

Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико-механический институт

Работа допущена к защите
Руководитель ОП
_____ К.Н. Козлов
« _____ » _____ 2023 г.

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ
РАСПОЗНАВАНИЕ ХИМИЧЕСКОЙ СТРУКТУРЫ МОЛЕКУЛЫ
МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ**

по направлению подготовки 01.04.02 Прикладная математика и информатика
Направленность (профиль) 01.04.02_02 Математические методы анализа и визуализации данных

Выполнил
студент гр. 5040102/10201

К.А. Дамасинский

Руководитель
старший преподаватель ВШПМиВФ

В.С. Чуканов

Консультант
по нормоконтролю

Л.А. Арефьева

Санкт-Петербург
2023

**САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
ПЕТРА ВЕЛИКОГО
Физико-механический институт**

УТВЕРЖДАЮ

Руководитель ОП

_____ К.Н. Козлов

« _____ » _____ 2023г.

**ЗАДАНИЕ
на выполнение выпускной квалификационной работы**

студенту Дамаскинскому Константину Александровичу гр. 5040102/10201

1. Тема работы: Распознавание химической структуры молекулы методами машинного обучения.
2. Срок сдачи студентом законченной работы июнь 2023.
3. Исходные данные по работе
 - Набор молекул, сохранённых в InChI формате, полученный из kaggle соревнования BMS [3.1].
 - Обзорные статьи:
 1. OCR-решения для распознавания молекул: [3.2]
 2. ML-решения для распознавания молекул: [3.2]
 - Инструментальные средства:
 - Языки программирования C++, Python, bash
 - Платформа для машинного обучения Google Cloud
 - Система контроля версий git
- 3.1. Kaggle соревнование BMS. — URL: <https://www.kaggle.com/c/bms-molecular-translation> (дата обращения: 28.05.2023).
- 3.2. *Vogt M.* Using deep neural networks to explore chemical space // Expert Opinion on Drug Discovery. — 2022. — Т. 17, № 3. — С. 297—304.
4. Содержание работы (перечень подлежащих разработке вопросов):
 - 4.1. Введение. Обоснование актуальности
 - 4.2. Постановка задачи

- 4.3. Обзор существующих решений
 - 4.4. Разработка моделей машинного обучения
 - 4.5. Применение
 - 4.6. Результаты и их сравнительный анализ
 - 4.7. Выводы
 - 4.8. Заключение
5. Дата выдачи задания: 13.02.2023.

Руководитель ВКР _____ В.С. Чуканов

Задание принял к исполнению 13.02.2023

Студент _____ К.А. Дамасинский

РЕФЕРАТ

На 27 с., 6 рисунков, 6 таблиц, 2 приложения

КЛЮЧЕВЫЕ СЛОВА: ЗАДАЧА РАСПОЗНАВАНИЯ, ХИМИЯ, INCHI, МАШИННОЕ ОБУЧЕНИЕ.

Тема выпускной квалификационной работы: «Распознавание химической структуры молекулы методами машинного обучения»¹.

Целью данной работы является построение программного продукта, который позволяет распознавать структуру химических, в том числе стереоорганических, молекул. Основным требованием является способность распознавать изображения, полученные из сканов печатной литературы и электронных документов.

При выполнении работы произведён анализ программных продуктов, выполняющих задачу распознавания структуры молекулы, основанных на принципах оптического распознавания знаков (алгоритмический подход) и машинного обучения. Разработано решение в парадигме машинного обучения: построен конвейер из трёх моделей машинного обучения с архитектурами Faster R-CNN, CNN и CNN. В открытых источниках найден набор InChI-последовательностей для обучения и тестирования моделей. Реализован генератор изображений молекул с различными аугментациями, такими как изменение разрешения изображения, толщина и длина линий связей, шрифты атомов, гауссовский шум на изображении. Построен алгоритм, осуществляющий сборку молекулы из распознанных компонентов и валидацию построенного химического соединения. Оценено качество для всех этапов полученного решения. Произведено сравнение качества с существующими решениями. Выполнены замеры производительности программного продукта. Работа реализована на языке программирования Python. Обучение производилось на платформе Google Cloud и на персональном ноутбуке с видеокартой.

ABSTRACT

27 pages, 6 figures, 6 tables, 2 appendices

KEYWORDS: OBJECT DETECTION, CHEMISTRY, INCHI, MACHINE LEARNING.

¹Реферат **должен содержать:** предмет, тему, цель ВКР; метод или методологию проведения ВКР; результаты ВКР; область применения результатов ВКР; выводы.

The subject of the graduate qualification work is «Molecular structure detection by machine learning methods».

In the given work the essence of the approach to creation of a dynamic information portal on the basis of use of open technologies Apache, MySQL and PHP is stated. The general concepts and classification of IT-systems of such class are given. The analysis of systems-prototypes is lead. The technology of creation of the specified class of information systems is investigated. Concrete program realization of a dynamic information portal on an example of a portal of the chosen subjects is developed...

In the given work the essence of the approach to creation of a dynamic information portal on the basis of use of open technologies Apache, MySQL and PHP is stated. The general concepts and classification of IT-systems of such class are given. The analysis of systems-prototypes is lead. The technology of creation of the specified class of information systems is investigated. Concrete program realization of a dynamic information portal on an example of a portal of the chosen subjects is developed...

