετε να βρείτε όλες τις διευθύνσεις ηλεκτρονικού ταχυδρομείου; Ή μπορεί να θέλετε να αναζητήσετε μια λέξη, στην οποία να θέλετε να δεχτείτε κάποιες παραλλαγές στον τρόπο γραφής της, όπως ορθογραφικά λάθη· θέλετε να βρείτε τις εμφανίσεις της λέξης κτίριο, αλλά και της κτήριο.

Με τις κανονικές εκφράσεις χρησιμοποιούμε χαρακτήρες του αλφαβήτου και κάποιους άλλους, ειδικούς χαρακτήρες, για να περιγράψουμε τις συμβολοσειρές που θέλουμε. Έτσι:

- Ο ειδικος χαρακτήρας `*` σημαίνει ότι η μονάδα που προηγείται μπορεί να εμφανιστεί μηδέν ή περισσότερες φορές. Οπότε η κανονική έκφραση `ca*t` περιγράφει τις συμβολοσειρές `ct`, `cat`, `caat`, κ.λπ.
- Ο ειδικός χαρακτήρας `+` σημαίνει ότι η μονάδα που προηγείται μπορεί να εμφανιστεί μία ή περισσότερες φορές. Έτσι, η κανονική έκφραση `ca+t` περιγράφει τις συμβολοσειρές `cat`, `caat`, `caaat`, κ.λπ.
- Ο ειδικός χαρακτήρας `?`σημαίνει ότι η μονάδα που προηγείται μπορεί να εμφανιστεί μία ή καμμία φορά. Συνεπώς, η κανονική έκφραση `ca?t` περιγράφει τις συμβολοσειρές `ct` και `cat`.

Μία μονάδα μπορεί να είναι ένας χαρακτήρας, όπως στα παραδείγματά μας, ή μπορεί να είναι μια ομάδα χαρακτήρων, αν τους περικλύσουμε σε παρενθέσεις. Η κανονική έκφραση `(ba)+boom` περιγράφει τις συμβολοσειρές `baboom`, `bababoom`, `babababoom`, κ.λπ.

Έχοντας γνώση των κανονικών εκφράσεων, μπορείτε να λύσετε το σταυρόλεξο στην αρχή της εκφώνησης, αναπτύσσοντας τις κανονικές εκφράσεις που δίνονται από κάτω του όσο χρειάζεται για να γεμίζουν τα τετράγωνα του σταυρολέξου.

Μέθοδος Επίλυσης

Για να λύσουμε το σταυρόλεξο θα πρέπει να χρησιμοποιήσουμε έναν αλγόριθμο ο οποίος να δοκιμάζει πιθανές κανονικές εκφράσεις μέχρι να μπορέσει να γεμίσει όλο το σταυρόλεξο. Αυτό δεν μπορεί να γίνει τυχαία, δεδομένου ότι δεν θα είναι εφικτό να δοκιμαστούν όλες οι δυνατές συμβολοσειρές που προκύπτουν από όλες τις κανονικές εκφράσεις σε όλες τις θέσεις. Θα πρέπει να ακολουθήσετε μια πιο ευφυή διαδικασία, όπως η παρακάτω.

1. Όσο μένουν ακόμα λέξεις που δεν έχουν βρεθεί:
    2. Επέλεξε μια λέξη που δεν έχει λυθεί.
    3. Βρες τις κανονικές εκφράσεις που μπορούν να γεμίσουν τα τετράγωνα της συγκεκριμένης θέσης.

4. Για κάθε μία από τις υποψήφιες αυτές κανονικές εκφράσεις, συνέχισε αναδρομικά τη διαδικασία από το βήμα 1.

