Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Χειμερινό Εξάμηνο, 2020

Πρώτη Σειρά Ασκήσεων

Αυτή είναι η πρώτη σειρά ασκήσεων. Η προθεσμία για την παράδοση είναι στις 1 Δεκεμβρίου 11:59 μ.μ. Κάνετε turn-in τον κώδικα σας και τα αποτελέσματα σας. Για τις αποδείξεις συνίσταται θερμά να παραδώσετε κάποιο pdf με το κείμενο σας, αλλά αν δεν γίνεται, μπορείτε να παραδώσετε φωτογραφίες των γραπτών απαντήσεων σας. 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. Αν χρησιμοποιήσετε free passes θα πρέπει να το αναφέρετε στην αναφορά. Θα αφαιρεθούν από όλα τα μέλη της ομάδας. Λεπτομέρειες για το turn-in, και για το πώς να γράφετε αναφορές είναι στη σελίδα Ασκήσεις του μαθήματος. Οι ασκήσεις μπορούν να γίνουν σε ομάδες μέχρι δύο ατόμων.

Ερώτηση 1

Α. Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο Reservoir Sampling ώστε να κάνει δειγματοληψία $\boldsymbol{K}$ αντικειμένων ομοιόμορφα τυχαία από ένα ρεύμα N αντικειμένων, 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, ούτε να είναι η περιγραφή του κώδικα σε φυσική γλώσσα. Θα πρέπει να εξηγεί τη λογική του αλγορίθμου 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 μία συνάρτηση **sample σε Python** που υλοποιεί τον αλγόριθμο σας. Η συνάρτηση σας θα πρέπει να παίρνει σαν όρισμα το όνομα ενός αρχείου, και τον αριθμό K και να επιστρέφει ένα δείγμα με K τυχαίες γραμμές από το αρχείο. Χρησιμοποιήστε την συνάρτηση σας μέσα σε ένα πρόγραμμα και εκτυπώστε 10 τυχαίες γραμμές από ένα αρχείο που θα δώσετε σαν είσοδο. Μπορείτε να παρδώσετε είτε το πρόγραμμα σας σε python, είτε ένα notebook.

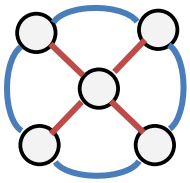
Β. Στην τάξη περιγράψαμε τον αλγόριθμο Reservoir Sampling για τη δειγματοληψία ενός αντικειμένου από ένα ρεύμα αντικειμένων. Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο ώστε να κάνει **σταθμισμένη δειγματοληψία**. Υποθέτουμε ότι το κάθε αντικείμενο i έχει βάρος w_i. Θα τροποποιήσετε τον αλγόριθμο δειγματοληψίας ώστε από ένα ρεύμα αντικειμένων με βάρη, να επιλέγει ένα αντικείμενο με πιθανότητα ανάλογη του βάρους του αντικειμένου. Δηλαδή, αν τελικά το ρεύμα έχει N αντικείμενα, και το συνολικό βάρος τους είναι $W = \sum_{i=1}^{N} w_i$ το αντικείμενο i θα πρέπει να έχει πιθανότητα w_i/W να επιλεγεί, για κάθε $1 \leq i \leq N$. Όπως και με τον κλασικό Reservoir Sampling αλγόριθμο, το N δεν είναι γνωστό εκ των προτέρων και ο αλγόριθμος θα πρέπει να δουλεύει με σταθερό χώρο μνήμης, ανεξάρτητο του N. Αποδείξτε την ορθότητα του αλγορίθμου σας.