СОДЕРЖАНИЕ

| | |
|---|----|
| Введение | 7 |
| Глава 1. Название первой главы: всестороннее изучение объекта и предмета исследования, анализ результатов, полученных другими авторами... | 9 |
| 1.1. Название параграфа | 9 |
| 1.1.1. Название первого подпараграфа первого параграфа первой главы для демонстрации переноса слов в содержании..... | 9 |
| 1.2. Название параграфа | 9 |
| 1.3. Выводы | 11 |
| Глава 2. Название второй главы: разработка метода, алгоритма, модели исследования..... | 11 |
| 2.1. Название параграфа | 11 |
| 2.2. Название параграфа | 12 |
| 2.2.1. Название подпараграфа | 12 |
| 2.3. Название параграфа | 13 |
| 2.4. Выводы | 20 |
| Глава 3. Название третьей главы: разработка программного обеспечения.. | 20 |
| 3.1. Название параграфа | 21 |
| 3.2. Название параграфа | 21 |
| 3.3. Выводы | 21 |
| Глава 4. Название четвёртой главы. Апробация результатов исследования, а именно: метода, алгоритма, модели исследования | 21 |
| 4.1. Название параграфа | 21 |
| 4.2. Название параграфа | 21 |
| 4.3. Выводы | 21 |
| Заключение | 22 |
| Список сокращений и условных обозначений | 23 |
| Словарь терминов..... | 24 |
| Список использованных источников..... | 25 |
| Приложение 1. Краткие инструкции по настройке издательской системы \LaTeX | 28 |
| Приложение 2. Некоторые дополнительные примеры | 32 |

ВВЕДЕНИЕ

Задача распознавания структуры химической молекулы является разновидностью задачи компьютерного зрения. До того, как машинное обучение получило широкое распространение, такие задачи чаще всего решали алгоритмическим путём. Так, большинство решений, упомянутых в [6], состояло из этапов предобработки (сглаживание, бинаризация), поиска связей как отрезков прямых и распознавания меток атомов OCR-методами. Решение, основанное на таком подходе, оказывается уязвимым к дефектам изображения, таким как шум, получаемый при сканировании, дефекты печати, следы эксплуатации журналов и книг (мятые или надорванные страницы). Кроме того, при создании алгоритмического решения оказывается гораздо сложнее поддерживать такие параметры, как положения букв на метках атомов, шрифты, расстояние между линиями на двойных и тройных связях, поскольку OCR-методы зачастую параметризуются различного рода порогами погрешностей.

Таким образом, программная поддержка OCR-решений является крайне трудоёмким процессом, поскольку при появлении новых, пусть даже слабо отличающихся от предыдущих, стандартов рендеринга изображений, в программу-распознаватель придётся дописывать блоки кода, поддерживающие эти изменения.

В то же время решение, построенное с помощью методов машинного обучения, даже не видя каких-то характерных особенностей при обучении, зачастую сможет провести корректное распознавание. Если же окажется, что изменение слишком сильное, достаточно добавить некоторое количество изображений с новой особенностью в обучающую выборку и дообучить сеть, что требует весьма небольших затрат ресурсов.

Задача распознавания структуры химической молекулы является частью комплексного решения по распознаванию документа и необходима преимущественно для распознавания научных статей по химии, биологии и медицине. Распознавание документов необходимо для следующих целей:

- Хранение «идеальных» рендеров: пользователь получает возможность читать документы без дефектов. Изображения молекул и другие векторные по своей природе схемы, отпечатанные на бумаге в растровом виде, сохраняются в векторном формате, что позволяет пользователю масштабировать изображение до необходимых ему пределов

- Хранение документов в plain-text формате позволяет осуществлять эффективный поиск информации. В частности, модуль распознавания структуры молекулы позволяет искать научные статьи, содержащие информацию о конкретной молекуле. Данная функциональность помогает фармацевтическим исследователям находить интересные их свойства конкретных веществ, что значительно ускоряет процесс создания лекарственных средств

В рамках данной работы предполагается, что распознавание структуры химической молекулы может производиться как для сканов печатных документов, так и для электронных документов, не содержащих информацию о структуре молекулы в плоском виде.

Таким образом, **исследование является актуальным**, поскольку построение решения, которое позволяет распознавать структуру химической молекулы на разнообразных, поможет создать высококачественный распознаватель документов химической тематики, что в свою очередь значительно улучшит качество и производительность работы исследователей в сфере химии, биологии и фармацевтики.

Целью проводимого исследования является создание описанного ранее решения, замеры его качества, оценка производительности. Также в рамках исследования будет оценена скорость обучения выбранных моделей и *масштабируемость* решения, то есть трудозатраты, требуемые для поддержки дополнительных элементов молекул, таких как RS-стереометки атомов и обозначения смесей различных геометрических изомеров.

ГЛАВА 1. НАЗВАНИЕ ПЕРВОЙ ГЛАВЫ: ВСЕСТОРОННЕЕ ИЗУЧЕНИЕ ОБЪЕКТА И ПРЕДМЕТА ИССЛЕДОВАНИЯ, АНАЛИЗ РЕЗУЛЬТАТОВ, ПОЛУЧЕННЫХ ДРУГИМИ АВТОРАМИ

Хорошим стилем является наличие введения к главе, которое *начинается непосредственно после названия главы, без оформления в виде отдельного параграфа*. Во введении может быть описана цель написания главы, а также приведена краткая структура главы. Например, в параграфе 1.1 приведены примеры оформления одиночных формул, рисунков и таблицы. Параграф 1.2 посвящён многострочным формулам и сложносоставным рисункам.

Текст данной главы призван привести *краткие* примеры оформления текстово-графических объектов. Более подробные примеры можно посмотреть в следующей главе, а также в рекомендациях студентам [5].

1.1. Название параграфа

1.1.1. Название первого подпараграфа первого параграфа первой главы для демонстрации переноса слов в содержании

Содержание первого подпараграфа первого параграфа первой главы.

Одиночные формулы оформляют в окружении `equation`, например, как указано в следующей одиночной нумерованной формуле:

$$\pi \approx 3,141. \quad (1.1)$$

На рис.1.1 изображена гидробашня СПбПУ, а в табл.2.2 приведены данные, на примере которых коротко и наглядно будет изложена суть ВКР.

