

Санкт-Петербургский государственный университет

Кафедра системного программирования

Группа 22.Б11-мм

Анализ неустойчивости выделенного ЮНИТ-теста

Демченко Артем Евгеньевич

Отчёт по производственной практике
в форме «Производственное задание»

Научный руководитель:
к. ф.-м. н., доц. Куликов Е. К.

Санкт-Петербург
2025

Оглавление

Введение	3
1. Постановка задачи	5
2. Обзор	6
2.1. Методы обнаружения нестабильных тестов	6
2.2. Используемые технологии	8
3. Реализация	10
3.1. Запуск пользовательских тестов	11
3.2. Анализатор тестов на предмет нестабильности	11
Заключение	13
Список литературы	14

Введение

Как известно, тестирование — необходимый этап разработки ПО, помогающий обеспечить качество создаваемого продукта. Тестирование можно разделить на два вида — ручное и автоматизированное. При автоматизированном подходе тестовые сценарии проверяются без участия человека. Этот процесс инициируется довольно часто, например, при непрерывной интеграции или выпуске очередной версии продукта. В первом случае тесты запускаются каждый раз при попытке внедрения новых изменений в проект. От их результатов зависит, готовы ли эти изменения к интеграции. Зачастую хотя бы один провалившийся тест означает, что код нужно дорабатывать.

Нестабильные тесты — тесты, результат которых может меняться без изменения исходного кода. Причины могут быть разные — недетерминированность в коде, зависимость от внешних компонентов, некорректная инициализация и/или очистка тестового окружения. Выходит, что если какой-то тест в проекте не проходит, то программистам необходимо определять, менялся ли исходный код, относящийся к провальному тесту, и на основе этого выяснять, является ли этот тест нестабильным. В статье блога компании Google, посвященному тестированию, отмечается, что примерно 1,5% от всех их тестов являются нестабильными [10]. Таким образом, для проекта с 1000 тестами примерно 15 дают нестабильные результаты при каждом запуске. На практике зачастую такие тесты помещают в карантин, то есть исключают из набора запускаемых тестов. Это приводит к тому, что часть функциональности остается не протестированной. При исправлении же нестабильного теста возникают дополнительные трудности по сравнению с исправлением теста, который всегда не проходит. Во-первых, ошибка может быть как в тестируемом коде, так и в коде самого теста. Во-вторых, провальный результат нестабильного теста не всегда можно воспроизвести. Следовательно, процесс определения, является ли тест нестабильным, а также его исправление сказывается на времени разработки, замедляя ее [12]. Возникает задача по созданию инструментов,

помогающих разработчикам определять нестабильные тесты на ранних этапах и исправлять их.

Explyt Test [3] — динамично развивающийся продукт, представляющий собой плагин для IntelliJ IDEA. На данный момент он в основном нацелен на автоматическую генерацию тестов для Java и Kotlin, но планируется, что Explyt Test станет инструментом всестороннего ухода за тестами, в том числе помощником при работе с нестабильными тестами. Поэтому анализатор тестов на предмет нестабильности должен стать частью этого инструмента. Такая проверка будет полезна как для уже существующих в проекте тестов, так и для вновь сгенерированных с помощью Explyt Test.

1. Постановка задачи

Цель работы — разработать анализатор, определяющий, является ли выделенный юнит-тест нестабильным, и предоставляющий полезную для исправления информацию. Для достижения цели были поставлены задачи:

1. рассмотреть существующие способы анализа нестабильности тестов;
2. реализовать анализатор юнит-теста на предмет нестабильности, извлекающий полезную для его исправления информацию;
3. исследовать возможности больших языковых моделей в анализе причин нестабильности тестов;
4. апробировать разработанный инструмент на наборе модельных примеров и на нестабильных тестах из реальных проектов;
5. реализовать механизм представления результатов анализа пользователю;
6. внедрить разработанные компоненты в Explyt Test.

2. Обзор

2.1. Методы обнаружения нестабильных тестов

Основным источником для поиска возможных подходов определения, является ли тест нестабильным, послужила большая обзорная статья [12] 2021 года, посвященная систематизации знаний из других 76 статей на тему нестабильных тестов. В статье также исследуются причины происхождения, последствия для проектов и способы исправления. Далее будут рассмотрены инструменты, нацеленные на обнаружение нестабильных тестов.

