Τεχνικές Εξόρυξης Δεδομένων
Εαρινό Εξάμηνο 2017-2018

2η Άσκηση, Ημερομηνία παράδοσης: 01/06/2018
Ομαδική Εργασία (2 Ατόμων)

Σκοπός της εργασίας

Σκοπός της εργασίας είναι η εξοικείωσή σας με τα βασικά στάδια της διαδικασίας που ακολουθούνται για την εφαρμογή τεχνικών εξόρυξης δεδομένων, ήτοι: εφαρμογή τεχνικών εξόρυξης δεδομένων και αξιολόγηση. Η υλοποίηση θα γίνει στην γλώσσα προγραμματισμού *Python*.

Περιγραφή των Δεδομένων

Η εργασία σχετίζεται με την κατηγοριοποίηση χωροχρονικών δεδομένων (τροχιές λεωφορείων). Τα Dataset δίνονται σε αρχεία CSV. Οι διαφορετικές στήλες είναι διαχωρισμένες με τον χαρακτήρα ‘,’.
Δίνονται δυο αρχεία:

1. **train_set.csv**:
   Το αρχείο αυτό θα χρησιμοποιηθεί για να εκπαιδεύσετε τους αλγόριθμους σας και περιέχει τα εξής πεδία:

- *TripId*: Μοναδικός αριθμός της τροχιάς που εξήχθη.
- *JourneyPatternId*: Η γραμμή στην οποία υπάγεται η τροχιά.
- $[[t_1, lon_1, lat_1],\ [t_2, lon_2, lat_2],\ ...,\ [t_N, lon_N, lat_N]]$: Η ακολουθία των σημείων της τροχιάς συνοδευόμενα από τον χρόνο.

0	*224-1*	$[[t_0, lon_0, lat_0], [t_1, lon_1, lat_1]]$
1	*224-0*	$[[t_2, lon_2, lat_2], [t_3, lon_3, lat_3], [t_4, lon_4, lat_4], [t_5, lon_5, lat_5]]$
2	*224-1*	$[[t_6, lon_6, lat_6], [t_7, lon_7, lat_7]]$

Το φορτώνεται με τον παρακάτω κώδικα σε pandas:

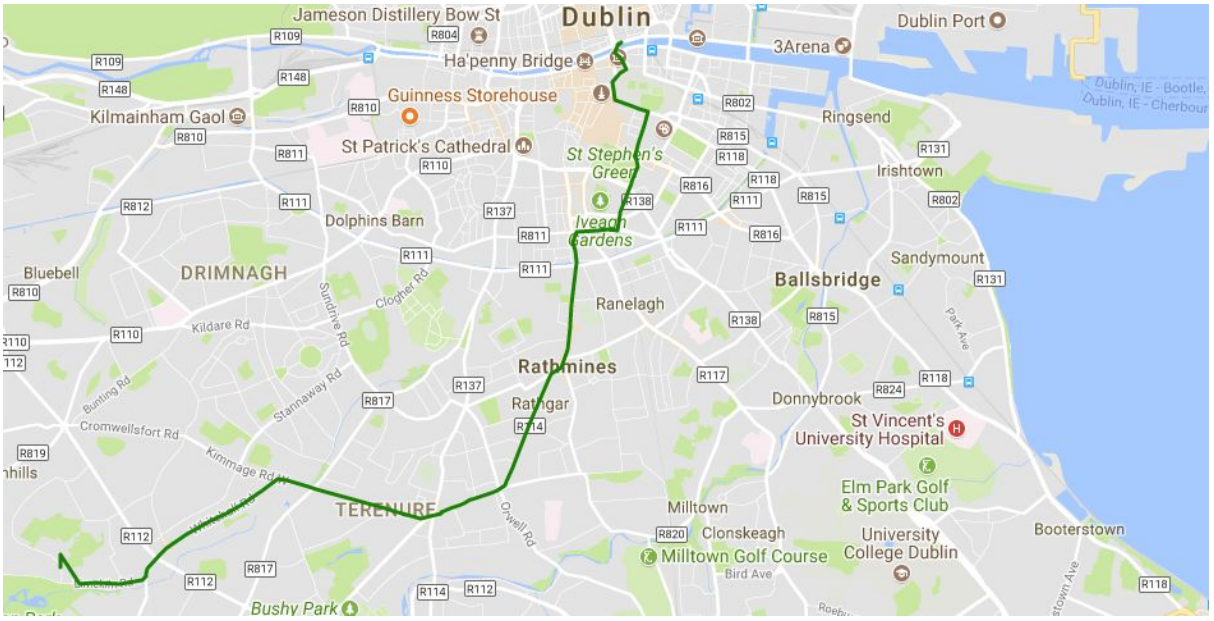
```
import pandas as pd
from ast import literal_eval

trainSet = pd.read_csv(
        'train_set.csv', # replace with the correct path
        converters={"Trajectory": literal_eval},
        index_col='tripId'
)
```

