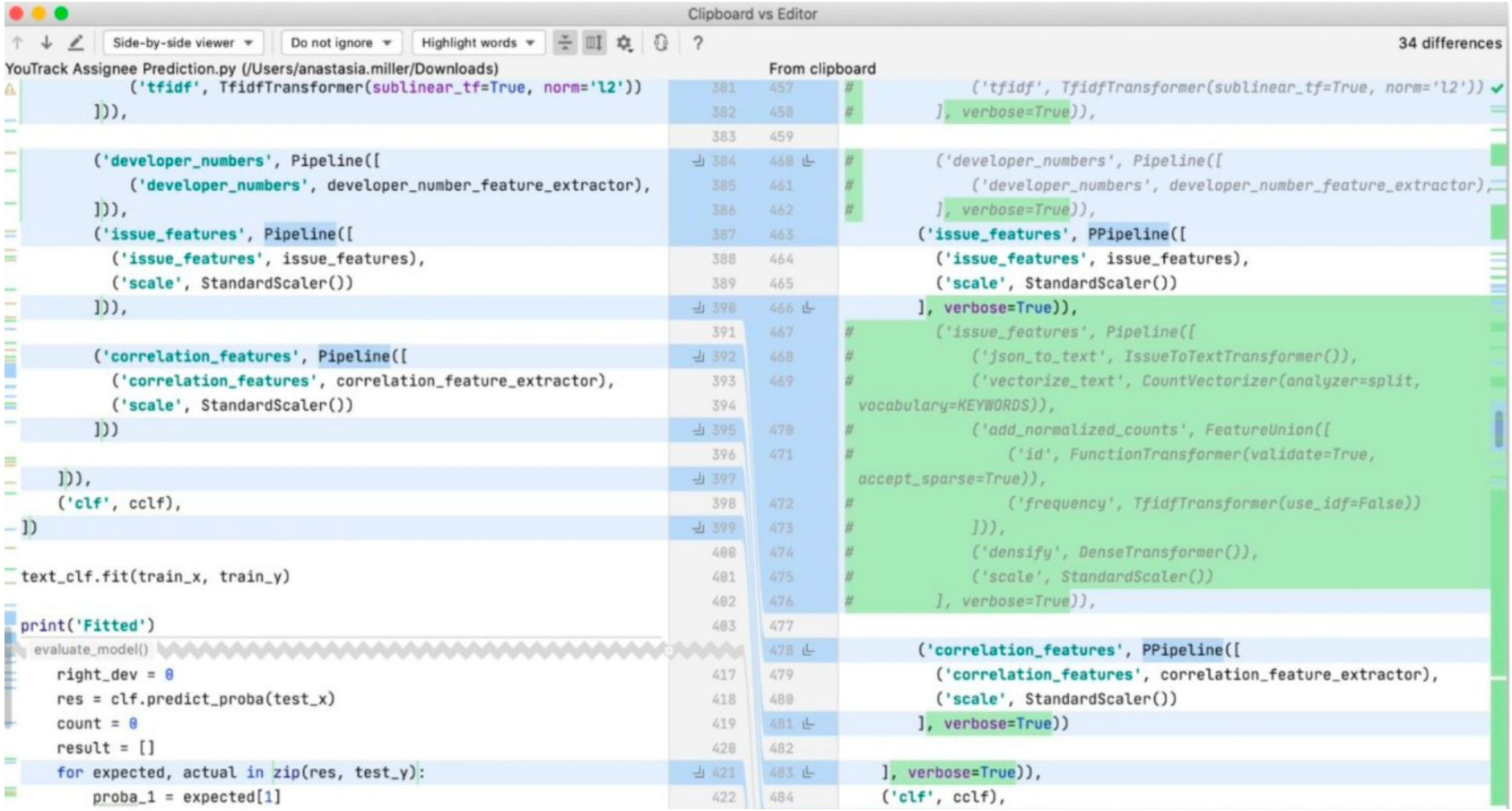
Инструмент для сравнения Jupyter ноутбуков в интегрированной среде разработки DataSpell