Απαιτήσεις Προγράμματος

Κάθε φοιτητής θα εργαστεί σε αποθετήριο στο GitHub. Για να αξιολογηθεί μια εργασία θα πρέπει να πληροί τις παρακάτω προϋποθέσεις:

- Για την υποβολή της εργασίας θα χρησιμοποιηθεί το ιδιωτικό αποθετήριο του φοιτητή που δημιουργήθηκε για τις ανάγκες του μαθήματος και του έχει αποδοθεί. Το αποθετήριο αυτό έχει όνομα του τύπου `username-algo-assignments`, όπου `username` είναι το όνομα του φοιτητή στο GitHub. Για παράδειγμα, το σχετικό αποθετήριο του διδάσκοντα θα ονομαζόταν `louridas-algo-assignments` και θα ήταν προσβάσιμο στο https://github.com/dmst-algorithms-course/louridas-algo-assignments. *Τυχόν άλλα αποθετήρια απλώς θα αγνοηθούν.*
- Μέσα στο αποθετήριο αυτό θα πρέπει να δημιουργηθεί ένας κατάλογος `assignment-2021-1`.
- Μέσα στον παραπάνω κατάλογο το πρόγραμμα θα πρέπει να αποθηκευτεί με το όνομα `re_crossword.py`.
- Δεν επιτρέπεται η χρήση έτοιμων βιβλιοθηκών γράφων ή τυχόν έτοιμων υλοποιήσεων των αλγορίθμων, ή τμημάτων αυτών, εκτός αν αναφέρεται ρητά ότι επιτρέπεται.
- Επιτρέπεται η χρήση δομών δεδομένων της Python όπως στοίβες, λεξικά, σύνολα, κ.λπ.
- Επιτρέπεται η χρήση της βιβλιοθήκης `re` και της βιβλιοθήκης `sre_yield`.
- Επιτρέπεται η χρήση της βιβλιοθήκης `csv`.
- Επιτρέπεται η χρήση της βιβλιοθήκης `argparse` ή της βιβλιοθήκης `sys` (συγκεκριμένα, της λίστας `sys.argv`) προκειμένου να διαβάσει το πρόγραμμα τις παραμέτρους εισόδου.
- Το πρόγραμμα θα πρέπει να είναι γραμμένο σε Python 3.

Για να λύσετε το σταυρόλεξο έχετε στη διάθεσή σας τις κανονικές εκφράσεις που θα χρησιμοποιήσετε. Αυτές θα πρέπει να τις αναπτύξετε ώστε να δείτε πού ταιριάζουν. Για να αναπτύξετε κανονικές εκφράσεις μπορείτε να χρησιμοποιήσετε τη βιβλιοθήκη `sre_yield`. Η βιβλιοθήκη αυτή σας δίνει τη συνάρτηση `sre_yield.AllStrings(pattern)` η οποία παράγει όλες τις συμβολοσειρές που αντιστοιχούν στην κανονική έκφραση `pattern`.

Το πρόγραμμα θα καλείται ως εξής (όπου `python` η κατάλληλη εντολή στο εκάστοτε σύστημα):

```
python re_crossword.py crossword_file regular_expressions_file
```

Η σημασία των παραμέτρων είναι η εξής:

- `crossword_file`: το αρχείο που περιγράφει τη δομή του σταυρολέξου. Προσοχή, δεν σημαίνει ότι το αρχείο θα ονομάζεται ντε και καλά `crossword_file`, μπορεί να έχει οποιοδήποτε όνομα.
- `regular_expressions_file`: το αρχείο που περιέχει τις κανονικές εκφράσεις που θα χρησιμοποιήσετε. Ισχύει η ίδια παρατήρηση όπως και παραπάνω για το όνομα.

Το αρχείο `crossword_file` θα είναι ένα αρχείο μορφής CSV (Comma Separated Values). Κάθε γραμμή θα αποτελείται από τα πεδία, που χωρίζονται από κόμμα:

1. Λέξη στο σταυρόλεξο.
2. Συμβολοσειρά που αναπαριστά τα περιεχόμενα της λέξης. Κάθε άγνωστος χαρακτήρας αναπαρίσταται με μία τελεία (`.`).
3. Ένα ή περισσότερα ζεύγη αριθμών. Ο πρώτος αριθμός κάθε ζεύγους δείχνει μια λέξη με την οποία τέμνεται η λέξη της γραμμής που βρισκόμαστε. Ο δεύτερος αριθμός κάθε ζεύγους δείχνει τη θέση του χαρακτήρα της τομής στην τεμνόμενη λέξη.

