Πρώτη Σειρά Ασκήσεων

Αυτή είναι η πρώτη σειρά ασκήσεων. Η προθεσμία για την παράδοση είναι στις 21 Νοεμβρίου 11:59 μ.μ. Κάνετε turn-in τον κώδικα σας και τα αποτελέσματα σας, και παραδώστε τις υπόλοιπες ερωτήσεις είτε ηλεκτρονικά, είτε σε χαρτί. Όπου ζητείται αναφορά θα πρέπει να είναι γραμμένη ηλεκτρονικά. Σε κάποιες περιπτώσεις η αναφορά θα είναι ενσωματωμένη στο notebook με τους υπολογισμούς. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Λεπτομέρειες για το turn-in, και για το πώς να γράφετε αναφορές είναι στη σελίδα Ασκήσεις του μαθήματος. Οι ασκήσεις θα πρέπει να γίνουν **ατομικά**.

Ερώτηση 1

Α. Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο Reservoir Sampling ώστε να κάνει δειγματοληψία $\boldsymbol{K}$ αντικειμένων από ένα ρεύμα N αντικειμένων ομοιόμορφα τυχαία, ώστε το κάθε αντικείμενο να έχει πιθανότητα K/N να εμφανιστεί στο δείγμα. Ο αλγόριθμος σας θα πρέπει να δουλεύει με ένα μόνο πέρασμα στα δεδομένα διαβάζοντας τα αντικείμενα ένα-ένα, χωρίς προηγούμενη γνώση του μεγέθους του ρεύματος, και να χρησιμοποιεί $O(K)$ μνήμη (υποθέστε ότι το μέγεθος του κάθε αντικειμένου είναι σταθερό).

1. Περιγράψετε τον αλγόριθμο που διαλέγει ένα ομοιόμορφο δείγμα K αντικειμένων από ένα ρεύμα N αντικειμένων. Η περιγραφή του αλγορίθμου **δεν** πρέπει να είναι σε κώδικα ή ψευδοκώδικα, αλλά να εξηγεί τη λογική του αλγορίθμου στα Ελληνικά με απλό τρόπο.
2. Αποδείξτε ότι ο αλγόριθμος σας παράγει ένα ομοιόμορφα τυχαίο δείγμα, δηλαδή, για κάθε $i, 1 \leq i \leq N$, το i -οστό στοιχείο έχει πιθανότητα K/N να εμφανιστεί στο δείγμα.
3. Γράψτε ένα πρόγραμμα **σε Python** που υλοποιεί τον αλγόριθμο σας. Το πρόγραμμα σας θα πρέπει να παράγει ένα δείγμα με K τυχαίες γραμμές από ένα αρχείο κειμένου. Θα πρέπει να μπορούμε να τρέξουμε το πρόγραμμα από την γραμμή εντολών, θα παίρνει σαν όρισμα εντολής την τιμή του K, θα διαβάζει γραμμές από το standard input και θα εκτυπώνει το δείγμα στο standard output. Για παράδειγμα, η παρακάτω εντολή θα πρέπει να τυπώνει στην οθόνη ένα τυχαίο δείγμα 10 γραμμών από το αρχείο input.txt:
   “`sample.py 10 < input.txt`”.

Β. Περιγράψετε ένα αλγόριθμο που κάνει δειγματοληψία ενός τυχαίου κόμβου από ένα δέντρο για το οποίο δεν ξέρετε εκ των προτέρων τον αριθμό των κόμβων που έχει, και δεν μπορείτε να το αποθηκεύσετε στην μνήμη.

Ερώτηση 2

Έστω $x_1, x_2, \dots, x_n$, n πραγματικές τιμές, ταξινομημένες σε αύξουσα σειρά, και υποθέστε ότι $n = 2k + 1$ είναι ένας περιττός αριθμός. Δείξετε ότι ο πραγματικός αριθμός z που ελαχιστοποιεί το άθροισμα των αποστάσεων

$$\sum_{i=1}^{n} |x_i - z|$$

είναι ο μέσος x_{k+1}.

