固定一个月，分析指标在样本排序上的关联性

Created By	
Stakeholders	
Status	
Type	Technical Spec
Created	@June 9, 2022 5:37 PM
Last Edited Time	@June 9, 2022 6:00 PM
Last Edited By	

目标
解决方案
    构思
问题

目标

- 排序名次上的关联性：排序名次上：基于不同的特征的值进行样本排序，样本在排序结果的名次上一致性高的特征组/对。容易解读。
- 固定一个月，输入样本集合，在每一个指标上进行样本排序，对比两个指标的样本排序结果，计算出两个指标的关联性。

解决方案

构思

- 方案一：计算两个排序的最大公共子序列、字串

- 方案二
  - 在每个排序上计算样本的“得分”（0-1；按名次或按数值）
  - 对比两个排序上对应样本的得分，取差的平方
  - 汇总差的平方，除以有效样本数

```
month252 = mr.read_one_month_values(252,
                                    govs_ids=[3, 4, 5, 1, 0],
                                    prod_indexes_ids=[4, 125, 71909, 63331])
month252
```

✓ 0.4s

	gov_id	month	63331	4	125	71909
0	0.0	252.0	1302461.0	NaN	NaN	NaN
1	1.0	252.0	29592.0	0.0385	7.7090	NaN
3	3.0	252.0	673.0	0.0000	9.8405	NaN
4	4.0	252.0	4277.0	0.0000	8.3459	NaN
5	5.0	252.0	2672.0	0.0000	7.2058	NaN

```
rankings = omo.order_govs_by_every_index(month252)
rankings
```

✓ 0.5s

```
{4: {1.0: 1.0, 3.0: 0.0, 4.0: 0.0, 5.0: 0.0},
 125: {1.0: 0.3333333333333333, 3.0: 1.0, 4.0: 0.6666666666666666, 5.0: 0.0},
 63331: {0.0: 1.0, 1.0: 0.75, 3.0: 0.0, 4.0: 0.5, 5.0: 0.25},
 71909: {}}
```

```
losses, counts = omo.calc_losses_between_indexes(rankings)
print(losses)
print(counts)
```

✓ 0.5s

```
{(63331, 4): 0.09375, (63331, 125): 0.3159722222222222, (4, 125): 0.47222222222222227}
{(63331, 4): 4, (63331, 125): 4, (4, 125): 4}
```

问题

- 有没有现成的计算“样本排序上的关联性”的包
- 方案二：按名次给均匀分布的得分，还是按值给变化较大的得分？
- 方案二：可能会有点慢