Отчет по производственной практике
Божнюк Александр Сергеевич
группа 19.Б11-мм
Консультант: DataSpell разработчик А. В. Вокин
Научный руководитель: ст. преп. С.Ю. Сартасов

Введение

- В наше время все большую популярность набирают ноутбуки Jupyter;
- Компанией JetBrains была создана IDE для Data Science и работы с ноутбуками Jupyter - DataSpell;
- В IntelliJ Platform существует инструмент для сравнения файлов (Diff Tool), который хорошо справляется со своей задачей;
- Однако в случае сравнения ноутбуков Jupyter есть проблема: отображается результат сравнения для JSON представления, вместо обычного представления ноутбуков. Это неудобно для пользователя и приводит к проблемам при работе с инструментом.

Проблема


Clipboard vs Editor
Side-by-side viewer
Do not ignore
Highlight words
34 differences
YouTrack Assignee Prediction.py (/Users/anastasia.miller/Downloads)
From clipboard

Цели и задачи

Целью данной работы является создание инструмента сравнения для ноутбуков Jupyter в интегрированной среде разработки DataSpell.

Для достижения цели были поставлены следующие задачи:

1. Провести обзор существующих подходов к алгоритмам сравнения;
2. Провести обзор конкурентов, имеющих свою реализацию инструмента сравнения файлов;
3. Провести обзор платформы IntelliJ;
4. Реализовать инструмент сравнения файлов (Diff Tool) для Jupyter ноутбуков и интегрировать его в DataSpell. При этом разобраться с такими сопутствующими задачами, как сравнение выводов ячеек и сравнение ячеек с кодом Markdown;
5. Реализовать инструмент слияния (Merge Tool) для интеграции с системами контроля версий.

Обзор конкурентов

Вопросы, которые были рассмотрены во время обзора конкурентов:

- Сравнение ячеек с кодом на языке Markdown;
- Сравнение выводов ячеек;
- Дополнительные возможности.

Рассмотренные конкуренты:

- Библиотека nbdime;
- Visual Studio Code;
- DagsHub;
- Curvenote;
- ReviewNB for GitHub.

Обзор конкурентов: результаты

- Сравнение ячеек с кодом Markdown: отключение обработки текста и последующее сравнение содержимого;
- Сравнение выводов ячеек: принцип было/стало;
- Дополнительные возможности: сравнение метаданных и свертка ячеек, в которых не произошло изменений.

Обзор платформы IntelliJ

Платформа IntelliJ - проект, предоставляющий инфраструктуру для разработки IDE.

Составляющие платформы IntelliJ, которые необходимы для дальнейшего хода работы:

- **Документ (Document):** позволяет работать с содержимым файла как с обычным текстом;
- **Редактор (Editor):** настройка и использование различной функциональности, связанной с IDE;
- **Точки расширения (Extension Points):** расширение функциональности одних плагинов другими.