Инструмент iDFlakies [15] предназначен для обнаружения и классификации нестабильных тестов. Он представляет собой maven-плагин, работающий с тестовым фреймворком JUnit. В этом заключается ограничение данного инструмента. Для определения того, является ли тест нестабильным, авторы используют стратегию **ReRun** — запуск тестов несколько раз и сравнение результатов. Таким образом, проверка выполняется по определению нестабильного теста. Плюс данной стратегии — несложная реализация, минус — затраты ресурсов на повторный запуск тестов.

DeFlaker [2] — еще один инструмент, целью которого является автоматическое обнаружение нестабильных тестов. Его преимуществом перед стратегией **ReRun** является то, что запуск тестов происходит только один раз. Принцип работы таков: анализ тестов происходит при внесении каких-либо изменений в систему контроля версий. Осуществляется инструментированный запуск тестов и на основе их работы строится покрытие. Затем анализируются полученные данные. Если тест провалился и измененный код не покрыт, то такой тест считается нестабильным. Такой подход имеет недостаток, так как нацелен на обнаружение нестабильных тестов, которые уже существуют в проекте.

NonDex [11] — инструмент, предназначенный для поиска тестов, реализация которых опирается на неверные предположения о детерминированности недетерминированных спецификаций в Java. Примером это-

му может служить порядок получения элементов `HashMap`. При этом зачастую на одинаковых платформах и версиях Java реализация спецификаций будет одной и той же, поэтому определять нестабильность тестов в этом случае становится трудно. Авторы инструмента запускают тесты, при этом подменяя вызов недетерминированного API своими реализациями. Например, реализация `HashMap` заменяется такой, что при итерации по элементам их получение происходит в случайном порядке. Всего в инструменте поддерживается 41 недетерминированных спецификаций Java [16]. Такой подход действительно находит неверные предположения о них, но в целом затрагивает довольно узкий спектр причин возникновения нестабильных тестов.

Также стоит отметить, что после 2021 года продолжают выходить статьи, посвященные анализу тестов на предмет нестабильности. В одной из них [4] описывается инструмент `Flakify` — анализатор, основанный на языковой модели `CodeBERT` [1], которая была дообучена авторами. `Flakify` определяет, является ли тест нестабильным, опираясь только на его исходный код. `CodeBERT` имеет ограничение на количество входных токенов (512). Авторы разработали эвристику уменьшения количества токенов, в которые конвертируется тестовый метод. Тем не менее полностью избавиться от этого ограничения нельзя. Код анализатора выгружен на GitHub [5], но он никак не поддерживается более двух лет. Хотя в README репозитория есть инструкция, как воспроизвести результаты, полученные авторами в статье, документации по использованию инструмента в других проектах нет. Исходный код также не документирован.

Таким образом, учитывая ограничения существующих инструментов, специфику `Explyt Test` и необходимость анализировать только что сгенерированные тесты, для анализатора была выбрана стратегия `ReRun`. Также такая стратегия позволяет получить информацию о трассах исполнения тестов, а затем анализировать расхождения успешного и провального запусков.

2.2. Используемые технологии

Анализатор должен быть интегрирован в Explyt Test — плагин для IntelliJ IDEA, построенный на IntelliJ Platform [6]. Поэтому разрабатываемый анализатор также будет использовать данное API. Написание плагинов поддерживается на языках Java и Kotlin. Для анализатора был выбран Kotlin, так как на нем написан Explyt Test. В Explyt Test основной способ взаимодействия с пользователями — использование Action-ов [7], предоставляемых IntelliJ Platform. Например, точка входа для генерации тестов использует этот механизм. Поэтому точка входа для анализа нестабильности юнит-теста должна быть реализована в виде Action-а, обеспечивая единый стиль взаимодействия с пользователем. При этом IntelliJ Platform также предоставляет Execution API [8], благодаря которому становится возможным запуск тестов в проекте пользователя и их анализ. Ключевые его интерфейсы:

- `RunConfiguration`. Описывает процесс, который может быть запущен. Позволяет указывать рабочую директорию, переменные среды, а также аргументы командной строки.
- `ProgramRunner`. Ответственен за процесс исполнения. Должен имплементировать точку расширения `com.intellij.programRunner`.
- `RunConfigurationExtension`. Позволяет модифицировать `RunConfiguration`, добавляя, например, аргумент командной строки.

