Analisi Database Crime Boston (sito kaggle)

Il database contiene 260.760 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.

To do:

* Parser per la data/ora oppure usare year month e hour
* 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)

PROBLEMI RICONTRATI:

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

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
* Calcolare PR (fai la funzione)
* Calcolare PI (fai funzione) [Alg4]
* VerifyCandidates(generando prima i candidati C\_2) [Alg2]
* Calculate Neighborhood [Alg5]
* Generare i candidati (Alg3)

Implementare struttura dati ad albero e trovare N-Top patterns