Για παράδειγμα, η πρώτη γραμμή του αρχείου `laughs.csv` περιγράφει τη λέξη 0 του σταυρολέξου:

```
0,......,15,0
```

Η λέξη έχει έξι άγνωστους χαρακτήρες (`......`) και τέμνει τη λέξη 15 του σταυρολέξου στη θέση 0.

Η προτελευταία γραμμή του ίδου αρχείου περιγράφει τη λέξη 13 του σταυρολέξου:

```
13,........,4,4,10,3
```

Η λέξη έχει οκτώ άγνωστους χαρακτήρες (`........`) και τέμνει τη λέξη 4 στη θέση 4 και τη λέξη 10 στη θέση 3.

Οι λέξεις θα αριθμούνται σειριακά με άρτιους αριθμούς οι οριζόντιες και περιττούς αριθμούς η κάθετες, όπως μπορείτε να δείτε στην αρχική εικόνα της εκφώνησης. Σε περίπτωση που ένα τετράγωνο είναι κοινή αρχή κοινή αρχή δύο λέξεων, το τετράγωνο θα περιέχει και τους δύο αριθμούς, όπως φαίνεται για το τετράγωνο με την επισημείωση 2/1.

Η έξοδος του προγράμματος θα είναι μια σειρά γραμμών της μορφής:

```
X regex word
```

όπου `X` ο αριθμός της λέξης, `regex` η κανονική έκφραση που ταιριάζει και `word` η λέξη που προκύπτει από το ανάπτυγμα της κανονικής έκφρασης. οι γραμμές θα πρέπει να είναι ταξινομημένες σε αύξουσα σειρά ως προς τη λέξη.

Λεπτομέρειες Υλοποίησης

- Στο βήμα 2 του αλγορίθμου που θα φτιάξετε θα πρέπει να επιλέγετε μια λέξη που δεν έχει βρεθεί ακόμα. Η σωστή επιλογή της εκάστοτε λέξης μπορεί να έχει πολύ μεγάλη επίπτωση στον αριθμό των δοκιμών που θα γίνουν και άρα στην ταχύτητα του προγράμματός σας. Μπορείτε να δοκιμάσετε ως στρατηγική να επιλέγετε τη λέξη με το μεγαλύτερο λόγο γνωστών γραμμάτων προς το μήκος της λέξης.
- Η συνάρτηση `sre_yield.AllStrings(pattern)` μπορεί, αναλόγως την κανονική έκφραση, να επιστρέψει πάρα πολύ μεγάλο αριθμό παραγόμενων συμβολοσειρών. Για να περιορίσουμε τον αριθμό αυτό μπορούμε να χρησιμοποιήσουμε την παράμετρο `max_count`. Για τις ανάγκες της εργασίας το `max_count=5`, δηλαδή `sre_yield.AllStrings(pattern, max_count=5)`, είναι αρκετό.
- Η συνάρτηση `sre_yield.AllStrings(pattern)` μπορεί να επιστρέψει την ίδια λέξη περισσότερες από μία φορές, οπότε θα πρέπει να λάβετε πρόνοια να παίρνετε μόνο μία φορά την κάθε μία.

