МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО

ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное

образовательное учреждение высшего образования

«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ

ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ»

Факультет информатики и прикладной математики

Кафедра прикладной математики и экономико-математических методов

**КУРСОВАЯ РАБОТА**

по дисциплине:

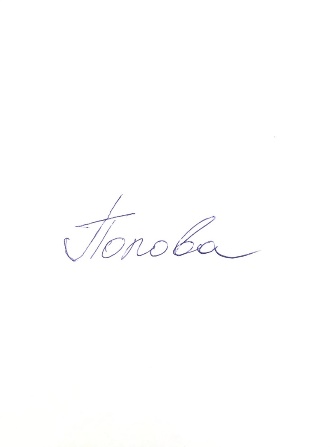
**«Системы компьютерной математики»**

Тема:

Создание обучающего курса «Алгоритмы кластеризации» в системе Wolfram Mathematica

Направление: 01.03.02 Прикладная математика и информатика

Студент: Попова Софья Ивановна

Группа: ПМ-1901 Подпись: 

Проверил: Фридман Григорий Морицович

Должность: Доктор технических наук, профессор

Оценка: Дата:

Подпись:

Санкт-Петербург

2020

СОДЕРЖАНИЕ

[**ВВЕДЕНИЕ** 3](#_Toc53051218)

[**1.** **ОБЗОР ОБУЧАЮЩЕГО КУРСА** 4](#_Toc53051219)

[**1.1 Основное меню** 4](#_Toc53051220)

[**1.2 Теоретический материал** 5](#_Toc53051221)

[**2.** **АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ** 9](#_Toc53051222)

[**2.1 Иерархические алгоритмы** 11](#_Toc53051223)

[**2.2 K–Means алгоритм** 12](#_Toc53051224)

[**2.3 Алгоритм минимально покрывающего дерева** 14](#_Toc53051225)

[**2.4 Алгоритм – метод ближайшего соседа** 16](#_Toc53051226)

[**2.5 Тестирование** 18](#_Toc53051227)

[**ЗАКЛЮЧЕНИЕ** 20](#_Toc53051228)

[**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ** 21](#_Toc53051229)

# **ВВЕДЕНИЕ**

Благодаря своему широкому применению, теория о кластеризации данных в последнее время интенсивно развивается.

У кластеризации существует большое количество практических применений как в информатике, так и в других областях. Примерами применения могут служить:

* Анализ данных;
* Извлечение и поиск информации;
* Группировка и распознавание объектов.

Кроме того, кластеризация является бурно развивающимся разделом современной теоретической информатики и в этой области можно получить ряд интересных исследовательских результатов.

Существует множество известных алгоритмов кластеризации. Такие как:

* иерархические алгоритмы;
* k - means алгоритм;
* минимальное покрывающее дерево;
* метод ближайшего соседа;
* алгоритмы нечеткой кластеризации;
* применение нейронных сетей;
* генетические алгоритмы;
* метод закалки.

Цель работы: разработать обучающий курс, используя возможности Wolfram Mathematica [2].

# **ОБЗОР ОБУЧАЮЩЕГО КУРСА**

## **1.1 Основное меню**

При открытии курса пользователь видит главную страницу обучающего курса, на которой находятся название и основное меню. (Рисунок 1)

