Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра алгоритмических языков

КУРСОВАЯ РАБОТА

Построение вопросно-ответной системы на основе программного обеспечения и его артефактов

Студент:
Жуков Павел Николаевич, 525 группа

Научный руководитель:
Головин Игорь Геннадьевич, к.ф.-м.н.

Введение

- **Вопросно-ответная система** — это информационная система, способная принимать вопросы и отвечать на них на естественном языке.

- В ходе коллективной разработки программного обеспечения в большинстве случаев используются **вспомогательные средства**, такие как системы контроля версий, системы отслеживания задач и ошибок и т.д.

- Также в результате разработки ПО появляется множество **артефактов**, например, непосредственно код, логи выполнения, документация, почтовая переписка и другие.

Актуальность работы

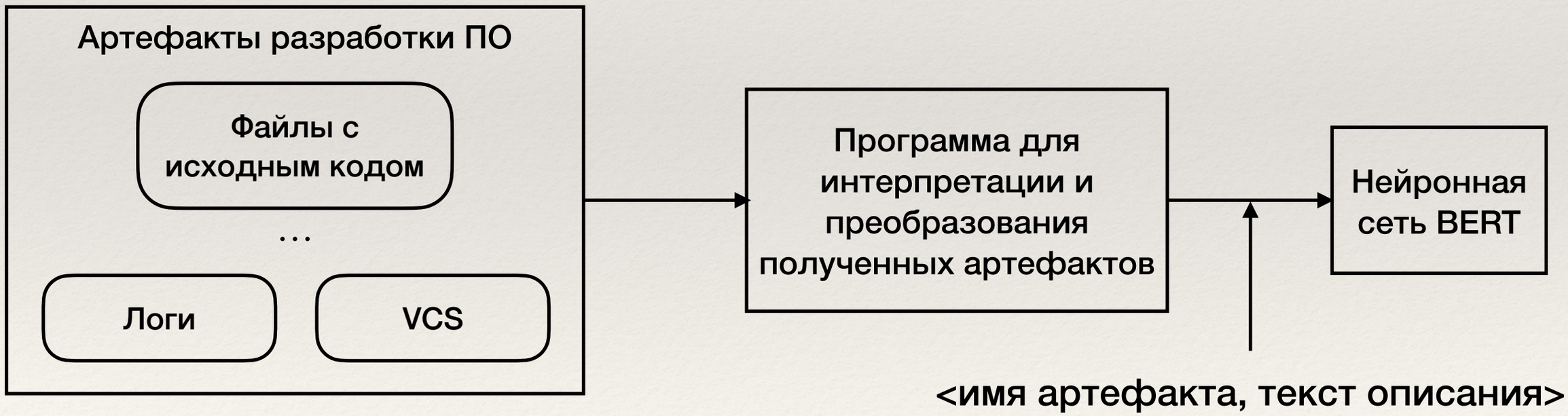
- Документация на естественном языке имеет свойство быстро устаревать, так как требует постоянной поддержки.
- Многие специалисты, такие как тестировщики ПО или аналитики, не обязаны понимать код и другие артефакты на уровне разработчиков, но некоторые данные им необходимы для эффективного выполнения своей работы.
- Процесс присоединения нового человека к команде разработки также затрудняется из-за больших объёмов информации.

Постановка задачи

Разработать вопросно-ответную систему, позволяющую узнавать информацию о программном обеспечении, которая изначально не доступна на естественном языке.

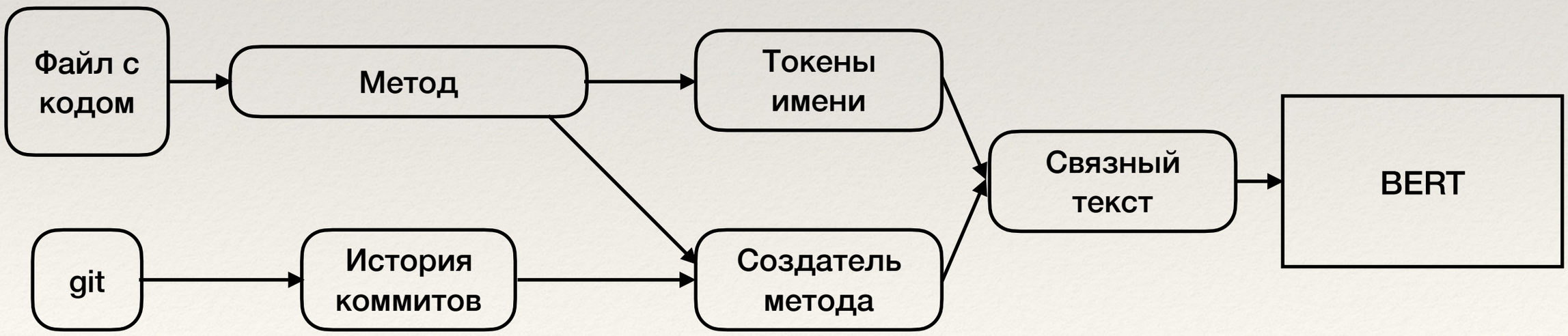
Подход к решению задачи (1)

- В результате исследования области было выявлено, что среди вопросно-ответных систем с большим отрывом побеждает нейронная сеть BERT.
- В соответствии с этим фактом было решено на основе артефактов ПО автоматически генерировать связный текст на английском языке и подавать его на вход предобученной нейросети.



Подход к решению задачи (2)

- Основной сложностью здесь является задача генерация такого текста, который был бы похож на естественный, использовал правильные признаки и был понятен нейронной сети.
- Было решено начать с задачи ответа на вопрос о создателе какого-либо метода, так как она включает в себя и анализ дерева git-коммитов, и проход по синтаксическому дереву программного кода.



Пример работы

Вопрос: Who did implement a method for sending nickname after it is changed?

Ответ: Dmitry Lyukov

Файл: ChangeIdCommand.java

Отрывок: Method "sendChangedNicknameMessage" was implemented by Dmitry Lyukov. Its purpose is send changed nickname message.

Результаты

- Для экспериментов и отладки использовался маленький, но хорошо знакомый автору проект на языке Java:
  - 37 файлов с кодом (не тесты)
  - из них 25 с имплементацией (не интерфейсы)
  - 56 извлечённых методов
    - из них у 26-и названия могут нести полезную информацию (не main, run и т.д.)
- Удалось узнать на естественном языке создателей 85% (22-х) таких методов.
- Проблемы обнаружились
  - при коллизии имён методов в разных файлах, т.к. имена файлов не учитывались
  - при недостаточной полноте вопроса (например, в случае методов changeNickname и sendChangedNicknameMessage)

Будущие исследования

Очевидным шагом к улучшению вопросно-ответной системы является сбор наиболее полной информации о сущностях из различных источников, например:

- документации в коде (для методов и классов)
- констант
- сообщений коммитов, их дат и других параметров
- сообщений поднимаемых исключений
- логов

Каждый из этих пунктов предоставляет довольно большой простор для исследований.

Возможные проблемы

- Ухудшение понимания генерируемого текста нейронной сетью при увеличении его размера
- Время работы
- Сложность выявления ложного ответа
- Нецелесообразность использования при генерации текста некоторых важных данных
- Неприменимость к некоторым проектам