Υπόδειξη: Τοποθετήστε το z στον μέσο και εξετάσετε τι γίνεται αν το μετακινήσετε προς τα αριστερά ή δεξιά.

Ερώτηση 3

Σας δίνεται ο παρακάτω πίνακας προτιμήσεων για χρήστες και ταινίες, με τις βαθμολογίες που έχουν δώσει οι χρήστες για κάθε ταινία.

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6
User X	5	3		1		?
User Y	4	2	1			1
User Z	3			1	3	3
User W	2	5	1	5	3	4

Στόχος μας είναι να προβλέψουμε την τιμή (X,6) του User X για το Movie 6.

Θα χρησιμοποιήσετε τον User-User Collaborative Filtering αλγόριθμο, σε δύο εκδοχές.
Στην πρώτη εκδοχή θα χρησιμοποιήσετε τον πίνακα όπως είναι, θα υπολογίσετε το cosine similarity μεταξύ του X και των υπολοίπων χρηστών, και θα υπολογίσετε την βαθμολογία για το κελί (X,6) ως τον σταθμισμένο μέσο όρο της βαθμολογίας των δυο πιο όμοιων χρηστών στον X.
Στην δεύτερη εκδοχή πρώτα θα αφαιρέσετε την μέση τιμή των βαθμολογιών από κάθε γραμμή, θα υπολογίσετε και πάλι το cosine similarity του X με τους υπολοίπους χρήστες, και τώρα θα υπολογίσετε την απόκλιση από τον μέσο όρο για το κελί (X,6) ως τον σταθμισμένο μέσο όρο της απόκλισης των δυο πιο όμοιων χρηστών στον X. Υπολογίσετε την βαθμολογία προσθέτοντας την απόκλιση που υπολογίσατε στην μέση τιμή του X.
Μπορείτε να κάνετε τους υπολογισμούς προγραμματιστικά ή με το χέρι. Αναφέρετε τα αποτελέσματα (ομοιότητες, εκτιμώμενη βαθμολογία και απόκλιση), και για τις δύο εκδοχές στην αναφορά σας.

Τι παρατηρείτε? Ποια κελιά έχουν μικρότερη ή μεγαλύτερη σημασία στον υπολογισμό της ομοιότητας σε κάθε εκδοχή? Πως αλλάζουν οι πιο κοντινοί γείτονες? Τι βαθμολογία υπολογίζετε σε κάθε περίπτωση? Εξηγήστε τις παρατηρήσεις σας.

Ερώτηση 4

Από το site του βιβλίου The Data Science Manual, από τη σελίδα με τα δεδομένα, κατεβάσετε τα Movie Data. Τα δεδομένα είναι σε μορφή csv αρχείου, όπου κάθε γραμμή είναι μία ταινία οι στήλες attributes που περιγράφονται στο Description link. Υπάρχουν δεδομένα 3201 ταινίες Ο στόχος είναι να κάνετε ανάλυση αυτών των δεδομένων, για να κατανοήσετε την μορφή τους, να διατυπώσουμε κάποια ερωτήματα και να προσπαθήσουμε να τα απαντήσουμε από τα δεδομένα.

Φορτώσετε τα δεδομένα σε ένα Pandas Dataframe. Τα δεδομένα, όπως και όλα τα πραγματικά δεδομένα είναι ημιτελή και έχουν θόρυβο, οπότε θα πρέπει να κάνετε κάποιο καθάρισμα όπου χρειάζεται. Υπάρχουν πολλοί τρόποι για να χειριστείτε αυτά τα προβλήματα, κάνετε σαφείς τις επιλογές σας, και γράψετε ξεκάθαρα τι επίδραση έχουν στα δεδομένα (π.χ., αν πετάξετε κάποια δεδομένα, αναφέρετε τα κριτήρια και πόσα δεδομένα μένουν, αν αντικαταστήσετε τιμές που λείπουν, πείτε με τι αντικαταστήσατε και τι επίδραση έχει αυτό). Καλό θα είναι επίσης να αλλάξετε και τα ονόματα των στηλών ώστε να σας βολεύουν καλύτερα.