Ερώτημα 1

Οπτικοποίηση των Δεδομένων

Στο σημείο αυτό καλείστε να οπτικοποιήσετε 5 διαδρομές από **διαφορετικές** γραμμές λεωφορείων (journeyPatternID), όπως φαίνεται στην παρακάτω εικόνα. Για την οπτικοποίηση πρέπει να χρησιμοποιήσετε τη βιβλιοθήκη της Python gmplot.



Ερώτημα 2

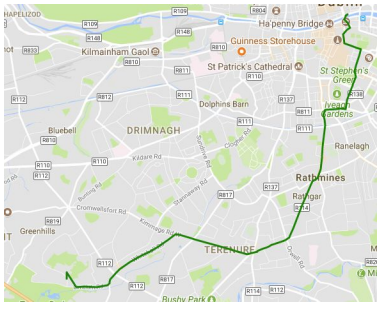
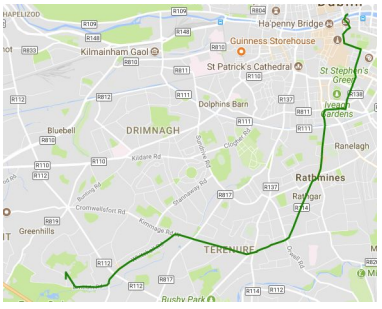
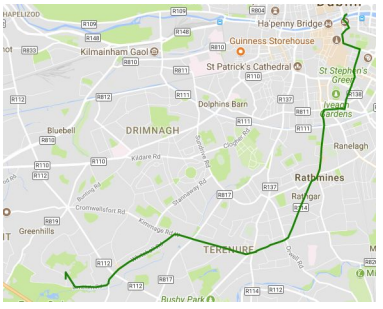
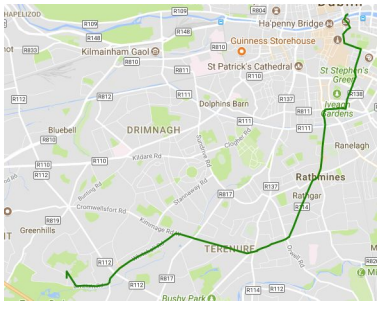
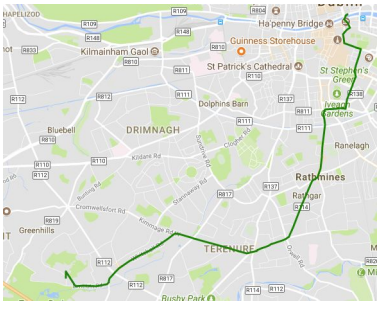
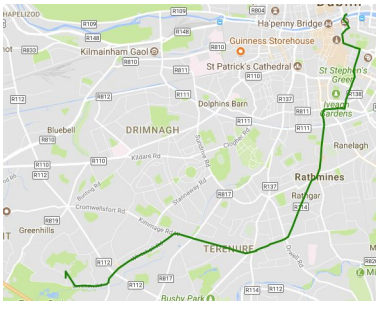
(Α-1) Εύρεση κοντινότερων γειτόνων

Στο σημείο αυτό καλείστε να εντοπίσετε τους κοντινότερους γείτονες χρησιμοποιώντας την τεχνική Dynamic Time Warping (DTW). Θα σας δοθεί το αρχείο “test_set_a1.csv” το οποίο θα περιέχει ένα σύνολο από διαδρομές. Για κάθε μια από τις διαδρομές αυτές θα πρέπει να βρείτε τους 5 κοντινότερος γείτονες από το dataset “train_set.csv”.