Для межпроцессного взаимодействия в Explyt Test используется фреймворк JetBrains RD (Reactive Distributed) [9]. RD предоставляет возможность создания модели данных, которая автоматически синхронизируется на обеих сторонах (клиентской и серверной). Также поддерживается автоматическая генерация исходного кода для классов модели и методов сериализации и десериализации этой модели на основе декларативного описания, что уменьшает количество кода и упрощает разработку.

Также в Explyt Test уже реализован компонент runtime, который предоставляет java agent, позволяющий совершать инструментированные запуски тестов и собирать трассы исполнения. Предполагается, что результаты работы тестов и собранные трассы исполнения передаются по протоколу RD.

Трассы исполнения, получаемые благодаря компоненту runtime, представляют собой список инструкций. Каждая является имплементацией интерфейса JcInst — части библиотеки JacoDB [14]. Это библиотека, предоставляющая способ получить некоторую информацию о байт-коде вне процесса JVM. Js-инструкции как раз будут полезны для предоставления пользователю информации о расхождении успешного и провального запусков нестабильного теста. Для этого нужно будет разработать способ сопоставлять конкретной Js-инструкции строчку в исходном файле.

Для отправки запросов в LLM и дальнейшего анализа результатов будет использоваться ai client — модуль в Explyt Test, отвечающий за взаимодействие с большими языковыми моделями.

3. Реализация

На рисунке 1 представлена архитектура инструмента с зависимыми модулями. Для осуществления стратегии ReRun в проекте был реализован модуль test runner, осуществляющий запуск тестов пользователя. Для сбора трасс исполнения тестов использовался существующий модуль runtime, предоставляющий java agent, который позволяет выполнять инструментированные запуски и собирать трассы. Изначально в проекте трассы собирались в виде покрытия кода. Поскольку для анализатора необходима полная последовательность выполненных инструкций, были внесены изменения в usvm [13] — библиотеку с открытым исходным кодом, которая используется для сбора трасс. Модули runtime и test runner взаимодействуют по протоколу RD. В основе взаимодействия лежит общая модель данных. На стороне test runner регистрируется обработчик изменений этой модели, а на стороне runtime в нее записываются результаты запуска тестов и трассы.

В итоге модуль с анализатором использует API, предоставляемое модулем test runner для повторного запуска тестов. При определении, что пользовательский тест является нестабильным, отправляется запрос в LLM с помощью модуля ai client.

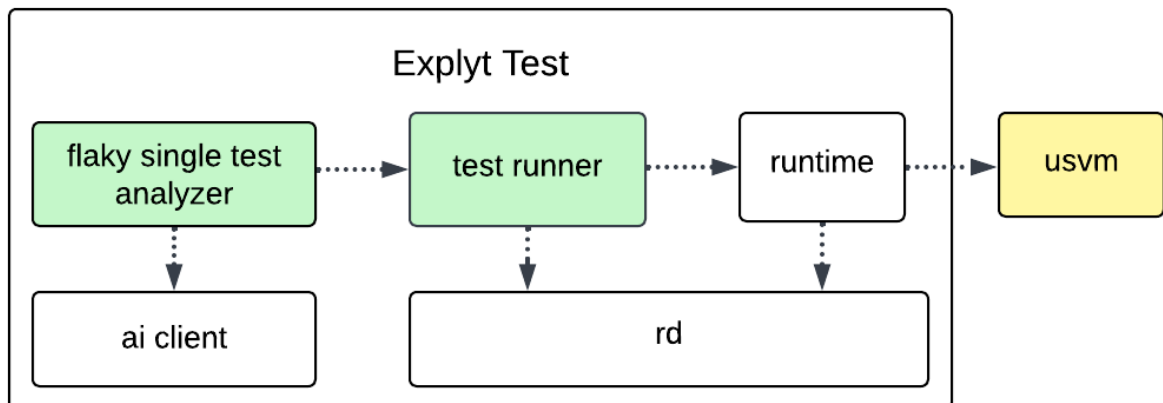


Рис. 1: Зависимость модулей в Explyt Test. Зеленым отмечены те, что были реализованы в данной работе. Желтым — те, в которые были внесены изменения

