# 检测工具使用说明

目录

**[运行说明](#_Toc24793_WPSOffice_Level1)** **[1](#_Toc24793_WPSOffice_Level1)**

**[1. 工具文件](#_Toc26483_WPSOffice_Level1)** **[1](#_Toc26483_WPSOffice_Level1)**

[(1) SACloneDetector.jar：工具JAR包](#_Toc26483_WPSOffice_Level2) [1](#_Toc26483_WPSOffice_Level2)

[(2) config.properties：配置文件](#_Toc4313_WPSOffice_Level2) [1](#_Toc4313_WPSOffice_Level2)

[(3) executable：文件夹内包含各系统的检测exe](#_Toc26898_WPSOffice_Level2) [1](#_Toc26898_WPSOffice_Level2)

**[2. 配置文件](#_Toc4313_WPSOffice_Level1)** **[1](#_Toc4313_WPSOffice_Level1)**

**[3. 使用方法](#_Toc26898_WPSOffice_Level1)** **[3](#_Toc26898_WPSOffice_Level1)**

[(1) 按照第2步修改配置文件](#_Toc17562_WPSOffice_Level2) [3](#_Toc17562_WPSOffice_Level2)

[(2) 运行jar包](#_Toc8474_WPSOffice_Level2) [3](#_Toc8474_WPSOffice_Level2)

[(3) java -jar SACloneDetector.jar dataset\_path](#_Toc11078_WPSOffice_Level2) [3](#_Toc11078_WPSOffice_Level2)

**[4. 注意](#_Toc17562_WPSOffice_Level1)** **[3](#_Toc17562_WPSOffice_Level1)**

**[输出文件说明](#_Toc8474_WPSOffice_Level1)** **[3](#_Toc8474_WPSOffice_Level1)**

**运行说明**

1. **工具文件**
   1. SACloneDetector.jar：工具JAR包
   2. config.properties：配置文件
   3. executable：文件夹内包含各系统的检测exe
2. **配置文件**

配置项说明：

标黄配置项为可能需要修改的项，未标黄配置项可默认不变。

1. minLine=2

一个方法的最小检测行数，可以通过设置这个参数筛除不检测的方法

1. mlcc=20

后缀数组中两个片段的token重叠部分数量，不连续

设置的数值越大，检测结果越倾向于高度相似的长方法

1. language=java

数据集源文件类型，需要改为待检测的语言，如”java”,”c”,”cpp”,”h”

1. threshold=0.7

克隆检测的阈值，取值范围0-1

阈值 = 重叠部分token数量 / 其中较长的那个片段的token数量

1. openStringHash=1

是否在token化中开启字符串哈希，0→关闭，1→开启

字符串哈希根据每一个字符串后三位计算哈希，-128~-3

用于提高字符串区分度

1. granularity=method

检测粒度，按需更改，包含”method”,”file”,”snippet”

1. extensions=java

待检测文件的后缀，按需更改，多个后缀间用逗号分隔

1. sepNum=200000000

数据块的分割大小，当token数大于sepNum时，存入磁盘

在GPU版本中，该参数若设置过大，将无法解析构造后缀数组

1. mlc=50

一个方法的最小token数，若一个方法的token数小于mlc，则丢弃

1. exe=.\\executable\\executable\_cpu\_win10.exe

可执行文件的路径，通过此参数决定使用特定系统的cpu或gpu检测

【注意】

* 配置项 (6) granularity 决定工具检测粒度
* 配置项 (10) exe 决定工具是CPU版本还是GPU版本，以及决定是否使用片段级克隆检测。如果使用片段级检测，granularity参数应选择snippet，exe文件应选择executable\_gpu\_win10\_snippet\_type12.exe或executable\_gpu\_win10\_snippet\_type123.exe，前者包含Type1、2克隆检测，后者包含Type1、2、3克隆检测。

1. **数据集格式**

待检测数据集推荐放在LargeScaleCodeClone/dataset下，若放在其他位置【也应按照以下目录结构】，dataset\_path为数据集根目录名称，如下：

dataset\_path/author1/project1

/project2

/project3

/author2/project4

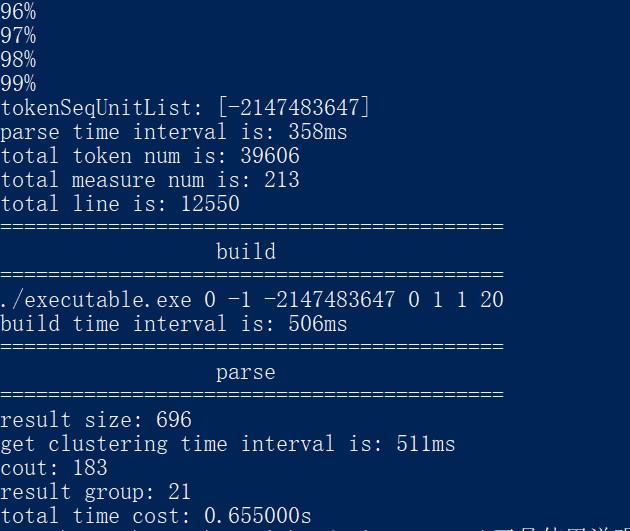
/project5

…

1. **使用方法**
   1. 按照第2步修改配置文件
   2. 按照第3步存放数据集
   3. 运行jar包 java -jar SACloneDetector.jar dataset\_path