1.2. Название параграфа

Формулы могут быть размещены в несколько строк. Чтобы выставить номер формулы напротив средней строки, используйте окружение `multlined` из пакета



Рис.1.1. Вид на гидробашню СПбПУ [30]

mathtools следующим образом [20]:

$$\begin{aligned}
 (A_1, B_1) &\leq (A_2, B_2) \Leftrightarrow \\
 &\Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow \\
 &\Leftrightarrow B_2 \subseteq B_1.
 \end{aligned}
 \tag{1.2}$$

Используя команду `\labelcref` из пакета `cleveref`, допустимо следующим образом оформлять ссылку на несколько формул: (1.1 и 1.2). На рис.1.2 приведены три картинки под общим номером и названием, но с отдельной нумерацией подрисунков посредством пакета `subcaption`.

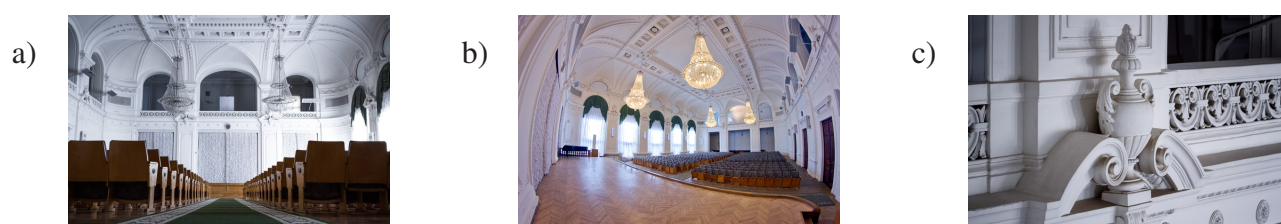


Рис.1.2. Фотографии Белого зала СПбПУ [30], в том числе: *a* — со стороны зрителей; *b* — со стороны сцены; *c* — барельеф

Далее можно ссылаться на три отдельных рисунка: рис.1.2а, рис.1.2б и рис.1.2с.

Пример ссылок [7; 9—11; 14—16; 19; 21—23; 25; 26; 32], а также ссылок с указанием страниц, на котором отображены номера страниц [28, с. 96] или в виде мультицитаты на несколько источников [28, с. 96; 20, с. 46]. Часть библиографических записей носит иллюстративный характер и не имеет отношения к реальной литературе.

1.3. Выводы

Текст выводов по главе 1.

Кроме названия параграфа «выводы» можно использовать (единообразно по всем главам) следующие подходы к именованию последних разделов с результатами по главам:

- «выводы по главе N», где N — номер соответствующей главы;
- «резюме»;
- «резюме по главе N», где N — номер соответствующей главы.

Параграф с изложением выводов по главе *является обязательным*.

ГЛАВА 2. НАЗВАНИЕ ВТОРОЙ ГЛАВЫ: РАЗРАБОТКА МЕТОДА, АЛГОРИТМА, МОДЕЛИ ИССЛЕДОВАНИЯ

Глава посвящена более подробным примерам оформления текстово-графических объектов.

В параграфе 2.1 приведены примеры оформления многострочной формулы и одиночного рисунка. Параграф 2.2 раскрывает правила оформления перечислений и псевдокода. В параграфе 2.3 приведены примеры оформления сложносоставных рисунков, длинных таблиц, а также теоремоподобных окружений.

2.1. Название параграфа

Все формулы, размещенные в отдельных строках, подлежат нумерации, например, как формулы (2.1) и (2.2) из [20].

$$A^\uparrow = \{m \in M \mid gIm \ \forall g \in A\}; \quad (2.1)$$

$$B^\downarrow = \{g \in G \mid gIm \ \forall m \in B\}. \quad (2.2)$$

Обратим внимание, что формулы содержат знаки препинания и что они выровнены по левому краю (с помощью знака & окружения align).

На рис.2.1 приведёна фотография Нового научно-исследовательского корпуса СПбПУ.



Рис.2.1. Новый научно-исследовательский корпус СПбПУ [30]

2.2. Название параграфа

Название параграфа оформляется с помощью команды `\section{...}`, название главы — `\chapter{...}`.

2.2.1. Название подпараграфа

Название подпараграфа оформляется с помощью команды `\subsection{...}`.

Использование подподпараграфов в основной части крайне не рекомендуется. В случае использования, необходимо вынести данный номер в содержание. Название подпараграфа оформляется с помощью команды `\subsubsection{...}`.

Вместо подподпараграфов рекомендовано использовать перечисления.

Перечисления могут быть с нумерационной частью и без неё и использоваться с иерархией и без иерархии. Нумерационная часть при этом формируется следующим способом:

1. в перечислениях *без иерархии* оформляется арабскими цифрами с точкой (или длинным тире).
2. В перечислениях *с иерархией* — в последовательности сначала прописных латинских букв с точкой, затем арабских цифр с точкой и далее — строчных латинских букв со скобкой.

Далее приведён пример перечислений с иерархией.

- A. Первый пункт.
- B. Второй пункт.
- C. Третий пункт.
- D. По ГОСТ 2.105–95 [2] первый уровень нумерации идёт буквами русского или латинского алфавитов (для *определённости* выбираем *английский алфавит*), а второй — цифрами.

1. В данном пункте лежит следующий нумерованный список:

- a) первый пункт;
- b) третий уровень нумерации не нормирован ГОСТ 2.105–95 (для *определенности* выбираем английский алфавит);
- c) обращаем внимание на строчность букв в этом нумерованном и следующем маркированном списке:
 - первый пункт маркированного списка.

Е. Пятый пункт верхнего уровня перечисления.

Маркированный список (без нумерационной части) используется, если нет необходимости ссылки на определенное положение в списке:

- первый пункт с *маленькой буквы* по правилам русского языка;
- второй пункт с *маленькой буквы* по правилам русского языка.