3.1. Запуск пользовательских тестов

На рисунке 2 представлена архитектура модуля test runner. Этот модуль реализован с целью предоставления публичного API для запуска пользовательских тестов и сбора трасс исполнения. Непосредственно запуск осуществляется механизмами IntelliJ Platform. Для возможности инструментированного запуска потребовалось реализовать выделенные компоненты. Например, в методе `updateJavaParameters(...)` реализована логика добавления опции `-javaagent` к команде процесса запуска тестов.

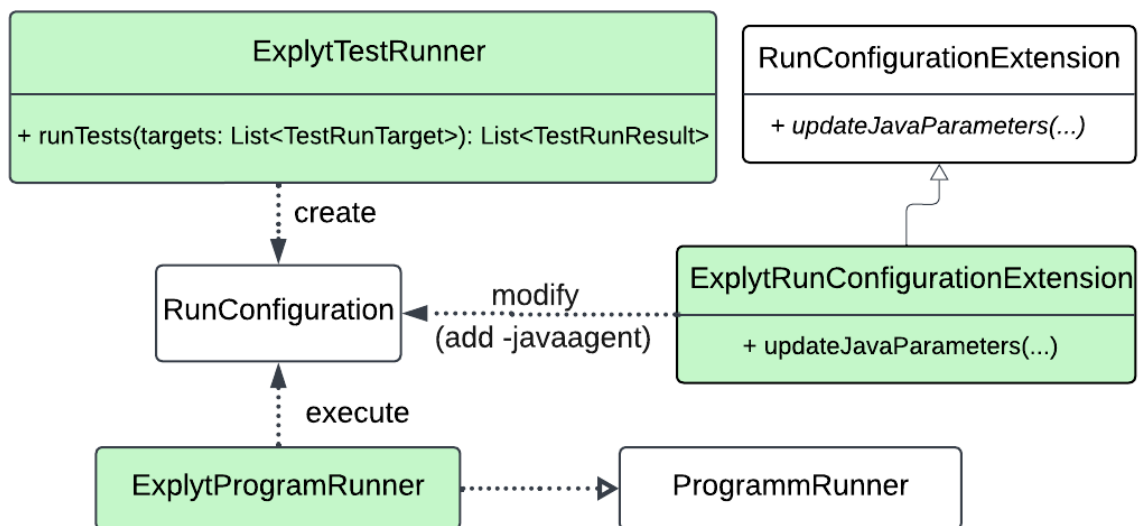


Рис. 2: Архитектура модуля test runner

3.2. Анализатор тестов на предмет нестабильности

Алгоритм анализа выделенного юнит-теста:

1. Тест запускается N раз.
2. Проверяется, есть ли разные результаты среди запусков. Если есть — тест распознается как нестабильный и совершается переход к следующему шагу, иначе алгоритм завершается.

3. Рассматриваются две трассы исполнения: успешная и провальная. Ищется наибольший общий совпадающий префикс. Таким образом выделяется последняя общая для двух трасс инструкция.
4. По префиксу трассы строится стек вызовов. Совершается линейный проход по списку инструкций. Если встречается инструкция «call», то она добавляется на стек. Встречается «return» — со стека удаляется последняя инструкция. Остальные инструкции игнорируются.
5. Каждой инструкции сопоставляется строчка в исходном файле.
6. Отправляется запрос в LLM с целью описать причину неустойчивости теста.

Заключение

В ходе данной производственной практики были решены следующие задачи:

1. Рассмотрены существующие способы анализа нестабильности тестов. Среди них для реализации был выбран тот, который лучше всего подходит для интеграции в Explyt Test.
2. Реализован анализатор, определяющий, является ли выделенный юнит-тест нестабильным, и предоставляющий данные для его исправления. Они включают первую строчку кода, после которой расходится исполнение провального и успешного запуска, и стек вызовов, приводящий к расхождению.
3. Реализована первая версия запроса к большой языковой модели и получение результатов.
4. Анализатор апробирован на модельном примере.