Основная идея

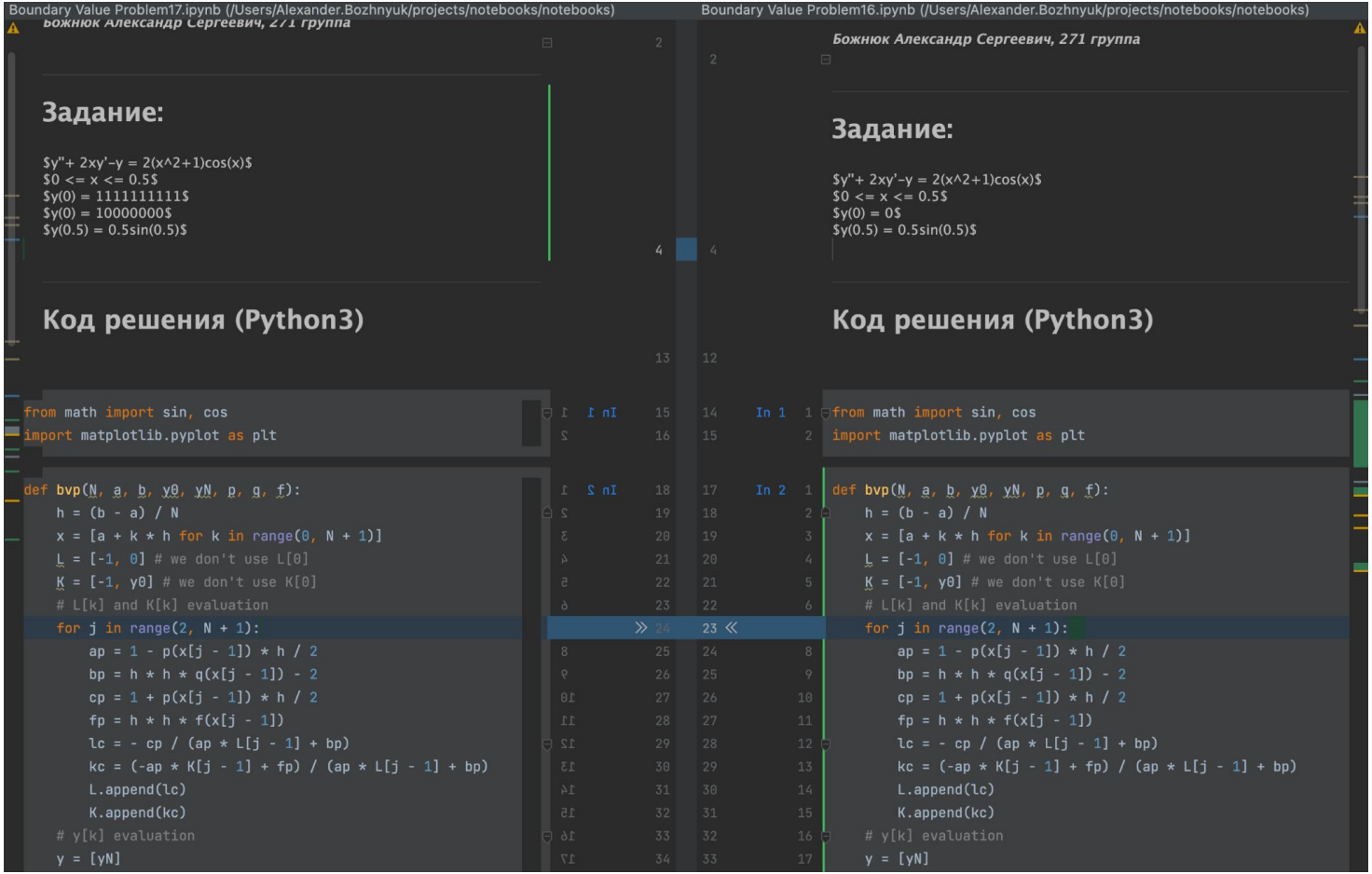
- С каждым файлом ноутбука связаны 2 документа;
- Первый: содержит JSON-представление;
- Второй: содержит “ноутбучное” представление;
- На вход инструмента сравнения приходит первый документ - нужно подменить его на второй.

```
#%% md
***Божнюк Александр Сергеевич, 271 группа***
#%%
from math import sin, cos
import matplotlib.pyplot as plt
```

Инструмент сравнения платформы

- **DiffRequest** - запрос, который содержит в себе информацию для сравнения;
- **DiffViewer** - компонент визуализации результатов сравнения;
- **DiffContext** - контекст, хранящий в себе информацию о текущем проекте и настройках;
- **DiffTool:** на основе **DiffContext** и **DiffRequest** создает **DiffViewer.**
- **Основная цель: подмена запроса, чтобы в нем хранились нужные документы;**
- **Для этого был реализован свой DiffTool: JupyterDiffTool.**

