Analisi Database Crime Boston (sito kaggle)

Il database contiene 327.821 record e 17 colonne.

Colonna delle UCR offence definitions: tipologie di crimini che in america vengono divise in categorie. Diviso in 3 gradi di gravità

UCR Part I sono i crimini più pesanti come omicidio, stupro, ecc (<https://www.ucrdatatool.gov/offenses.cfm>)

Indicazioni pratiche:

Utilizzo la colonna del **offense\_code\_group** per identificare le tipologie, sono 67 categorie!

Con le colonne **lat** e **long** ho la posizione degli eventi e con **occurred\_on\_date** ho la data e ora.

PROBLEMI RISCONTRATI:

nei dati latitudine e longitudine mancano dei record, in particolare o non c’è proprio il dato o segnala -100.000.000. TOLTI (nel database ufficiale)

Dati sono discontinui nel tempo nel senso che nel 2015 sono presenti solo i mesi di luglio e agosto. Nel 2016 e 2017 mancano dei mesi. Il database si ferma a ottobre 2018.

Nel 2016 manca: marzo, aprile, maggio, giugno, novembre, dicembre

Nel 2017 manca: febbraio, marzo

Nel 2018 manca: novembre e dicembre perché si interrompe prima

Struttura della mia tabulazione:

0.Incident\_Number, 1.Offence\_code\_group, 2.Offense\_descr, 3.date, 4.UCR\_part, 5.lat, 6.long, 7.location

Sequenza per implementazione:

* Trovare un modo per avere il set of instances (neighborhood) [Alg5]
* Calcolare PR (fai la funzione)
* Calcolare PI (fai funzione) [Alg4]
* VerifyCandidates(generando prima i candidati C\_2) [Alg2]
* Generare i candidati (Alg3)

Implementare struttura dati ad albero e trovare N-Top patterns

Per neighborhood:

ASSUNZIONI IMPORTANTI:

parser di Location e Time basate sulle stringhe particolari in input che ho (i vari caratteri della stringa), se si cambia zona di riferimento non funzionano!! (dovrei fare un altro tipo di parser)

ho reso intere le coordinate (10 cifre senza virgola) sottratte tra di esse e valutate il raggio, (raggio bene tenerlo sulle 7 cifre di range)

tempo ho trasformato in giorni e sottratto (attraverso la classe datatime)

Creato altra funzione che calcola meglio la distanza (file “calculate\_Location”) però non so se vale la pena utilizzarlo

To do:

* CAPIRE COME CALCOLARE LA DISTANZA Gps [fatto]
* Parser per la data/ora oppure usare year month e hour [fatto con DATATIME]
* Capire se usare anche la descrizione
* Capire se usare anche il giorno della settimana e la street
* Settare e preparare bene una tabella dove ho tabulato tutte le cose nell’ordine e modo più comodo a me, (leggere tutto il file parsare quello che mi serve e salvare tutto in un altro file) [dataset.csv]
* Capire se calcolo del neighborhood va bene e come utilizzarlo per andare avanti