Оформление псевдокода необходимо осуществлять с помощью пакета `algorithm2e` в окружении `algorithm`. Данное окружение интерпретируется в шаблоне как рисунок. Пример оформления псевдокода алгоритма приведён на рис.2.2.

Обратим внимание, что можно сослаться на строчку 1 псевдокода из рис.2.2.

2.3. Название параграфа

Одиночные формулы также, как и отдельные формулы в составе группы, могут быть размещены в несколько строк. Чтобы выставить номер формулы напротив средней строки, используйте окружение `multlined` из пакета `mathtools` следующим образом [20]:

$$\begin{aligned}
 (A_1, B_1) &\leq (A_2, B_2) \Leftrightarrow \\
 &\Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow \\
 &\Leftrightarrow B_2 \subseteq B_1.
 \end{aligned}
 \tag{2.3}$$

Используя команду `\labelcref{...}` из пакета `cleveref`, допустимо оформить ссылку на несколько формул, например, (2.1–2.3).

Пример оформления четырёх иллюстраций в одном текстово-графическом объекте приведён на рис.2.3. Это возможно благодаря использованию пакета `subcaption`.

Algorithm

Input: the many-valued context $M \stackrel{\text{def}}{=} (G, M, W, J)$, the class membership $\varepsilon : G \rightarrow K$

Output: positive and negative binary contexts $\overline{K}_+ \stackrel{\text{def}}{=} (\overline{G}_+, M, I_+)$, $\overline{K}_- \stackrel{\text{def}}{=} (\overline{G}_-, M, I_-)$ such that i-tests found in \overline{K}_+ are diagnostic tests in M , and objects from \overline{K}_- are counter-examples

```

1. for  $\forall g_i, g_j \in G$  do
2.   if  $i < j$  then
3.      $\overline{G} \leftarrow (g_i, g_j);$ 
4.   for  $\forall (g_i, g_j) \in \overline{G}$  do
5.     if  $m(g_i) = m(g_j)$  then
6.        $(g_i, g_j)Im;$ 
7.     if  $\varepsilon(g_i) = \varepsilon(g_j)$  then
8.        $\overline{G}_+ \leftarrow (g_i, g_j);$ 
9.     else  $\overline{G}_- \leftarrow (g_i, g_j);$ 
10.   $I_+ = I \cap (\overline{G}_+ \times M), I_- = I \cap (\overline{G}_- \times M);$ 
11.  for  $\forall \overline{g}_+ \in \overline{G}_+, \forall \overline{g}_- \in \overline{G}_-$  do
12.    if  $\overline{g}_+^\uparrow \subseteq \overline{g}_-^\uparrow$  then
13.       $\overline{G}_+ \leftarrow \overline{G}_+ \setminus \overline{g}_+;$ 

```

Рис.2.2. Псевдокод алгоритма DiagnosticTestsScalingAndInferring [28]

Далее можно ссылаться на составные части данного рисунка как на самостоятельные объекты: рис.2.3а, рис.2.3б, рис.2.3с, рис.2.3д или на три из четырёх изображений одновременно: рис.2.3а–2.3с.

Приведём пример табличного представления данных с записью продолжения на следующей странице на табл.2.1.

Таблица 2.1

Пример задания данных из [29] (с повтором для переноса таблицы на новую страницу)

| G | m_1 | m_2 | m_3 | m_4 | K |
|-------|-------|-------|-------|-------|-----|
| 1 | 2 | 3 | 4 | 5 | 6 |
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |

Продолжение табл.2.1

| 1 | 2 | 3 | 4 | 5 | 6 |
|-------|---|---|---|---|---|
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |

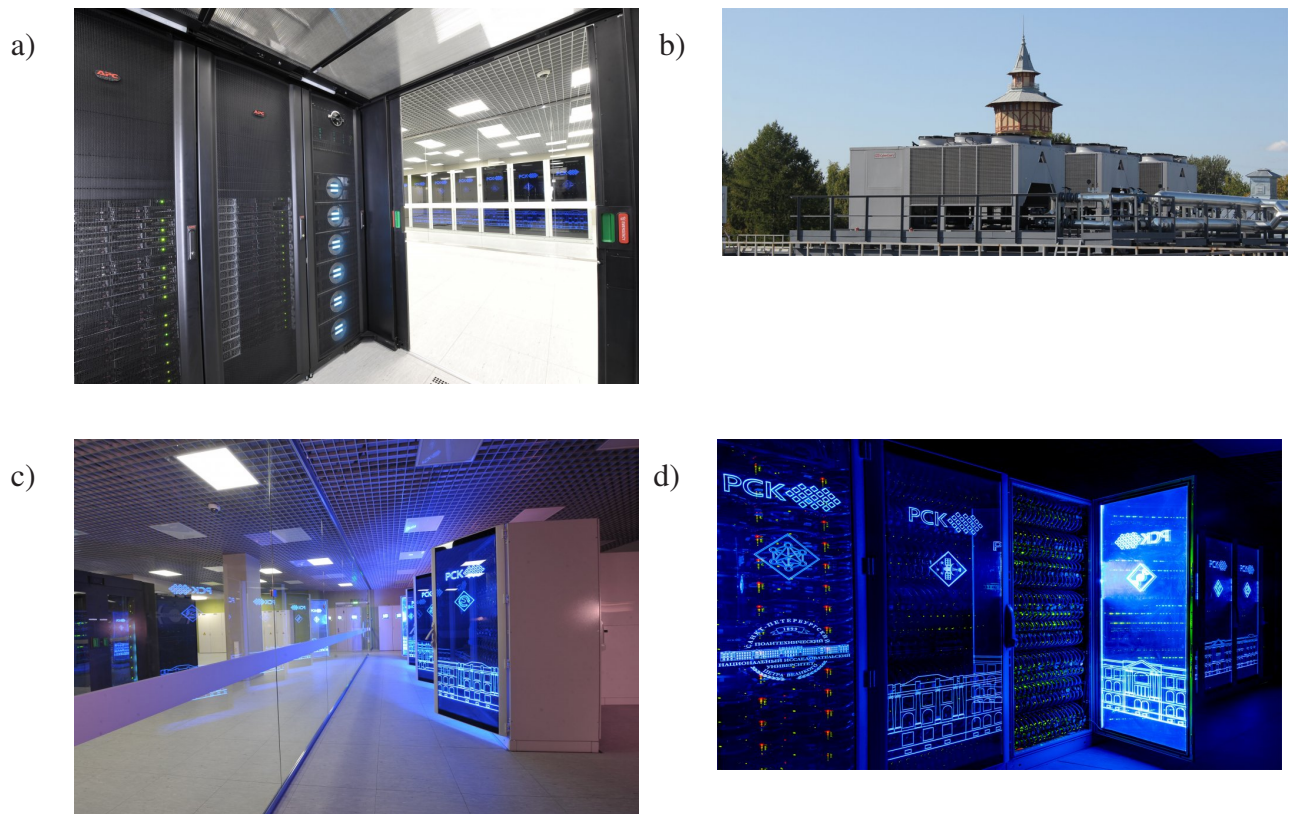


Рис.2.3. Фотографии суперкомпьютерного центра СПбПУ [30]: *a* — система хранения данных и узлы NUMA-вычислителя; *b* — холодильные машины на крыше научно-исследовательского корпуса; *c* — машинный зал; *d* — элементы вычислительных устройств

Таблица 2.2

Пример представления данных для сквозного примера по ВКР [29]

| G | m_1 | m_2 | m_3 | m_4 | K |
|-------|-------|-------|-------|-------|-----|
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |

Таблица 2.3

Пример задания данных в табличном виде из [29] (с помощью окружения minipage)

| G | m_1 | m_2 | m_3 | m_4 | K |
|-------|-------|-------|-------|-------|-----|
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |



Рис.2.4. Новый научно-исследовательский корпус СПбПУ [30] (с помощью окружения minipage)

Вопросы форматирования текстово-графических объектов (окружений) не регламентированы в известных нам ГОСТах, поэтому предлагаем придерживаться следующих правил:

- **полужирный текст** рекомендуем использовать только для названий стандартных окружений с нумерационной частью, например, для представления *впервые*: **определение 1.1, теорема 2.2, пример 2.3, лемма 4.5**;
- *курсив* рекомендуем использовать только для выделения переменных в формулах, служебной информации об авторах главы (статьи), важных терминов, представляемых по тексту, а также для всего тела окружений, связанных с получением *новых существенных результатов и их доказательством*: теорема, лемма, следствие, утверждение и другие.

По аналогии с нумерацией формул, рисунков и таблиц нумеруются и иные текстово-графические объекты, то есть включаем в нумерацию номер главы, например: теорема 3.1. для первой теоремы третьей главы монографии. Команды \LaTeX выставляют нумерацию и форматирование автоматически. Полный перечень команд для подготовки текстово-графических и иных объектов находится в подробных методических рекомендациях [8].

Для удобства авторов названия стандартных окружений, рекомендованных к использованию, приведены в табл.2.4, а в табл.2.5 перечислены имена специально разработанных окружений для шаблонов SPbPU.

На базе пакета `tikz` разработано большое количество расширений [17], например, `tikzcd`, которые мы рекомендуем использовать для оформления иллюстраций.

В случае, если авторам потребовалось новое окружение, то создать его можно в файле `my_folder/my_settings.tex` согласно правилам, приведённым ниже.

Таблица 2.4

Стандартные окружения

| Название окружения | Назначение |
|--------------------|--|
| center | центрирование, аналог команды <code>\centering</code> , но с добавлением нежелательного пробела, поэтому лучше избегать применения <code>center</code> |
| itemize | перечисления, в которых нет необходимости нумеровать пункты (немаркированные списки) |
| enumerate | перечисления с нумерацией (немаркированные списки) |
| refsection | создание отдельных библиографических списков для глав |
| tabular | оформление таблиц |
| table | автоматическое перемещение по тексту таблиц, оформленных, например, с помощью <code>tabular</code> , для минимизации пустых пространств |
| longtable | оформление многостраничных таблиц |
| tikzpicture | создание иллюстраций с помощью пакета <code>tikz</code> [17] |
| figure | автоматическое перемещение по тексту рисунков, оформленных например, с помощью <code>tikz</code> или подключенных с помощью команды <code>\includegraphics</code> , для минимизации пустых пространств |
| subfigure | оформление вложенных рисунков в составе <code>figure</code> |
| algorithm | оформление псевдокода на основе пакета <code>algorithm2e</code> [18] |
| minipage | оформление рисунков и таблиц без функций автоматического перемещения по тексту для минимизации пустых пространств |
| equation | оформление выключенных (не встроенных в текст с помощью <code>\$...\$</code>) одиночных формул на одной строке |
| multilined | оформление выключенных (не встроенных в текст с помощью <code>\$...\$</code>) одиночных формул в несколько строк |
| aligned | оформление нескольких формул с выравниванием по символу <code>&</code> . |

Таблица 2.5

Специальные окружения

| Название окружения | Текстово-графический объект |
|--------------------|-----------------------------|
| abstr | реферат (abstract) |
| m-theorem | теорема |
| m-corollary | следствие |
| m-proposition | утверждение |
| m-lemma | лемма |
| m-axiom | аксиома |
| m-example | пример |
| m-definition | определение |
| m-condition | условие |
| m-problem | проблема |
| m-exercise | упражнение |
| m-question | вопрос |
| m-hypothesis | гипотеза |

1. Для перехода в режим создания окружений следует указать:
 - `\theoremstyle{myplain}` — окружения с доказательствами или аксиомами
 - `\theoremstyle{mydefinition}` — окружения, не связанные с доказательствами или аксиомами.
2. В команде создания окружения следует ввести краткий псевдоним (`m-new-env`) и отображаемое в pdf имя окружения (Название_окружения):
 - `\newtheorem{m-new-env-second}{Название_окружения} - [chapter]`.

