Τεχνικές Εξόρυξης Δεδομένων.
1η Άσκηση
Ομαδική Εργασία (2 Ατόμων)

Στην εργασία αυτή θα ασχοληθείτε με δεδομένα από γνωστή εφαρμογή ενοικίασης κατοικιών. Συγκεκριμένα σας δίνονται τα δεδομένα για την περιοχή της Αθήνας για 3 μήνες του 2019. Τα δεδομένα είναι σε μορφή csv και θα χρησιμοποιήσετε Python για να απαντήσετε στα παρακάτω ερωτήματα.

Ερώτημα 1ο: Ανάλυση Δεδομένων (Data exploration)

Τα δεδομένα που σας δίνονται είναι οργανωμένα σε 3 φακέλους (april,march, february). Κάθε φάκελος περιέχει διαφορετικά csv αρχεία τα οποία πρέπει να συνδυάσετε και να συνενώσετε κατάλληλα χρησιμοποιώντας την python και το pandas. Συγκεκριμένα θα χρειαστεί να δημιουργήσετε ένα ενιαίο csv αρχείο το οποίο να περιέχει τις παρακάτω στήλες.

id
zipcode
transit
Bedrooms
Beds
Review_scores_rating
Number_of_reviews
Neighbourhood
Name
Latitude
Longitude
Last_review
Instant_bookable
Host_since
Host_response_rate
Host_identity_verified
Host_has_profile_pic
First_review
Description
City
cancellation_policy
Bed_type

Bathrooms
Accommodates
Amenities
Room_type
Property_type
price
Availability_365
Minimum_nights

Σε αυτό το αρχείο (ονομάστε το train.csv) θα χρειαστεί να μελετησετε αν υπάρχουν missing data. Αποφασίστε πως θα τα χειριστείτε και απαλείψτε τα ή συμπληρώστε τα κατάλληλα στο αρχείο train.csv.

Απαντήσετε στα παρακάτω ερωτήματα χρησιμοποιώντας γραφήματα, ιστογράμματα, heat maps, κ.α. είτε με το matplotlib είτε με το seaborn ή και όποια άλλη βιβλιοθήκη επιθυμείτε.

1.1 Ποιός είναι ο πιο συχνός τύπος room_type για τα δεδομένα σας ;
1.2 Φτιάξτε γράφημα ή γραφήματα που δείχνουν την πορεία των τιμών για το διάστημα των 3 μηνών.
1.3 Ποιές είναι οι 5 πρώτες γειτονιές με τις περισσότερες κριτικές;
1.4 Ποιά είναι η γειτονιά με τις περισσότερες καταχωρήσεις ακινήτων;
1.5 Πόσες είναι οι καταχωρήσεις ανά γειτονιά και ανά μήνα;
1.6 Σχεδιάστε το ιστόγραμμα της μεταβλητής neighborhood
1.7 Ποιος είναι ο πιο συχνός τύπος δωματίου (room_type) σε κάθε γειτονιά (neighborhood);
1.8 Ποιός είναι ο πιο ακριβός τύπος δωματίου;
1.9 Χρησιμοποιήστε τη βιβλιοθήκη Folium Map με τις στήλες latitude/longitude και εμφανίστε σε ένα χάρτη για ένα μήνα της επιλογής σας τα ακίνητα και στα popup στον χάρτη επιλέξτε όποια άλλη πληροφορία θέλετε να εμφανίζεται για το ακίνητο (πχ bed_type, room_type, transit κτλ) .
1.10 Φτιάξτε διαφορετικά wordclouds με τα δεδομένα από τη στήλη neighbourhood, transit, description, last_review.
1.12 Ποιά άλλη πληροφορία θα μπορούσατε να εξάγετε από τα δεδομένα αυτά ;
Σκεφτείτε 2 επιπλέον ερωτήσεις για την περιοχή της Αθήνας και εμφανίστε τα αποτελέσματα (μπορείτε και να συνδυάσετε περισσότερες από 2 στήλες).