Θα ασχοληθείτε με τα εξής ερωτήματα:

1. Δημιουργήστε ιστογράμματα για τα attributes “Worldwide Gross” (τα έσοδα της ταινίας σε όλο τον κόσμο), “Rotten Tomatoes Rating”, “IMDB Rating”, “IMDB votes” (προσοχή γιατί το “Worldwide Gross” αποθηκεύεται ως string). Τι παρατηρείτε? Πως φαίνονται οι κατανομές των τιμών?
   Για το attribute “Major Genre” υπολογίσετε για κάθε genre τον αριθμό των ταινιών με αυτό το genre. Φτιάξετε ένα πίνακα και ένα bar plot που να δείχνει την κατανομή. Τι παρατηρείτε για την δημοφιλία των genres?
   Για τα παραπάνω χρησιμοποιείστε τις λειτουργίες της βιβλιοθήκης Pandas.
   Στη συνέχεια, για τα attributes “Worldwide Gross” και “IMDB Votes” θα δημιουργήστε εσείς ένα ιστόγραμμα και θα κάνετε ένα plot σε log-log κλίμακα. Για το ιστόγραμμα σας, θα φτιάξετε bins τα οποία να αυξάνουν σε μέγεθος εκθετικά. Για το plot στον x άξονα θα έχετε την μέση τιμή του διαστήματος του bin (έσοδα ή ψήφοι), και στον y άξονα τον αριθμό των ταινιών που πέφτουν σε αυτό το bin. Κάνετε ένα scatter plot με τις τιμές, με λογαριθμική κλίμακα και στους δύο άξονες. Τι παρατηρείτε? Τι κατανομή φαίνεται να ακολουθούν τα έσοδα και ο αριθμός των IMDB votes?
2. Μετά την ανάλυση των ιστογραμμάτων ο στόχος μας είναι να βρούμε αν υπάρχουν συσχετίσεις μεταξύ των attributes. Καταρχάς θα εξετάσουμε τα αριθμητικά attributes “Worldwide Gross”, “Rotten Tomatoes Rating”, “IMDB Rating”, “IMDB votes”. Το πρώτο αντιπροσωπεύει την δημοφιλία της ταινίας στα σινεμά, ενώ το τελευταίο την δημοφιλία online. Τα άλλα δύο αντιπροσωπεύουν την ποιότητα της ταινίας όπως την κρίνουν οι ειδικοί και το κοινό. Θα εξετάσετε ανα δύο τα attributes αν υπάρχει συσχέτιση. Παράγετε ένα scatter plot για κάθε ζεύγος (τα Pandas σας επιτρέπουν να τα βάλετε όλα μαζί σε ένα σχήμα αν θέλετε), και υπολογίστε τα Pearson και Spearman correlation coefficients, καθώς και τα p-values. Τι συσχετίσεις βλέπετε και πόσο ισχυρές είναι? Τι παρατηρείτε στα τα scatter plot με το “World Gross”? Πως μπορείτε να κάνετε οπτικά πιο κατανοητά?
   Στη συνέχεια θα εξετάσετε αν υπάρχει συσχέτιση μεταξύ του genre και των εισόδων της ταινίας. Μπορείτε να αγνοήσετε πολύ σπάνια genres, αλλά κάνετε σαφές τις επιλογές σας στην αναφορά σας. Δημιουργήστε ένα bar plot (με confidence intervals) με το μέσο worldwide gross για κάθε genre.