（dataset\_path为数据集根目录名称）

检测结果：



1. **注意**

再次运行前请先删除所有生成的文件（result和tokenData文件夹）

**输出文件说明**

1. tokenData文件夹：存放中间文件
2. result文件夹：存放结果文件
   1. files.txt

进行克隆检测的文件路径

* 1. MeasureIndex.csv

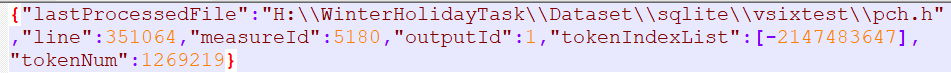
每条数据有4个数据列，分别表示：

方法ID，方法所在路径，方法起始行，方法结束行



* 1. state.json

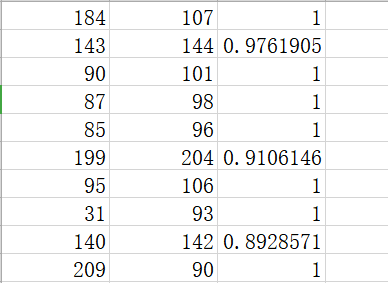
工具检测项目的基本信息



* 1. type123\_method\_result.csv

克隆对结果，每条数据有3个数据列，表示：

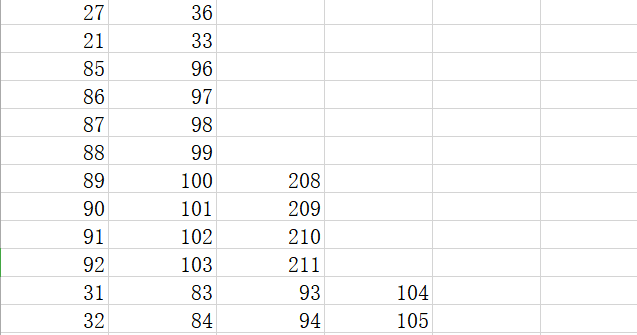
一对相似克隆片段的方法ID，以及它们的相似程度



* 1. type123\_method\_group\_result.csv

克隆组结果，每条数据有2个及以上数据列，表示：

一组相似克隆片段的方法ID（对应MeasureIndex.csv中的方法ID）

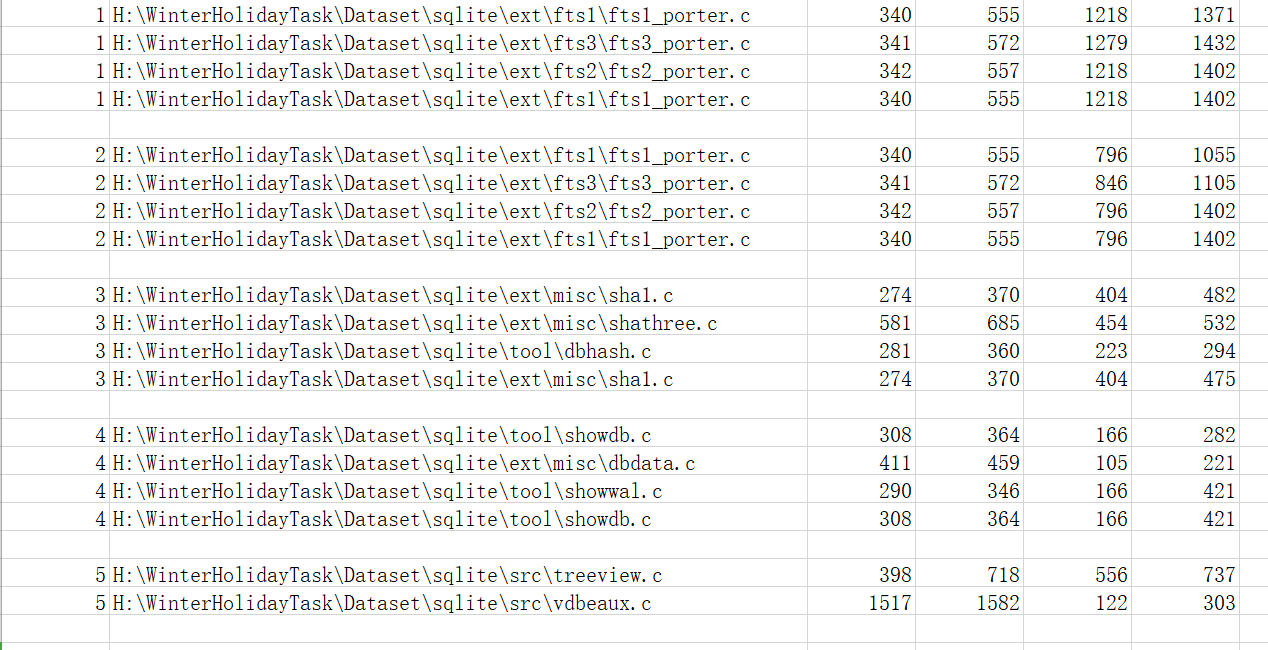


1. 如果使用snippet片段级检测，则不生成结果文件2.(4) type123\_method\_result.csv和2.(5) type123\_method\_group\_result.csv，取而代之的是：
   1. type12\_snippet\_result.csv

Type1、2克隆检测结果，每条数据有6个数据列，分别表示：

克隆组ID，文件路径，起始行，结束行，起始token，结束token

（起始token和结束token指片段再该方法token串的序号）



* 1. type3\_snippet\_result.csv

Type3克隆检测结果，格式同上。