Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων

Εξέταση εργαστηρίου Python, 19/11/2020, Νίκος Παπασπύρου.

Πρόβλημα "contact-tracking"

Σας ζητείται να επιστρατεύσετε τις ικανότητές σας στην επεξεργασία δεδομένων για να βοηθήσετε στην καταγραφή κρουσμάτων του κορονωϊού, μέσω της ιχνηλάτησης των επαφών των φορέων του.

Οι άνθρωποι σήμερα δεν πάνε πουθενά χωρίς τα κινητά τους τηλέφωνα. Και τα περισσότερα από αυτά έχουν GPS και ανοιχτή την υπηρεσία τοποθεσίας (geo-location), με αποτέλεσμα οι πάροχοι υπηρεσιών κινητής τηλεφωνίας να γνωρίζουν ανά πάσα στιγμή πού βρισκόμαστε. Ο εθνικός οργανισμός δημόσιας υγείας ζήτησε λοιπόν τη βοήθεια των εταιριών κινητής τηλεφωνίας και έχει τώρα στη διάθεσή του ένα τεράστιο αρχείο δεδομένων που περιέχει μία γραμμή για κάθε καταγεγραμμένη "επαφή" δύο κατόχων κινητών τηλεφώνων.

Σας δίνεται αυτό το αρχείο σε μορφή CSV (βλ. παράδειγμα). Θα έχει τρεις στήλες. Η πρώτη στήλη περιέχει την ημερομηνία και ώρα (YYYY-MM-DD hh:mm:ss) μίας επαφής, ενώ οι δύο άλλες θα έχουν τους αριθμούς των κινητών τηλεφώνων που έρχονται σε επαφή. Οι γραμμές του αρχείου είναι σε τυχαία σειρά. Ας υποθέσουμε ότι τα δεδομένα επαφών αφορούν το διάστημα από την ημέρα d_s μέχρι την ημέρα d_e.

```
$ head -5 dataset-20.csv
When,Phone1,Phone2
2020-01-06 10:21:21,6700000008,6700000009
2020-01-05 06:03:16,6700000000,6700000007
2020-01-09 14:54:03,6700000001,6700000000
2020-01-03 07:58:18,6700000000,6700000005
```

Ζητείται να γράψετε ένα πρόγραμμα σε Python που να δέχεται στο command line τέσσερις παραμέτρους, κατά σειρά:

- το όνομα του αρχείου CSV με τα δεδομένα
- τον αριθμό τηλεφώνου του πρώτου κρούσματος
- τον αριθμό ημερών επώασης του ιού (έστω X) — θα είναι $0 \leq X$
- τον αριθμό ημερών ανάρρωσης από τον ιό (έστω Y) — θα είναι $X \leq Y$

Υποθέσεις:

- Κάθε κινητό τηλέφωνο αντιστοιχεί σε διαφορετικό άνθρωπο.
- Θεωρούμε ότι το πρώτο κρούσμα έρχεται σε επαφή με τον ιό αμέσως πριν την αρχή της ημέρας d_s.
- Αν κάποιος εκτεθεί στον ιό την ημέρα d, τότε αρχίζει να τον μεταδίδει στην αρχή της ημέρας $d + X$ και σταματά να τον μεταδίδει στο τέλος της ημέρας $d + Y$. (Στην ειδική περίπτωση που $X = 0$, η μετάδοση του ιού κατά την πρώτη μέρα δεν μπορεί φυσικά να προηγείται της έκθεσης σε αυτόν.)
- Κάθε άνθρωπος μπορεί να κολλήσει τον ιό το πολύ μία φορά.
- "Νέο κρούσμα" (case) κατά τη διάρκεια της ημέρας d ονομάζεται ένας άνθρωπος που ήταν υγιής την ημέρα $d - 1$ και εκτίθεται για πρώτη φορά στον ιό κατά τη διάρκεια της ημέρας d.