Ερώτηση 2

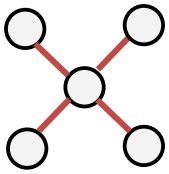
Θέλουμε να λύσουμε το παρακάτω πρόβλημα ανάθεσης: Έχουμε N αντικείμενα τα οποία κατοικούν σε ένα μετρικό χώρο (metric space). Δηλαδή, ορίζεται μια απόσταση μεταξύ τους που είναι μετρική. Έχουμε επίσης και μια τοπολογική διάταξη που ορίζεται από ένα γράφημα. Θέλουμε να αναθέσουμε αντικείμενα στις κορυφές του γραφήματος, ώστε να ελαχιστοποιήσουμε το άθροισμα των αποστάσεων μεταξύ των γειτονικών αντικειμένων στην διάταξη.

Μαθηματικά, έχουμε N αντικείμενα και ένα $N \times N$ πίνακα d με τις αποστάσεις των αντικειμένων (οι οποίες ικανοποιούν τις συνθήκες της μετρικής). Έχουμε επίσης ένα γράφημα $G = (V, E)$. Θέλουμε να βρούμε μια ερριπτική συνάρτηση $f: V \to A$ η οποία αναθέτει σε κάθε κόμβο v ένα αντικείμενο $f(v)$, διαφορετικό για κάθε κόμβο, έτσι ώστε να ελαχιστοποιεί το άθροισμα $C(f) = \sum_{(u,v)\in E} d(f(u), f(v))$. Η συνάρτηση f είναι η ανάθεση, και το $C(f)$ είναι το κόστος της ανάθεσης. Στην γενική περίπτωση το πρόβλημα είναι NP-complete. Μπορούμε όμως να βρούμε την βέλτιστη λύση αποτελεσματικά για την περίπτωση που το γράφημα G είναι δέντρο.

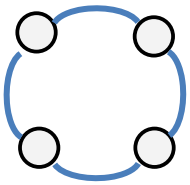
Θεωρείστε την ειδική περίπτωση που το γράφημα είναι μια "ρόδα" $W = (V, E_W)$, μεγέθους k. Το γράφημα W μπορούμε να το σκεφτούμε ως την ένωση δύο γραφημάτων: του γραφήματος $S = (V, E_S)$, που είναι ένα αστέρι με k ακτίνες, και του γραφήματος $R = (V, E_R)$ που είναι ένα κύκλος με k κόμβους. Οπότε $W = (V, E_s \cup E_R)$. Στο σχήμα φαινεται μια ρόδα μεγέθους 4.



Το γράφημα ρόδα W μεγέθους 4



Το αστέρι S



Ο κύκλος R

Αποδείξτε τα παρακάτω:

1. Έστω $f: V \to A$ μια ανάθεση στους κόμβους του W. Έστω $C_S(f) = \sum_{(u,v)\in E_S} d(f(u), f(v))$ το κόστος της ανάθεσης για το αστέρι S, και $C_W(f) = \sum_{(u,v)\in E_S} d(f(u), f(v))$ για όλο το γράφημα. Δείξτε ότι $C_W(f) \leq 3C_S(f)$.
2. Έστω $t: V \to A$ η βέλτιστη ανάθεση για το αστέρι S (η ανάθεση με το μικρότερο κόστος) και $o: V \to A$ η βέλτιστη ανάθεση για το γράφημα W. Δείξτε ότι $C_W(t) \leq 3C_W(o)$.

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.

1. 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.
2. Στην δεύτερη εκδοχή πρώτα θα αφαιρέσετε την μέση τιμή των βαθμολογιών από κάθε γραμμή (θα χρησιμοποιήσετε μόνο τις μη μηδενικές τιμές), 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

Σε αυτή την ερώτηση θα μελετήσουμε δεδομένα για την πρόσφατη πανδημία του Covid-19. Ο στόχος είναι να δούμε αν υπάρχουν κάποιες συσχετίσεις μεταξύ του αριθμού των κρουσμάτων και θυμάτων με χαρακτηριστικά των περιοχών. Η ερώτηση αποτελείται από τα παρακάτω κομμάτια.