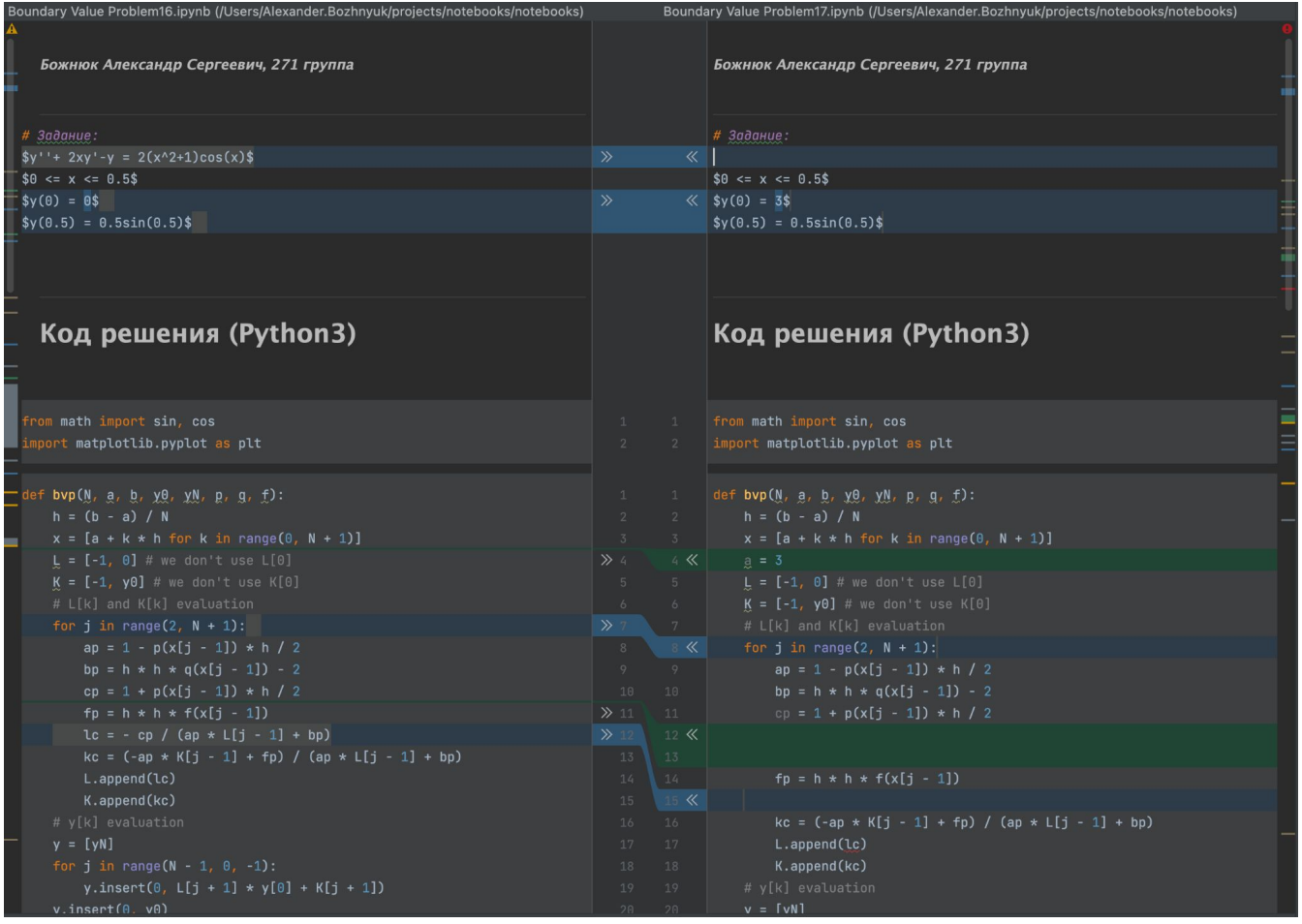
Результат подмены



Исправление визуальных проблем

- Внесение изменений в **модуль визуализации ноутбуков** (требовалось адекватная работа в режиме сравнения);
- **Определение своих DiffViewer** для более гибкой настройки визуализации (например, решение проблемы с сравнением ячеек с кодом на языке Markdown).