Οι γεωγραφικές αποστάσεις ανάμεσα σε δυο σημεία GPS θα πρέπει να υπολογιστούν με τον τύπο Harversine εκφρασμένες σε km.
Για κάθε μια από τις διαδρομές του αρχείου “test_set_a1.csv” θα πρέπει να παρουσιάσετε τα παρακάτω:

- Το JourneyPatternId για κάθε έναν από τους γείτονες που εντοπίστηκαν.
- Την DTW απόσταση με κάθε έναν από τους 5 γείτονες.
- Την οπτικοποίηση της διαδρομής που εξετάστηκε και επίσης την οπτικοποίηση των πέντε γειτόνων (σύνολο 6 εικόνες).
- Το συνολικό χρόνο (Δt) που απαιτήθηκε από το το πρόγραμμα σας για τον εντοπισμό των πλησιέστερων γειτόνων.

Χρησιμοποιήστε το παρακάτω format για την παρουσίαση των αποτελεσμάτων σας.

 Test Trip 1 Δt=600sec	 *Neighbor 1* JP_ID: 224-0 DTW: 20.5km	 *Neighbor 2* JP_ID: 224-0 DTW: 20.7km
 *Neighbor 3* JP_ID: 224-0	 *Neighbor 4* JP_ID: 224-0	 *Neighbor 5* JP_ID: 224-0

DTW: 24km	DTW: 24.5km	DTW: 25.5km

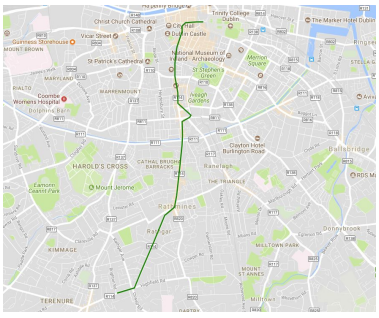
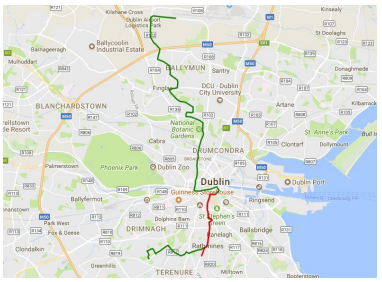
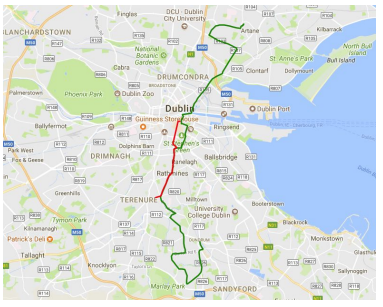
(Α-2) Εύρεση κοντινότερων υποδιαδρομών

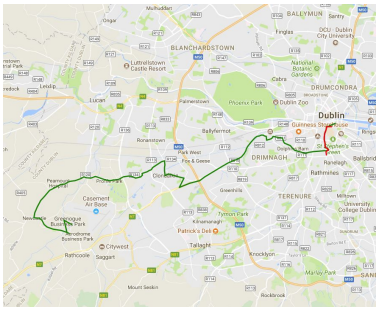
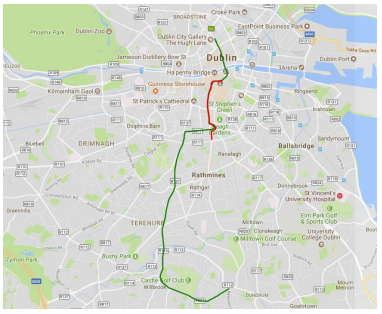
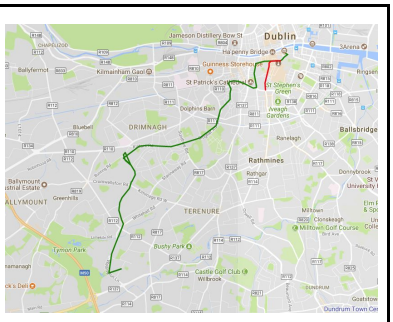
Στο σημείο αυτό θα σας δοθεί το αρχείο “test_set_a2.csv” το οποίο περιέχει ένα σύνολο από διαδρομές. Καλείστε για κάθε μια από αυτές να εντοπίσετε τα *k* κομμάτια των διαδρομών (του αρχείου “train_set.csv”) που είναι παρόμοια. Στο ερώτημα αυτό θα χρησιμοποιήσετε την τεχνική Longest Common Subsequence (LCSS). Δυο σημεία θα γίνονται *match* εάν η απόσταση του δε ξεπερνάει τα 200m.
Οι γεωγραφικές αποστάσεις ανάμεσα σε δυο σημεία GPS θα πρέπει να υπολογιστούν με τον τύπο Harversine εκφρασμένες σε km.