Теорема 2.1 (о чем-то конкретном). *Текст теоремы полностью выделен курсивом. Допустимо математические символы не выделять курсивом, если это искажает их значения. Используется абзацный отступ, так как “Абзацы в тексте начинают отступом” в соответствии с ГОСТ 2.105–95. Название теоремы допустимо убрать. Доказательство окончено.*

Доказательство теоремы 2.1, леммы, утверждений, следствий и других подобных окружений (в последнем абзаце) завершаем предложением в котором сказано, что доказательство окончено. Например, доказательство теоремы 2.1 окончено.

Тело доказательства не выделяется курсивом. Тело следующих окружений также не выделяется сплошным курсивом: определение, условие, проблема, пример, упражнение, вопрос, гипотеза и другие.

Определение 2.1 (термин). В тексте определения только *важные термины* выделяются курсивом. Если определение носит лишь вспомогательный характер, то допустимо не использовать окружение `m-definition`, представляя текст определения в обычном абзаце. Ключевые термины при этом обязательно выделяются курсивом.

Вместо теоремо-подобных окружений для вставки небольших текстово-графических объектов иногда используются команды. Типичным примером такого подхода является команда `\footnote{text}`², где в аргументе `text` указывают текст *подстрочной ссылки (сноски)*. В них *нельзя добавлять веб-ссылки или цитировать литературу*. Для этих целей используется список литературы. Нумерация сносок сквозная по ВКР без точки на конце выставляется в шаблоне автоматически, однако в каждом приложении к ВКР нумерация, зависящая от

²Внимание! Команда вставляется непосредственно после слова, куда вставляется сноска (без пробела). Лишние пробелы также не указываются внутри команды перед и после фигурных скобок.

номера приложения, выставляется префикс «П», например «П1.1» — первая сноска первого приложения.

2.4. Выводы

Текст заключения ко второй главе. Пример ссылок [7; 9—11; 14—16; 19; 21—23; 25; 26; 32], а также ссылок с указанием страниц, на котором отображены те или иные текстово-графические объекты [28, с. 96] или в виде мультицитаты на несколько источников [28, с. 96; 20, с. 46]. Часть библиографических записей носит иллюстративный характер и не имеет отношения к реальной литературе.

Короткое имя каждого библиографического источника содержится в специальном файле `my_biblio.bib`, расположенном в папке `my_folder`. Там же находятся исходные данные, которые с помощью программы `Biber` и стилевого файла `Biblatex-GOST` [12] приведены в списке использованных источников согласно ГОСТ 7.0.5-2008. Многообразные реальные примеры исходных библиографических данных можно посмотреть по ссылке [13].

Как правило, ВКР должна состоять из четырех глав. Оставшиеся главы можно создать по образцу первых двух и подключить с помощью команды `\input` к исходному коду ВКР. Далее в приложении 1 приведены краткие инструкции запуска исходного кода ВКР [27; 31].

В приложении 2 приведено подключение некоторых текстово-графических объектов. Они оформляются по приведенным ранее правилам. В качестве номера структурного элемента вместо номера главы используется «П» с номером главы. Текстово-графические объекты из приложений не учитываются в реферате.

ГЛАВА 3. НАЗВАНИЕ ТРЕТЬЕЙ ГЛАВЫ: РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Хорошим стилем является наличие введения к главе. Во введении может быть описана цель написания главы, а также приведена краткая структура главы.

3.1. Название параграфа

3.2. Название параграфа

3.3. Выводы

Текст выводов по главе 3.

ГЛАВА 4. НАЗВАНИЕ ЧЕТВЁРТОЙ ГЛАВЫ. АПРОБАЦИЯ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ, А ИМЕННО: МЕТОДА, АЛГОРИТМА, МОДЕЛИ ИССЛЕДОВАНИЯ

Хорошим стилем является наличие введения к главе. Во введении может быть описана цель написания главы, а также приведена краткая структура главы.

4.1. Название параграфа

4.2. Название параграфа

Пример ссылки на литературу [1; 3; 4; 24].

4.3. Выводы

Текст выводов по главе 4.

ЗАКЛЮЧЕНИЕ

Заключение (2 – 5 страниц) обязательно содержит выводы по теме работы, *конкретные предложения и рекомендации* по исследуемым вопросам. Количество общих выводов должно вытекать из количества задач, сформулированных во введении выпускной квалификационной работы.

Предложения и рекомендации должны быть органически увязаны с выводами и направлены на улучшение функционирования исследуемого объекта. При разработке предложений и рекомендаций обращается внимание на их обоснованность, реальность и практическую приемлемость.

Заключение не должно содержать новой информации, положений, выводов и т. д., которые до этого не рассматривались в выпускной квалификационной работе. Рекомендуются писать заключение в виде тезисов.

Последним абзацем в заключении можно выразить благодарность всем людям, которые помогали автору в написании ВКР.

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

DOI Digital Object Identifier.

WoS Web of Science.

ВКР Выпускная квалификационная работа.

ТГ-объект Текстово-графический объект.

СЛОВАРЬ ТЕРМИНОВ

TeX — язык вёрстки текста и издательская система, разработанные Дональдом Кнутом.