Παραδείγματα

Παράδειγμα 1

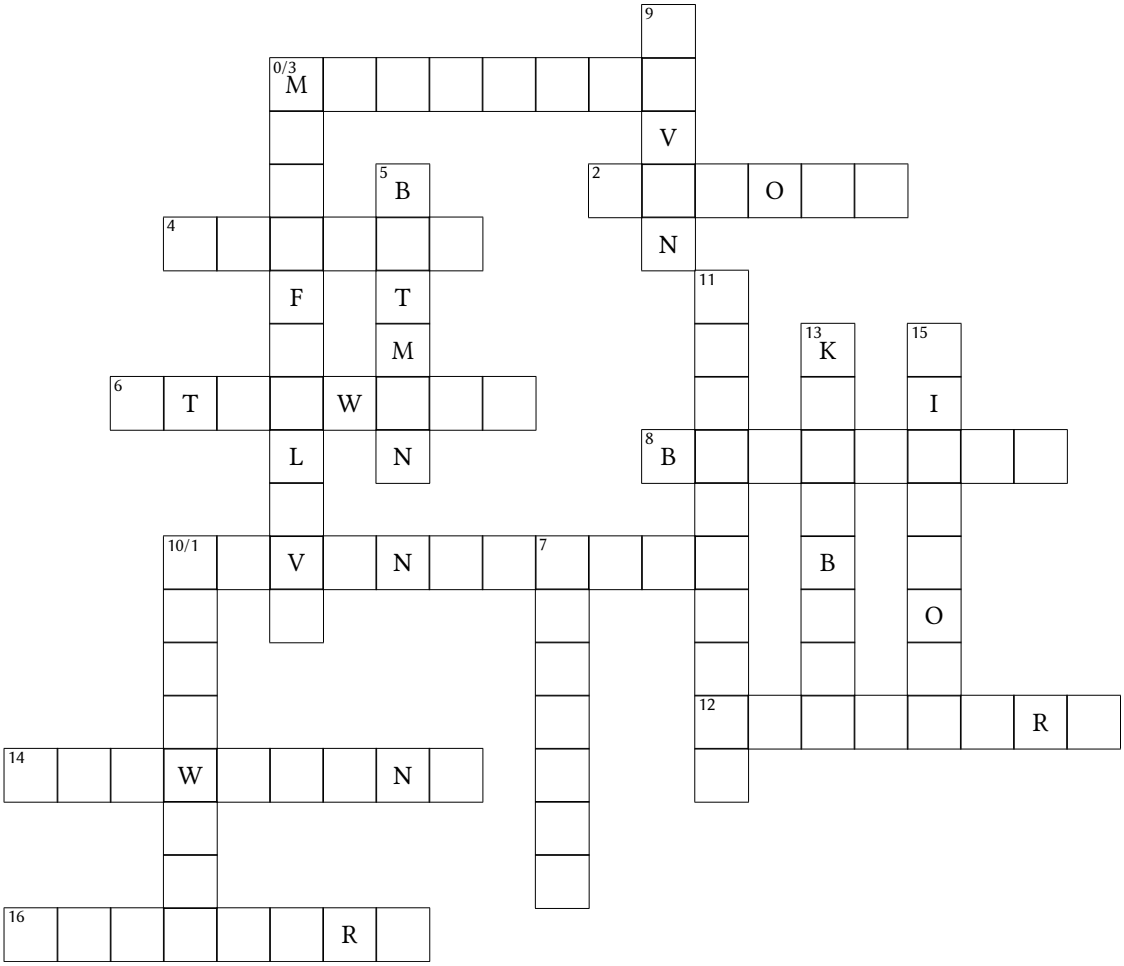
Αν ο χρήστης του προγράμματος δώσει:

```
python re_crossword.py laughs.csv laughs.txt
```

τότε το πρόγραμμα θα διαβάσει το αρχείο `laughs.csv` που περιγράφει το σταυρόλεξο και το αρχείο `laughs.txt` που περιγράφει τις κανονικές εκφράσεις και θα πρέπει να εμφανίσει στην οθόνη *ακριβώς* τα παρακάτω (και τίποτε παραπάνω):

```
0 HAHA* HAHAAA
1 HE(HE)+ HEHEHE
2 (HO)+ HOHO
3 HA+ HAA
4 TE(HE+)+ TEHEHEHE
5 LO+L LOOL
6 L(OL)+ LOLOLOL
7 LULZ LULZ
8 K(EK)*E KEKE
9 ROT?FL ROTFL
10 MWA(HA)+ MWAHA
11 LAW*L LAWL
12 H(EH)+ HEH
13 (HAR+)+ HARRHARR
14 (JA)+ JAJAJAJAJA
15 (AH)+A+ AHAHA
```