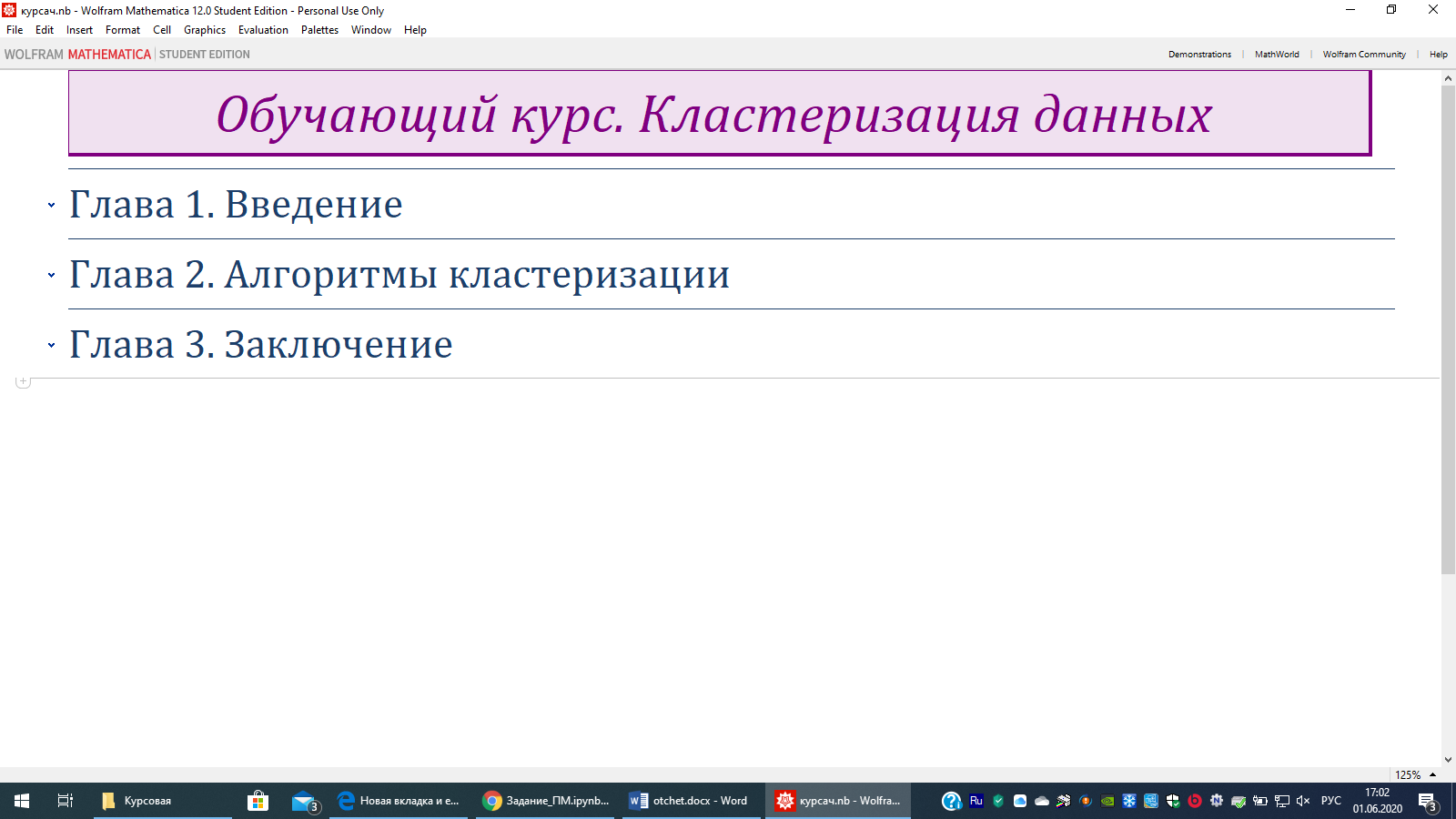


Рисунок 1 - Основное меню интерактивного курса

Далее при нажатии на любую из трех глав открывается ее содержимое. Например, первая глава содержит теоретический материал, разделенный на два раздела. (Рисунок 2)