LaTeX — язык вёрстки текста и издательская система, разработанные Лэсли Лампортом как надстройка над TeX.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Автономова Н. С. Философский язык Жака Деррида. — М.: Российская политическая энциклопедия (РОССПЭН), 2011. — 510 с. — (Сер.: Российские Пропилеи).
2. Единая система конструкторской документации. Общие требования к текстовым документам [Текст]: ГОСТ 2.105–95. — Взамен ГОСТ 2.105—79, ГОСТ 2.906—71 ; введ. 1996—07—01. — Минск : Межгос. совет по стандартизации, метрологии и сертификации, 2002. — 31 с. — (Сер.: Межгосударственный стандарт).
3. Котельников И. А., Чеботаев П. З. LaTeX по-русски. — 3-е изд. — Новосибирск: Сибирский Хронограф, 2004. — 496 с. — URL: <http://www.tex.uniyar.ac.ru/doc/kotelnikovchebotaev2004b.pdf> (дата обращения: 06.03.2019).
4. Песков Н. В. Поиск информативных фрагментов описаний объектов в задачах распознавания: дис. . . . канд. физ.-мат. наук: 05.13.17 / Песков Николай Владимирович. — М., 2004. — 102 с.
5. Руководство студента СПбПУ по подготовке выпускной квалификационной работы и сопутствующих документов с помощью LaTeX / В. А. Пархоменко [и др.]. — 2018. — URL: https://github.com/ParkhomenkoV/SPbPU-student-thesis-template/blob/master/Author_guide_SPbPU-student-thesis.pdf (дата обращения: 06.03.2019).
6. A review of optical chemical structure recognition tools / K. Rajan [и др.] // Journal of Cheminformatics. — 2020. — Т. 12, № 1. — С. 1—13.
7. Adams P. The title of the work // The name of the journal. — 1993. — Vol. 4, no. 2. — P. 201–213.
8. Author and editor guide to prepare and submit the academic SPbPU editions to Clarivate Analytics: Book Citation Index Web of Science / V. Parkhomenko [et al.]. — 2018. — URL: https://github.com/ParkhomenkoV/SPbPU-BCI-template/blob/master/Author_guide_SPbPU-BCI.pdf (visited on 06.03.2019).
9. Babington P. The title of the work. Vol. 4. — 3rd ed. — The address: The name of the publisher, 1993. — 255 p. — (Ser.: 10).
10. Badiou A. Briefings on Existence: A Short Treatise on Transitory Ontology / ed. and trans. from the French, with an introd., by N. Madarasz. — NY: SUNY Press, 2006. — 190 p. — URL: https://books.google.ru/books?id=7HNkAT%5C_NFksC (visited on 05.12.2017).

11. *Caxton P.* The title of the work. — The address of the publisher, 1993. — 255 p.
12. *Domanov O.* BibLATEX support for GOST standard bibliographies. — URL: <https://ctan.org/pkg/biblatex-gost> (visited on 06.03.2019).
13. *Domanov O.* Biblatex-GOST examples. — URL: <http://ctan.altspu.ru/macros/latex/contrib/biblatex-contrib/biblatex-gost/doc/biblatex-gost-examples.pdf> (visited on 06.03.2019).
14. *Draper P.* The title of the work // The title of the book. Vol. 4 / ed. by T. editor. — The organization. The address of the publisher: The publisher, 1993. — (Ser.: 5).
15. *Eston P.* The title of the work // Book title. Vol. 4. — 3rd ed. — The address of the publisher: The name of the publisher, 1993. — Chap. 8 — P. 201–213. — (Ser.: 5).
16. *Farindon P.* The title of the work // The title of the book. Vol. 4 / ed. by T. editor. — 3rd ed. — The address of the publisher: The name of the publisher, 1993. — Chap. 8 — P. 201–213. — (Ser.: 5).
17. *Feuersanger C., Tantau T.* The TikZ and PGF packages. — URL: <https://ctan.org/pkg/pgf> (visited on 06.03.2019).
18. *Fiorio C.* The algorithm2e package. — URL: <https://ctan.org/pkg/algorithm2e> (visited on 06.03.2019).
19. *Gainsford P.* The title of the work / The organization. — 3rd ed. — The address of the publisher, 1993. — 255 p.
20. *Ganter B., Wille R.* Formal concept analysis: mathematical foundations. — Springer, Berlin, 1999. — 284 p.
21. *Harwood P.* The title of the work: Master's thesis / Harwood Peter. — The address of the publisher: The school where the thesis was written, 1993. — 255 p.
22. *Isley P.* The title of the work. — 1993.
23. *Joslin P.* The title of the work: diss. ... PhD in Engineering / Joslin Peter. — The address of the publisher: The school where the thesis was written, 1993. — 255 p.
24. *Kotelnikov I. A., Chebotaev P. Z.* LaTeX in Russian. — 3rd ed. — Novosibirsk: Sibiskiy Hronograph, 2004. — 496 p. — URL: <http://www.tex.uniyar.ac.ru/doc/kotelnikovchebotaev2004b.pdf> (visited on 06.03.2019); (in Russian).
25. *Lambert P.* The title of the work: tech. rep. / The institution that published. — The address of the publisher, 1993. — 255 p. — No. 2.
26. *Marcheford P.* The title of the work. — 1993.
27. MiKTeX web site. — URL: <https://miktex.org/> (visited on 06.03.2019).

28. Notes on relation between symbolic classifiers / X. Naidenova [et al.] // CEUR Workshop Proceedings / ed. by K. S. Watson B.W. — 2017. — Vol. 1921. — P. 88–103. — URL: <http://ceur-ws.org/Vol-1921/paper9.pdf> (visited on 19.12.2017).

29. *Peskov N. V.* Searching for informative fragments of object descriptions in the recognition tasks: diss. ... cand. phys.-math. sci.: 05.13.17 / Peskov Nickolay Vladimirovich. — M., 2004. — 102 p. — (in Russian).