Το πρόγραμμά σας πρέπει να υπολογίζει:

1. Πόσοι άνθρωποι συμμετέχουν στο αρχείο καταγραφής.
2. Το μέγιστο, το ελάχιστο και το μέσο πλήθος επαφών ανά ημέρα, όπως προκύπτουν από το αρχείο καταγραφής. Οι πρώτες δύο τιμές είναι ακέραιοι αριθμοί και η τελευταία ζητείται με στρογγυλοποίηση σε 3 δεκαδικά ψηφία.
3. Το συνολικό πλήθος ανθρώπων που έχουν εκτεθεί στον ιό μετά την τελευταία ημέρα καταγραφής d_e.
4. Το μέγιστο πλήθος νέων κρουσμάτων σε μια μέρα.
5. Δύο διαγράμματα που να απεικονίζουν το πλήθος νέων κρουσμάτων (cases) ανά ημέρα και το πλήθος των μολυσμένων ανθρώπων (infected) ανά ημέρα.

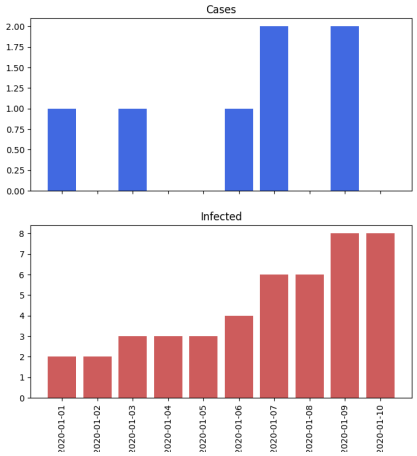
Ακολουθούν μερικά παραδείγματα χρήσης του προγράμματος. Τα ζητούμενα (1)-(4) πρέπει να εκτυπώνονται στην οθόνη **ακριβώς** στην παρακάτω μορφή. Τα ζητούμενα διαγράμματα (5) πρέπει να αποθηκεύονται σε ένα αρχείο `diagrams.png` που να μοιάζει με αυτό των Σχημάτων 1, 2 και 3.

```
$ ./contact-tracking.py dataset-20.csv 6700000001 0 7
people = 10
max contacts per day = 4
min contacts per day = 1
average contacts per day = 2.000
total infected = 8
max cases per day = 2
```

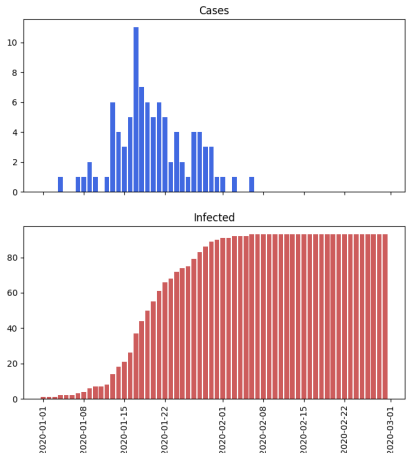
```
$ ./contact-tracking.py dataset-1000.csv 6700000001 0 7
people = 100
max contacts per day = 28
min contacts per day = 8
average contacts per day = 16.667
total infected = 93
max cases per day = 11

$ ./contact-tracking.py dataset-1000.csv 6700000042 3 10
people = 100
max contacts per day = 28
min contacts per day = 8
average contacts per day = 16.667
total infected = 92
max cases per day = 7
```

Το πρόγραμμά σας θα εκτελεστεί με πολλά διαφορετικά αρχεία CSV, το μεγαλύτερο από τα οποία θα περιέχει 10^6 γραμμές καταγεγραμμένων επαφών.

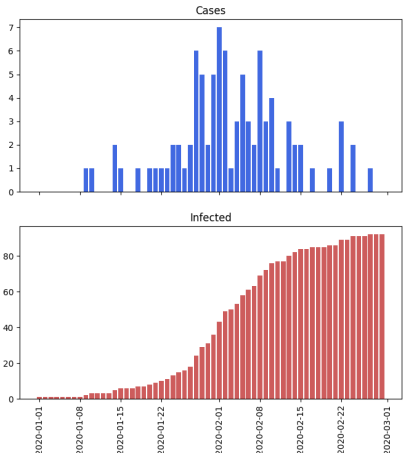
Προσοχή: Δείτε στην ιστοσελίδα του μαθήματος τι πρέπει να παραδώσετε και πώς!



Σχήμα 1: Διαγράμματα για `dataset-20.csv 6700000001 0 7`



Σχήμα 2: Διαγράμματα για `dataset-1000.csv 6700000001 0 7`



Σχήμα 3: Διαγράμματα για `dataset-1000.csv 6700000042 3 10`

Καλή επιτυχία!