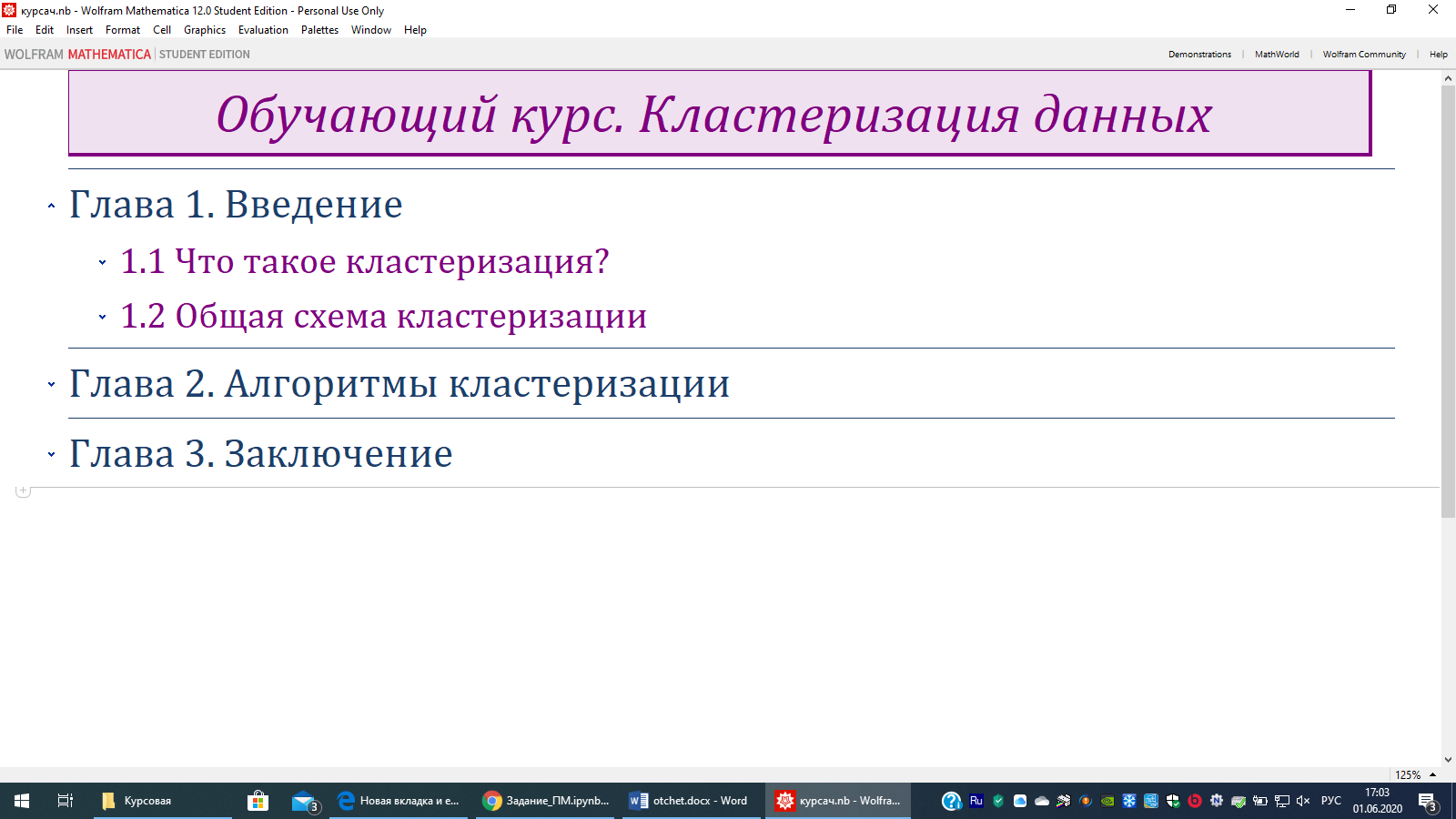


Рисунок 2 – Содержание главы 1

## **1.2 Теоретический материал**

Изучая информацию, которая находится в первой главе, пользователь получит общие сведения о кластеризации: определения, область применения, общая схема и примеры. (Рисунок 3)