Ερώτημα 3: Recommendation System

Σε αυτό το ερώτημα θα χρειαστείτε τις στήλες
Id
Name
Description

Σκοπός είναι να εξάγετε χρήσιμη πληροφορία από αυτά τα δεδομένα και να προσπαθήσετε να φτιάξετε ένα πρόγραμμα το οποίο θα παράγει προτάσεις (recommendations) για την περιοχή της Αθήνας. Στο πρώτο ερώτημα έχετε ήδη φτιάξει τα wordclouds για τη στήλη description. Σε αυτό το ερώτημα αφαιρέστε τα stop words, πειραματιστείτε με τις παραμέτρους του wordcloud και εντοπίστε τις πιο χαρακτηριστικές λέξεις που χρησιμοποιεί ο επισκέπτης για την περιοχή της Αθήνας. Στη συνέχεια δημιουργήστε μία νέα στήλη που θα έχει την ένωση (concatenation) των στηλών name και description (fill NA with NULL). Απαντήστε στα παρακάτω:

1. Δημιουργήστε τον TF-IDF (Term Frequency - Inverse Document Frequency) πίνακα των unigrams και των bigrams για τη νέα στήλη (χρησιμοποιήστε την παράμετρο stop_word του TfidfVectorizer).
2. Cosine Similarity: Η μετρική αυτή υπολογίζει την ομοιότητα μεταξύ δύο διανυσμάτων x,y, χρησιμοποιώντας τη γωνία μεταξύ τους (όταν η γωνία είναι 0 σημαίνει ότι τα x και y είναι ίσα , αν εξαιρέσουμε το μήκος τους). Διατρέξτε τον TF-IDF πίνακα και υπολογίστε το similarity καθενός ακινήτου με τα υπόλοιπα. Αποθηκεύστε σε ένα python dictionary τα 100 πιο όμοια ακίνητα.
3. Πρόβλεψη : Φτιάξτε μία συνάρτηση η οποία παίρνει σαν είσοδο ένα id και ένα ακέραιο αριθμό N , και επιστρέφει τα N πιο όμοια ακίνητα.

   ```
   recommend(item_id = 4085439, num = 5)
   ```

   Η έξοδος της συνάρτησης να είναι της παρακάτω μορφής μορφής

   Recommending 5 listings similar to Studio

   Recommended: NAME
   Description: DESCRIPTION
   (score:0.12235188993161432)

   Recommended: NAME
   Description: DESCRIPTION
   (score:0.12235188993161432)
   ……..

4. Λέξεις που εμφανίζονται συχνά μαζί με άλλες λέξεις (collocation). Χρησιμοποιήστε τον BigramCollocationFinder για να βρείτε 10 words που “τείνουν” να εμφανίζονται συχνά μαζί.

Παραδοτέο:

Η εργασία μπορεί να εκπονηθεί **ατομικά ή σε ομάδες 2 ατόμων**.

Θα ανεβάσετε στο eclass ένα φάκελο της μορφής sdixxxx.zip (όπου sdi το ΑΜ ενός εκ των ατόμων της ομάδας) ο οποίος θα περιέχει μόνο τον κώδικά σας σε μορφή **Ipython notebook** (προσοχή: δεν χρειάζεται να ανεβάσετε και το αρχείο train.csv).

Το notebook πρέπει να έχει “τρέξει” ώστε να φαίνονται τα αποτελέσματα της εργασίας σας. Το notebook αποτελεί και την ολοκληρωμένη αναφορά για την εργασία σας (δεν θα παραδώσετε τίποτα σε doc, pdf) , σχεδιάστε το με προσοχή, να θυμάστε να γράψετε μία περιγραφή σε κάθε βήμα για το τι κάνει ο κώδικάς σας σε κάθε κελί.

Διευκρινίσεις για την εργασία θα δίνονται μέσω του e-class.