Παράδειγμα 2



Αν ο χρήστης του προγράμματος δώσει:

```
python re_crossword.py films.csv films.txt
```

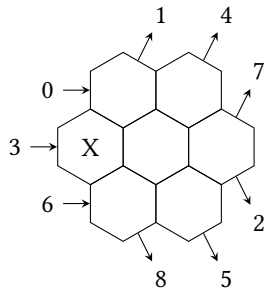
τότε το πρόγραμμα θα διαβάσει το αρχείο `films.csv` που περιγράφει το σταυρόλεξο και το αρχείο `films.txt` που περιγράφει τις κανονικές εκφράσεις και θα πρέπει να εμφανίσει στην οθόνη *ακριβώς* τα παρακάτω (και τίποτε παραπάνω):

```
0 (M?ON)+CLE MONONCLE
1 SIDEWAYS SIDEWAYS
2 (((OU)|(GE))T)+ GETOUT
3 MO+D([FL]O[RV])+E MOODFORLOVE
4 (MAD?)+X MADMAX
5 ([BM]A[NT])+ BATMAN
```

```
6 S([TW]AR)+S STARWARS
7 SH(IN)+G SHINING
8 ([BL](AN?D))+S BADLANDS
9 S((EN)|(EV))+ SEVEN
10 S((EN)|(EV))+THSEAL SEVENTHSEAL
11 CA((S|BL|NC)A)+ CASABLANCA
12 ([CG]A[LR]I)+ CALIGARI
13 ([KB](ILL))+ KILLBILL
14 (SHA[WN])+K SHAWSHANK
15 (K[IO]NG)+ KINGKONG
16 (S?TOR?Y)+ TOYSTORY
```

Παράδειγμα 3

Τα σταυρόλεξά μας μπορεί να είναι και λίγο πιο παράξενα, αν επιτρέψουμε κάθε θέση να μπορεί να τέμνεται με *δύο* άλλες, άρα αντί για τετραγωνάκια να έχουμε εξάγωνα, όπως στο παρακάτω:



Στο σταυρόλεξο αυτό δεν έχουμε πλέον οριζόντια και κάθετα, έχουμε μια διάσταση x, η οποία παίρνει τιμές τους αριθμούς που το υπόλοιπό τους με το 3 είναι 0, μια διάσταση y, η οποία παίρνει τιμές τους αριθμούς που το υπόλοιπό τους με το 3 είναι 1, και μια διάσταση z, η οποία παίρνει τιμές τους αριθμούς που το υπόλοιπό τους με το 3 είναι το 2. Αν αυτό σας φαίνεται παράξενο σκεφτείτε ότι και στα προηγούμενα παραδείγματα τον ίδιο κανόνα χρησιμοποιήσαμε, μόνο που παίρναμε το υπόλοιπο με το 2.

Αν ο χρήστης του προγράμματος δώσει:

```
python re_crossword.py hex.csv hex.txt
```

τότε το πρόγραμμα θα διαβάσει το αρχείο `hex.csv` που περιγράφει το σταυρόλεξο και το αρχείο `hex.txt` που περιγράφει τις κανονικές εκφράσεις και θα πρέπει να εμφανίσει στην οθόνη *ακριβώς* τα παρακάτω (και τίποτε παραπάνω):

```
0 HE|HA|OH HE
1 NP|HX|SP HX
2 FN|EG|GN EG
3 .?[SAX][GAS].? XAG
4 .?[FAT][HOT].? EAO
```

```
5 .?[AND][NOT].? HAN
6 NO|ON|AT ON
7 XO|GN|PO GN
8 RO|GN|XO XO
```

Καλή Επιτυχία.

Για Περισσότερες Πληροφορίες

Το σταυρόλεξο στην αρχή της εκφώνησης είναι από το βιβλίο του Alex Bellos *The Language Lover's PuzzleBook: Lexical Perplexities and Cracking Conundrums from Across the Globe*, Guardian Faber Publishing, 2020.