Α. Από το site ourworldindata.org κατεβάσετε τα δεδομένα για τα καθημερινά κρούσματα του ιού. Κατεβάσετε τα δεδομένα από εδώ, σε csv, xls, ή json μορφή και φορτώστε τα σε ένα Pandas DataFrame. Καταρχάς θα

κοιτάξουμε την κατάσταση σε μια συγκεκριμένη χρονική στιγμή, την 1/11/2020, και θα μελετήσουμε τους εξής δείκτες για την πανδημία: τον συνολικό αριθμό των κρουσμάτων ανά εκατομμύριο (total_cases_per_million), τον συνολικό αριθμό των θανάτων ανά εκατομμύριο (total_deaths_per_million), και το ποσοστό θνησιμότητας (συνολικός αριθμός θανάτων προς συνολικό αριθμό κρουσμάτων).

Το πρώτο που θα εξετάσουμε είναι αν υπάρχει συσχέτιση μεταξύ των παραπάνω δεικτών και χαρακτηριστικών των διαφορετικών χωρών. Θα εστιάσουμε σε 3 χαρακτηριστικά από τα δεδομένα: ΑΕΠ (gdp_per_capita), αριθμό κλινών (hospital_beds_per_thousand) και πυκνότητα πληθυσμού (population_density). Θα μελετήσουμε την συσχέτιση με δύο τρόπους. Ο πρώτος είναι μέσω οπτικοποίησης, δημιουργώντας scatter plots για κάθε ζευγάρι δείκτη και χαρακτηριστικού. Στα plots οι χώρες από διαφορετικές ηπείρους θα πρέπει να φαίνονται με διαφορετικό χρώμα (χρησιμοποιείστε το seaborn και το hue). Εμφανίστε τα plots σε ένα grid 3X3. Ο δεύτερος είναι υπολογίζοντας το Pearson Correlation Coefficient, και το p-value (στην περίπτωση αυτή θα πρέπει να αφαιρέσετε null values). Σχολιάστε τα αποτελέσματα και ξεχωρίστε τις συσχετίσεις που είναι στατιστικά σημαντικές (p-value < 0.05). Τι παρατηρείτε στα plots και τις μετρήσεις για το population density? Επαναλάβετε το πείραμα παίρνοντας τον λογάριθμο του population density.

Τι παρατηρείτε στα plots για τις χώρες τις Αφρικής? Επαναλάβετε όλα τα παραπάνω αφαιρώντας τις χώρες τις Αφρικής. Τι παρατηρείτε στα νέα αποτελέσματα? Σχολιάστε τα νέα αποτελέσματα και πως αλλάζουν σε σχέση με πριν.

Τέλος επαναλάβετε όλα τα παραπάνω μόνο για τις χώρες τις Ευρώπης. Τι παρατηρείτε? Πως αλλάζουν τα αποτελέσματα? Σχολιάστε τα νέα αποτελέσματα και πως αλλάζουν σε σχέση με πριν.

Β. Στη συνέχεια εξετάστε τις διαφορές μεταξύ των δεικτών για τις διαφορετικές ηπείρους. Δημιουργείστε bar plots με τις average τιμές ανά ήπειρο, με error bars (χρησιμοποιείστε το seaborn). Τι βλέπετε? Χρησιμοποιήστε το t-test για να εξετάσετε ποιες διαφορές είναι στατιστικά σημαντικές.

C. Σε αυτό το κομμάτι θα χρησιμοποιήσετε το αρχείο που κατεβάσατε στο προηγούμενο βήμα για να μελετήσετε την εξέλιξη των δεικτών στο χρόνο, συνολικά και για κάθε ήπειρο. Θα σπάσετε τα δεδομένα ανά μήνα (ξεκινώντας με τον Ιανουάριο του 2020). Για να επεξεργαστείτε τα δεδομένα θα πρέπει να μετατρέψετε τις ημερομηνίες στην στήλη date σε datetime αντικείμενα. Για κάθε μέρα θα υπολογίσετε τον συνολικό αριθμό νέων κρουσμάτων, συνολικό αριθμό νεων θανάτων, και τον λόγο τους. Θα κάνετε ένα plot με τις μέσες τιμές ανά μήνα, με confidence intervals (χρησιμοποιείστε το seaborn). Σχολιάστε τα αποτελέσματα. Κάνετε ένα επιπλέον plot για να εξετάσετε αν ο αριθμός των κρουσμάτων αυξάνεται εκθετικά.