Για κάθε μια από τις διαδρομές του αρχείου “test_set_a2.csv” θα πρέπει να παρουσιάσετε τα παρακάτω:

- Το JourneyPatternId για κάθε έναν από τους γείτονες που εντοπίστηκαν.
- Τον αριθμό των σημείων που έχουν γίνει *match* με κάθε έναν από τους 5 γείτονες.
- Την οπτικοποίηση της διαδρομής που δίνεται και επίσης την οπτικοποίηση των πέντε πλησιέστερων υποδιαδρομών που εντοπίστηκαν με κόκκινο χρώμα και με πράσινο χρώμα ολόκληρη τη διαδρομή του γείτονα (6 εικόνες).
- Το συνολικό χρόνο (Δt) που απαιτήθηκε από το το πρόγραμμα σας για τον εντοπισμό των πλησιέστερων γειτόνων.

Χρησιμοποιήστε το παρακάτω format για την παρουσίαση των αποτελεσμάτων σας.

 Test Trip 1 Δt = 700sec	 *Neighbor 1* JP_ID: 224-0 #Matching Points: 25	 *Neighbor 2* JP_ID: 224-0 #Matching Points: 22

 *Neighbor 3* JP_ID: 224-0 #Matching Points: 16	 *Neighbor 4* JP_ID: 224-0 #Matching Points: 16	 *Neighbor 5* JP_ID: 224-0 #Matching Points: 7.

Ερώτημα 3

Κατηγοριοποίηση

Σε αυτό το ερώτημα θα πρέπει να εφαρμόσετε την μέθοδο κοντινότερων γειτόνων k-nn χρησιμοποιώντας την τιμή 5 για το k και την συνάρτηση ομοιότητας DTW για την πρόβλεψη των γραμμών των διαδρομών που περιέχονται στο αρχείο *"test_set.csv"*. Τις προβλέψεις σας θα πρέπει να τις εισάγετε στο αρχείο *"testSet_JourneyPatternIDs.csv"*. Το format του αρχείου *"testSet_JourneyPatternIDs.csv"*, το οποίο θα περιέχει τις κατηγορίες των διαδρομών που δίνονται στο Test set φαίνεται παρακάτω:

Test_Trip_ID	Predicted_JourneyPatternID
1	224-0
2	250-0
...	

Για το αρχείο *"testSet_JourneyPatternIDs.csv"* θα πρέπει να χρησιμοποιηθεί αυστηρά η παραπάνω μορφοποίηση διαχωρίζοντας τα δυο πεδία με τον χαρακτήρα TAB ('\t') και επίσης θα πρέπει στην πρώτη γραμμή να υπάρχουν οι δυο επικεφαλίδες (Test_Trip_ID και Predicted_JourneyPatternID) και ακολούθως οι προβλέψεις του μοντέλου σας στις επόμενες γραμμές διευκρινίζοντας το ID της διαδρομής από το test set και το αντίστοιχο JourneyPatternID.

Επίσης θα πρέπει να αξιολογήσετε και να καταγράψετε την απόδοση του μοντέλου σας χρησιμοποιώντας 10-fold Cross Validation με τη μετρική Accuracy.

Σχετικά με το παραδοτέο

Ο φάκελος που θα παραδώσετε θα έχει το όνομα Ass2_όνοματεπώνυμο1_ΑΜ1_ονοματεπώνυμο2_ΑΜ2. Ο φάκελος θα περιέχει:

1. Ένα κείμενο με τον σχολιασμό στα πειράματα που κάνατε και στις μεθόδους που δοκιμάσατε σε μορφή PDF. Η αναφορά σας θα πρέπει να περιέχει και τα αποτελέσματα και δε θα πρέπει να ξεπερνάει τις 30 σελίδες.
2. Τα ζητούμενα αρχεία εξόδου.
3. Τους χρόνους εκτέλεσης για τα ερωτήματα 2,3
4. Τα αρχεία κώδικα που γράψατε.

Το εκτενές κείμενο που θα παραδώσετε, θα περιέχει την περιγραφή των δοκιμών σας και οτιδήποτε σκεφτείτε για να δείξετε τι δοκιμές κάνατε, για ποιο λόγο έχουν τα συγκεκριμένα αποτελέσματα οι μέθοδοι που επιλέξατε, πως λειτουργούν αυτές οι μέθοδοι και σχολιασμό των αποτελεσμάτων σας. Όλες οι εργασίες θα αξιολογηθούν στη βάση της σωστής τεκμηρίωσης και στο βαθμό που υλοποιούν τα ζητούμενα της εργασίας.