Per ogni dataset, di dimensione n=1M, troverai un file senza estensione con le chiavi ordinate (formato binario uint64 little endian) e un file csv con i segmenti.

Il file binario lo puoi leggere ad esempio con Python:

>>> import numpy as np

>>> np.fromfile("wiki\_ts\_1M\_uint64", dtype=np.uint64)

array([ 979672113, 979672371, 979676171, ..., 1050558645, 1050558649,

1050558664], dtype=uint64)

Nel caso in cui doveste fare qualche tipo di training con le coppie (chiave, pos), ti ricordo che alcuni dataset contengono duplicati e dunque devono essere trasformati in coppie (chiave, pos prima occorrenza).

Il file csv contiene colonne:

* data\_index, l’indice di partenza del segmento (0..n-1).
* key, la prima chiave coperta dal segmento, ossia la chiave in posizione data\_index.
* slope, lo slope del segmento.
* intercept, l’intercetta del segmento.

la stima della posizione di una chiave x\_i coperta da un segmento che inizia dalla chiave **key** va fatta tramite:

f(x\_i) = (x\_i-**key**)\*slope + intercept. Key = chiave inizio bin

Fatto questo, dovrebbe valere

|f(x\_i) - rank(x\_i)| ≤ ε = 64,

dove rank(x\_i) è la posizione della prima occorrenza della chiave x\_i. Se nel dataset non ci sono chiavi duplicate vale rank(x\_i)=i.