Χρησιμοποιείστε το t-test για να εξετάσετε ποιες από τις διαφορές που παρατηρείτε είναι στατιστικά σημαντικές. Τι συμπέρασμα βγάζετε?

3. Συχνά λέγεται ότι «Δεν κάνουν πια καλές ταινίες όπως παλιά». Χρησιμοποιείστε τα attributes Release date, Rotten Tomatoes Rating και IMDB Rating για να εξετάσετε αν αυτό ισχύει. Δημιουργήστε ένα γράφημα με την μέση ποιότητα (όπως αυτή μετριέται με το Rotten Tomatoes Rating και το IMDB Rating) και κάνετε γραφήματα για το πώς αλλάζουν σε διαφορετικές δεκαετίες (μπορείτε να αγνοήσετε δεκαετίες με πολύ λίγα δεδομένα). Τι παρατηρείτε? Ποιες ειναι κάποιες πιθανές εξηγήσεις για τα γραφήματα που παρατηρείτε?
4. Διατυπώστε ένα δικό σας ερώτημα και εξερευνήσετε το με τα δεδομένα. Γράψετε με σαφήνεια ποιο ερώτημα εξετάζετε, τι μετρήσεις θα κάνετε για να το απαντήσετε, παρουσιάσετε γραφήματα και νούμερα που δείχνουν την ανάλυση που κάνατε και υποστηρίζουν την απάντηση σας (η οποία ενδεχομένως να είναι αρνητική, αλλά και σε αυτή την περίπτωση θα πρέπει να υπάρχουν τα στοιχεία που να το δείχνουν), και σχολιάσετε τα αποτελέσματα.

Παραδώσετε ένα notebook με όλους τους υπολογισμούς, τα γραφήματα και το κείμενο της αναφοράς. Αν θέλετε μπορείτε να βάλετε την αναφορά σε ξεχωριστό κείμενο, αλλά όποτε χρειάζονται γραφήματα για να δείξετε κάτι, θα πρέπει να είναι μαζί με το κείμενο.

Ερώτηση 5 (bonus)

Τα Google trends είναι μια υπηρεσία της Google που μας επιτρέπει να εξετάσουμε την συχνότητα ερωτημάτων στην μηχανή αναζήτησης σε διαφορετικές στιγμές στον χρόνο (και στον χώρο για κάποιες χώρες). Π.χ., σε αυτό το blog post, χρησιμοποιούν τα Google trends για να εξετάσουν την ώρα και μέρα που ο κόσμος αποφασίζει να χωρίσει (επίσης το blog post έχει πληροφορίες για την βιβλιοθήκη Pytrends που σας επιτρέπει να τραβάτε δεδομένα). Διατυπώσετε ένα δικό σας ερώτημα, και προσπαθήσετε να το απαντήσετε χρησιμοποιώντας δεδομένα από το Google Trends. Το ερώτημα μπορεί να έχει σχέση με οτιδήποτε (π.χ., πολιτική, αθλητικά, τηλεόραση, καθημερινότητα, κλπ). Στην αναφορά σας γράψετε με σαφήνεια ποιο ερώτημα εξετάζετε, τι μετρήσεις θα κάνετε για να το απαντήσετε, παρουσιάσετε γραφήματα και νούμερα που δείχνουν την ανάλυση που κάνατε και υποστηρίζουν την απάντηση σας (η οποία ενδεχομένως να είναι αρνητική, αλλά και σε αυτή την περίπτωση θα πρέπει να υπάρχουν τα στοιχεία που να το δείχνουν), και σχολιάσετε τα αποτελέσματα.

Για την ανάλυση σας μπορείτε να χρησιμοποιήσετε κάποια βιβλιοθήκη όπως η Pytrends, ή να κάνετε τα ερωτήματα με το χέρι και να τραβήξετε τα δεδομένα. Περιγράψετε τι κάνατε στην αναφορά σας. Παραδώσετε την αναφορά, και τον κώδικα και τα δεδομένα.