dataSampleUtil

Spark自身提供了sample方法，使用签名是：

|  |
| --- |
| /\*\*  \* Returns a new Dataset by sampling a fraction of rows.  \*  \* @param withReplacement Sample with replacement or not.  \* @param fraction Fraction of rows to generate.  \* @param seed Seed for sampling.  \*  \* @group typedrel  \* @since 1.6.0  \*/  def sample(withReplacement: Boolean, fraction: Double, seed: Long): Dataset[T] = withTypedPlan {  Sample(0.0, fraction, withReplacement, seed, logicalPlan)() } |

1. 是否有放回对结果的影响

在使用中如果使用不放回的操作，在相同种子下，每次执行产生的结果是相同的，

在使用有放回的操作时，在相同的种子下会有重复，每次执行产生的结果是相同的；

1. 种子对结果的影响

种子决定了数据采样的数据量和数据分布，但是并没有什么特定关系，在使用中使用默认参数就可以。