Дальше планируется решить следующие задачи:

1. Разработать специализированный промпт для взаимодействия с большой языковой моделью.
2. Исследовать возможности больших языковых моделей в анализе причин нестабильности тестов.
3. Апробировать разработанный инструмент на нестабильных тестах из реальных проектов.
4. Реализовать механизм представления результатов анализа пользователю.
5. Внедрить разработанные компоненты в Explyt Test.

Код анализатора находится под договором о неразглашении, поэтому не может быть предоставлен.

Код изменений в UnitTestBot usvm доступен по ссылке: <https://github.com/UnitTestBot/usvm/pull/229>

Список литературы

- [1] [CodeBERT: A Pre-Trained Model for Programming and Natural Languages](#) / Zhangyin Feng, Daya Guo, Duyu Tang et al. — 2020. — 01. — P. 1536–1547.
- [2] [DeFlaker: automatically detecting flaky tests](#) / Jonathan Bell, Owolabi Legunsen, Michael Hilton et al. // Proceedings of the 40th International Conference on Software Engineering. — ICSE '18. — New York, NY, USA : Association for Computing Machinery, 2018. — P. 433–444. — URL: <https://doi.org/10.1145/3180155.3180164>.
- [3] Explyt Test. — URL: <https://explyt.com/test/> (дата обращения: 5 января 2025 г.).
- [4] Fatima Sakina, Ghaleb Taher A., Briand Lionel. Flakify: A Black-Box, Language Model-Based Predictor for Flaky Tests // [IEEE Transactions on Software Engineering](#). — 2023. — Vol. 49, no. 4. — P. 1912–1927.
- [5] Flakify Test Smell Detector. — URL: <https://github.com/uOttawa-Nanda-Lab/Flakify> (дата обращения: 5 января 2025 г.).
- [6] IntelliJ Platform. — URL: <https://plugins.jetbrains.com/docs/intellij/welcome.html> (дата обращения: 5 января 2025 г.).
- [7] IntelliJ Platform Actions. — URL: <https://plugins.jetbrains.com/docs/intellij/basic-action-system.html> (дата обращения: 5 января 2025 г.).
- [8] IntelliJ Platform Execution API. — URL: <https://plugins.jetbrains.com/docs/intellij/execution.html> (дата обращения: 5 января 2025 г.).
- [9] JetBrains RD. — URL: <https://github.com/JetBrains/rd> (дата обращения: 5 января 2025 г.).

- [10] Micco John. Flaky Tests at Google and How We Mitigate Them. — 2016. — URL: <https://testing.googleblog.com/2016/05/flaky-tests-at-google-and-how-we.html> (дата обращения: 5 января 2025 г.).
- [11] NonDex: a tool for detecting and debugging wrong assumptions on Java API specifications / Alex Gyori, Ben Lambeth, August Shi et al. // Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering. — FSE 2016. — New York, NY, USA : Association for Computing Machinery, 2016. — P. 993–997. — URL: <https://doi.org/10.1145/2950290.2983932>.
- [12] A Survey of Flaky Tests / Owain Parry, Gregory M. Kapfhammer, Michael Hilton, Phil McMinn // ACM Trans. Softw. Eng. Methodol. — 2021. — . — Vol. 31, no. 1. — 74 p. — URL: <https://doi.org/10.1145/3476105>.
- [13] Unit Test Bot USVM. — URL: <https://github.com/UnitTestBot/usvm> (дата обращения: 5 января 2025 г.).
- [14] UnitTestBot JacoDB. — URL: <https://github.com/UnitTestBot/jacodb> (дата обращения: 5 января 2025 г.).
- [15] iDFlakies: A Framework for Detecting and Partially Classifying Flaky Tests / Wing Lam, Reed Oei, August Shi et al. // 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST). — 2019. — P. 312–322.
- [16] Исходный код инструмента NonDex. — URL: <https://github.com/TestingResearchIllinois/NonDex> (дата обращения: 5 января 2025 г.).