Результат после исправлений



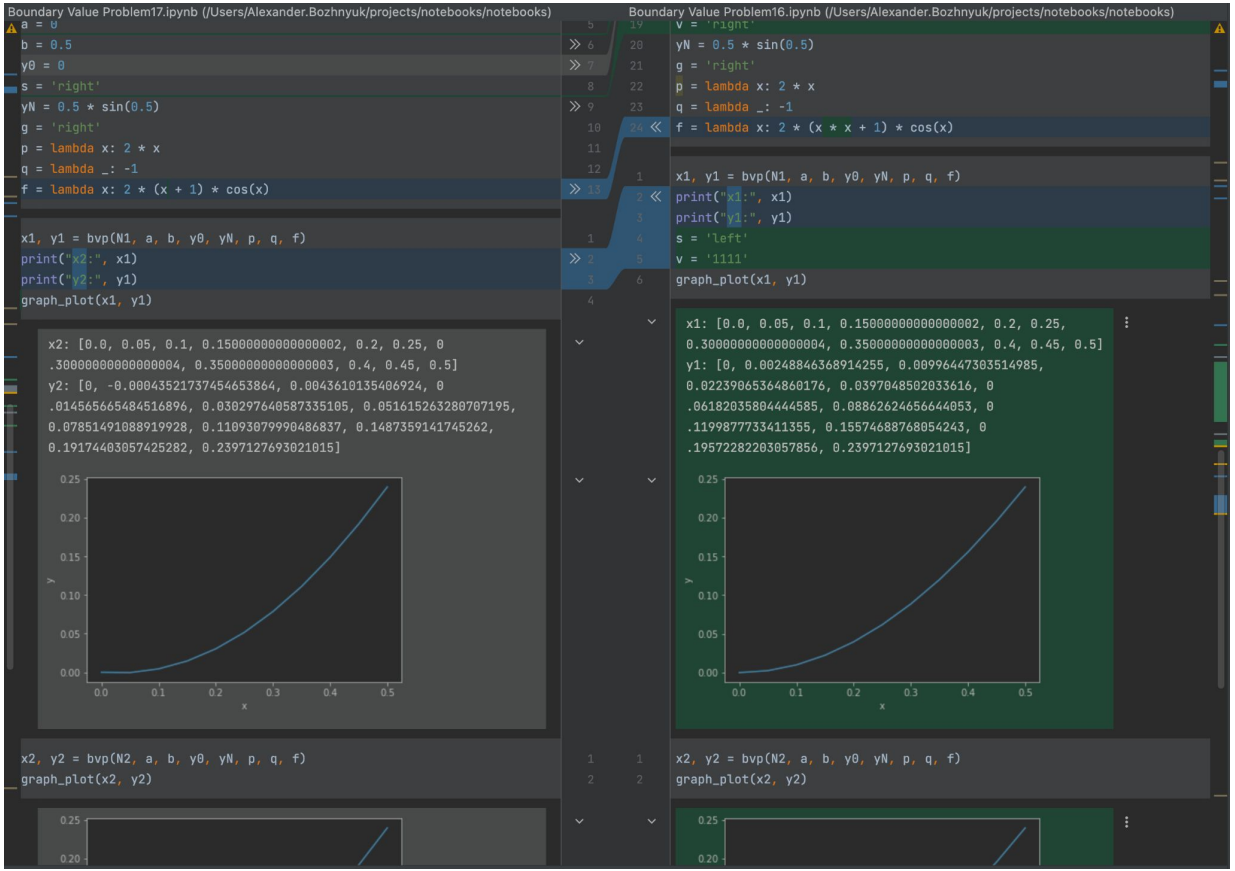
Сравнение выводов ячеек

Для сравнения выводов ячеек требовалось сделать следующее:

- **Извлечение информации** о выводах ячеек в удобном для сравнения виде;
- **Сравнение** извлеченной информации при помощи модуля сравнения IntelliJ. Определение того, какие выводы ячеек были вставлены/удалены/изменены;
- Основываясь на результатах сравнения, **перекрасить необходимые компоненты**, отвечающие за визуализацию выводов ячеек.

Все это было реализовано через переопределенные **DiffViewer**.

