

Every source: 21×1000 ↑ sample

每个 sample 由 { A: source
B: sample 唯一-确定
G: volume (冲突剧烈)

S: 天数

21天, 每天 1000支股票

对 truth:

R: 天数

每天有 100个 样本, 但原data中有1000个
从 1000 → 100 (?) (昨天的问题)

现对 truth 按A列排序: 如下图

② 对原 data 中 S 个数据源 ($S=55$), 从 $V \in S$ 抽取 $V_m \in M$, 得到第 m 号样本的 S 条数据
ps: 这里应保证每次抽取是同一天 (随机确定一天)
即同一天, 同一支, 不同源, 对应了 truth 中唯一一条
设该条为 t

m_1 : S 条

m_2

\vdots

$m_{|M|}$

③ 对 V_m , 其 S 条同源数据和 t , 利用 Jaccard
计算出一条向量 v

$\{ \{s_i, s_j, 1\}, \dots \}$ s_i, s_j 为对应列 k 值与
 t 相同的数据源标号
(列 k)

④ embedding 相似度比较:

可利用 embedding 生成样本和 t 的 embedding.

得到二者相似度, 和 ③ 中的 v (比较)
(比较方法?)

⑤ 可视化 (<簇状>)