30. SPbPU photo gallery. — URL: <http://www.spbstu.ru/media/photo-gallery/> (visited on 06.03.2019).

31. TeXstudio web site. — URL: <https://www.texstudio.org/> (visited on 06.03.2019).

32. The title of the work. Vol. 4 / ed. by P. Kidwelly. — The organization. The address of the publisher: The name of the publisher, 1993. — 255 p. — (Ser.: 5).

Приложение 1

Краткие инструкции по настройке издательской системы L^AT_EX

В SPbPU-BCI-template автоматически выставляются необходимые настройки и в исходном тексте шаблона приведены примеры оформления текстово-графических объектов, поэтому авторам достаточно заполнить имеющийся шаблон текстом главы (статьи), не вдаваясь в детали оформления, описанные далее. Возможный «быстрый старт» оформления главы (статьи) под Windows следующий^{П1.1}:

- A. Установка полной версии MikTeX [27]. В процессе установки лучше выставить параметр доустановки пакетов «на лету».
- B. Установка TexStudio [31].
- C. Запуск TexStudio и компиляция `my_chapter.tex` с помощью команды «Build&View» (например, с помощью двойной зелёной стрелки в верхней панели). Иногда, для достижения нужного результата необходимо несколько раз скомпилировать документ.
- D. В случае, если не отобразилась библиография, можно
 - воспользоваться командой Tools → Commands → Biber, затем запустив Build&View;
 - настроить автоматическое включение библиографии в настройках Options → Configure TexStudio → Build → Build&View (оставить по умолчанию, если сборка происходит слишком долго): `txs:///pdflatex | txs:///biber | txs:///pdflatex | txs:///pdflatex | txs:///view-pdf`.

В случае возникновения ошибок, попробуйте скомпилировать документ до последних действий или внимательно ознакомьтесь с описанием проблемы в log-файле. Бывает полезным переход (по подсказке TexStudio) в нужную строку в pdf-файле или запрос с текстом ошибки в поисковиках. Наиболее вероятной проблемой при первой компиляции может быть отсутствие какого-либо установленного пакета L^AT_EX.

В случае корректной работы настройки «установка на лету» все дополнительные пакеты будут скачиваться и устанавливаться в автоматическом режиме. Если доустановка пакетов осуществляется медленно (несколько пакетов за один запуск

^{П1.1} Вниманию! Пример оформления подстрочной ссылки (сноски).

компилятора), то можно попробовать установить их в ручном режиме следующим образом:

1. Запустите программу: меню → все программы → MikTeX → Maintenance (Admin) → MiKTeX Package Manager (Admin).
2. Пользуясь поиском, убедитесь, что нужный пакет присутствует, но не установлен (если пакет отсутствует воспользуйтесь сначала MiKTeX Update (Admin)).
3. Выделив строку с пакетом (возможно выбрать несколько или вообще все неустановленные пакеты), выполните установку Tools → Install или с помощью контекстного меню.
4. После завершения установки запустите программу MiKTeX Settings (Admin).
5. Обновите базу данных имен файлов Refresh FNDB.

Для проверки текста статьи на русском языке полезно также воспользоваться настройками Options → Configure TexStudio → Language Checking → Default Language. Если русский язык «ru_RU» не будет доступен в меню выбора, то необходимо вначале выполнить Import Dictionary, скачав из интернета любой русскоязычный словарь.

Далее приведены формулы (П1.2), (П1.1), рис.П1.2, рис.П1.1, табл.П1.2, табл.П1.1.

$$\pi \approx 3,141. \quad (\text{П1.1})$$



Рис.П1.1. Вид на гидробашню СПбПУ [30]

Представление данных для сквозного примера по ВКР [29]

| G | m_1 | m_2 | m_3 | m_4 | K |
|-------|-------|-------|-------|-------|-----|
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |

П1.1. Параграф приложения

П1.1.1. Название подпараграфа

Название подпараграфа оформляется с помощью команды `\subsection{...}`.
Использование подподпараграфов в основной части крайне не рекомендуется.

П1.1.1.1. Название подподпараграфа

$$\pi \approx 3,141. \quad (\text{П1.2})$$



Рис.П1.2. Вид на гидробашню СПбПУ [30]

Таблица П1.2

Представление данных для сквозного примера по ВКР [29]

| G | m_1 | m_2 | m_3 | m_4 | K |
|-------|-------|-------|-------|-------|-----|
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |

Приложение 2

Некоторые дополнительные примеры

В приложении^{П2.1} приведены формулы (П2.2), (П2.1), рис.П2.2, рис.П2.1, табл.П2.2, табл.П2.1

$$\pi \approx 3,141.$$

(П2.1)



Рис.П2.1. Вид на гидробашню СПбПУ [30]

Таблица П2.1

Представление данных для сквозного примера по ВКР [29]

| G | m_1 | m_2 | m_3 | m_4 | K |
|-------|-------|-------|-------|-------|-----|
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |

^{П2.1}Внимание! Пример оформления подстрочной ссылки (сноски).

П2.1. Подраздел приложения

$$\pi \approx 3,141.$$

(П2.2)



Рис.П2.2. Вид на гидробашню СПбПУ [30]

Таблица П2.2

Представление данных для сквозного примера по ВКР [29]

| G | m_1 | m_2 | m_3 | m_4 | K |
|-------|-------|-------|-------|-------|-----|
| g_1 | 0 | 1 | 1 | 0 | 1 |
| g_2 | 1 | 2 | 0 | 1 | 1 |
| g_3 | 0 | 1 | 0 | 1 | 1 |
| g_4 | 1 | 2 | 1 | 0 | 2 |
| g_5 | 1 | 1 | 0 | 1 | 2 |
| g_6 | 1 | 1 | 1 | 2 | 2 |