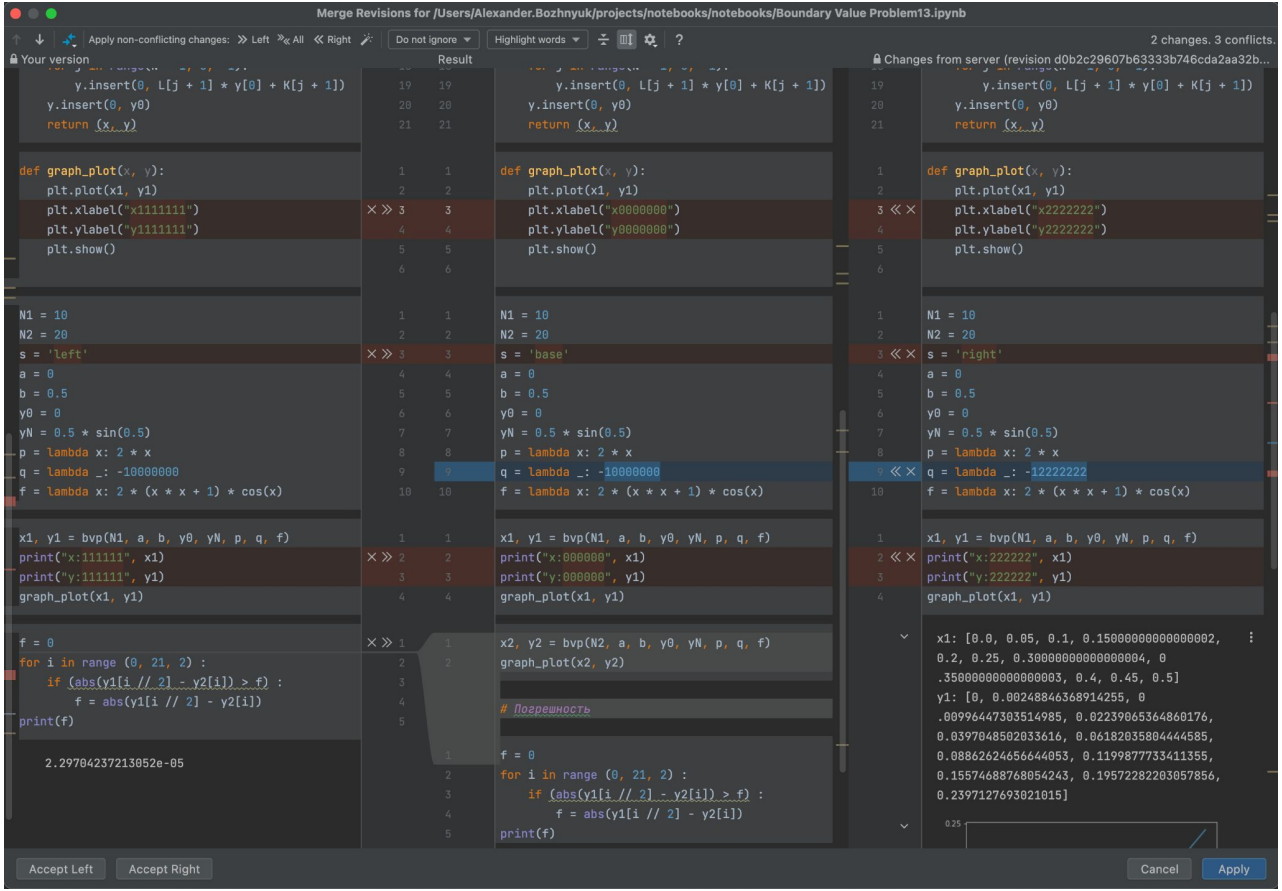
Результат сравнения выводов ячеек



Инструмент слияния

- Основная идея не отличается от **Diff Tool;**
- Реализация своего **MergeTool**: **JupyterMergeTool**, в нем производить подмену **MergeRequest;**
- Реализация своего **MergeView**, для этого требовалось произвести рефакторинг в модуле сравнения платформы. Это позволило внести правки в визуализацию.

Результат: инструмент сравнения



Результаты

1. Проведен обзор проекта Jupyter и алгоритмов для сравнения файлов;
2. Проведен обзор конкурентов;
3. Проведен обзор платформы IntelliJ;
4. Реализован инструмент сравнения ноутбуков. Исправлены многочисленные проблемы, связанные с визуализацией ноутбуков Jupyter в режиме сравнения. Реализованы система сравнения ячеек с кодом на языке Markdown и система сравнения выводов ячеек;
5. Реализован инструмент слияния веток.