Επαναλάβετε τις μετρήσεις ανά ήπειρο, και φτιάξετε ένα διάγραμμα με όλες τις ηπείρους μαζί (χρησιμοποιήσετε το hue). Τι παρατηρείτε για την εξέλιξη της πανδημίας ανά ήπειρο?

D. Στο τελευταίο κομμάτι της άσκησης θα εξετάσουμε την πανδημία στις Ηνωμένες Πολιτείες. Κατεβάστε από την σελίδα του CDC τα δεδομένα για την τρέχουσα κατάσταση της πανδημίας ανά πολιτεία. Επίσης εδώ μπορείτε να βρείτε ένα αρχείο με τα αποτελέσματα από τις πρόσφατες εκλογές, όπου για κάθε πολιτεία έχουμε την πληροφορία αν εξέλεξαν Δημοκρατικούς (D) ή Ρεπουμπλικάνους (R). Μας ενδιαφέρει να εξετάσουμε αν υπάρχει στατιστικά ενδιαφέρουσα διαφορά στον βαθμό της έξαρσης της πανδημίας ανάλογα με το τι ψηφίζει η πολιτεία. Σχεδιάστε και υλοποιήστε μετρήσεις για να εξετάσετε αυτή την υπόθεση.

Σημείωση: Στα δεδομένα του CDC η πολιτεία της Νέα Υόρκης και η πόλη της Νέας Υόρκης είναι ξεχωριστά. Θα πρέπει να φροντίσετε τα νούμερα για την πόλη της Νέας Υόρκης να προσμετρηθούν στις μετρήσεις για την πολιτεία της Νέα Υόρκης.

Γενικές οδηγίες: 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. Θα πρέπει να παραδώσετε όλο τον κώδικα σας για την επεξεργασία και τις μετατροπές που κάνετε στα δεδομένα ώστε να παράγετε τις απαιτούμενες μετρήσεις και γραφικές παραστάσεις. Στο σχολιασμό μπορείτε να χρησιμοποιήσετε και γνώση που έχετε για την πανδημία (π.χ., που και πότε έγινε lockdown, πως μεταφέρθηκε η πανδημία, κλπ). Επειδή σας ζητείτε να παράγετε πολλές μετρήσεις και διαγράμματα είναι σημαντικό να γράψετε κώδικα που να αυτοματοποιεί αυτή την διαδικασία. Τοποθετείστε τα διαγράμματα σε grid όποτε θέλετε να δείξετε πολλά διαγράμματα μαζί. Ρυθμίστε το μέγεθος του διαγράμματος ώστε να φαίνονται καλά τα διαγράμματα. Αντίστοιχα μπορείτε να κάνετε πίνακες για να δείξετε τα αποτελέσματα των διαφορετικών στατιστικών τεστ, και τα p-values. Θα πρέπει ο αναγνώστης να μπορεί εύκολα να δει για κάποιο ζεύγος ποια είναι η τιμή του τεστ, και ποιο το p-value.

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). Ο στόχος είναι να χρησιμοποιήστε το Google Trends σε συνδυασμό με τα Covid δεδομένα. Διατυπώσετε ένα δικό σας ερώτημα, και προσπαθήσετε να το απαντήσετε χρησιμοποιώντας δεδομένα από το Google Trends και διαθέσιμα δεδομένα για την πανδημία. 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. Στην μελέτη σας ίσως να σας είναι χρήσιμο να εξετάσετε και την συμπεριφορά σε περιόδους που δεν υπήρχε η πανδημία.

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. Ο βαθμός που θα πάρετε εξαρτάται από την πολυπλοκότητα της ερώτησης που θα εξετάσετε και την δυσκολία της εκτέλεσης.