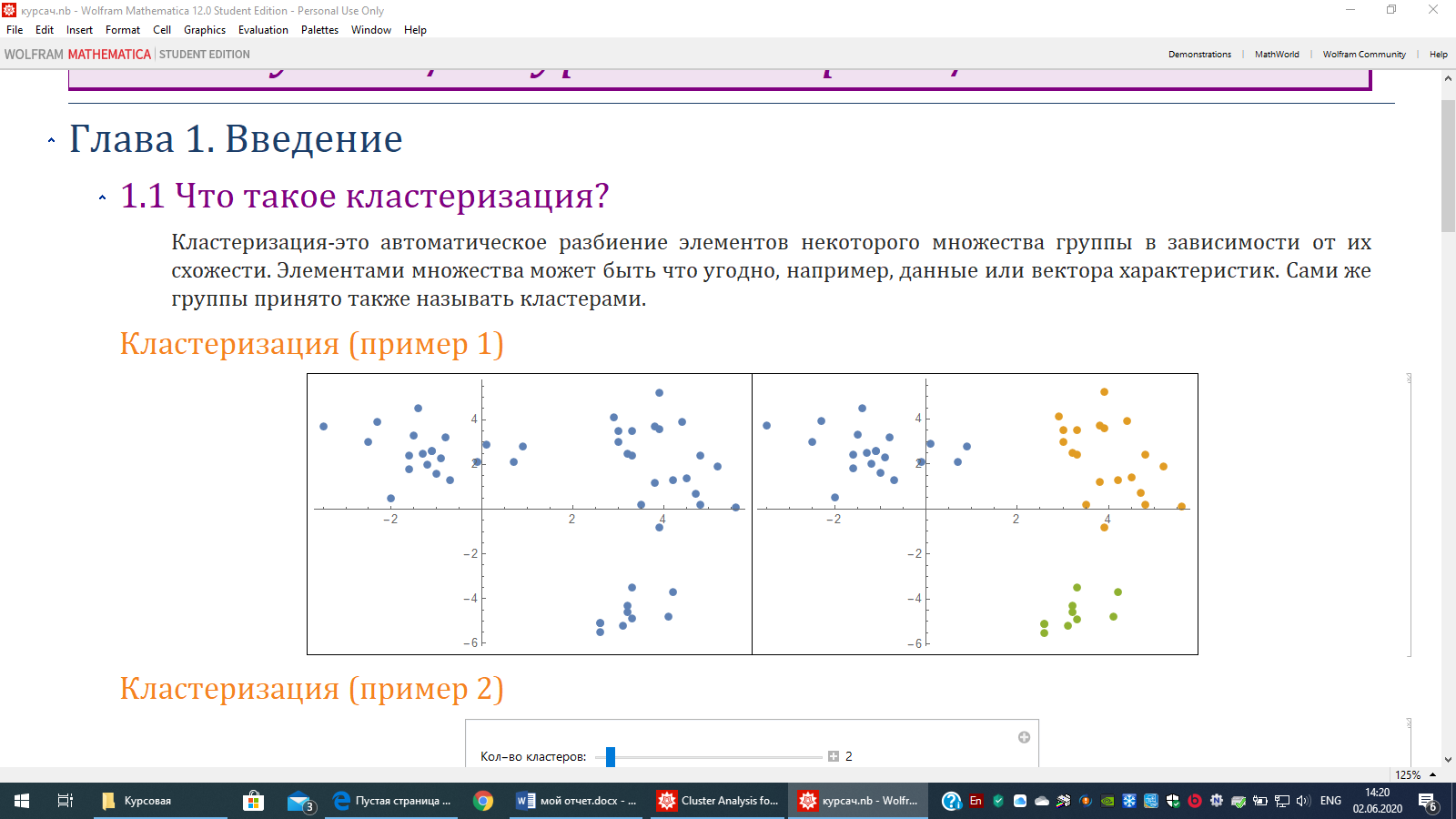


Рисунок 3 – Пример множества точек до и после применения кластеризации

Введем понятия и определения, которые необходимы для прохождения курса.

Кластеризация – это автоматическое разбиение элементов некоторого множества группы в зависимости от их схожести. Элементами множества может быть что угодно, например, данные или вектора характеристик. Сами же группы принято также называть кластерами. На рисунке 4 показано разбиение множества 122 точек на 7 кластеров с помощью Manipulate, также пользователь может сам установить параметры и посмотреть результат.

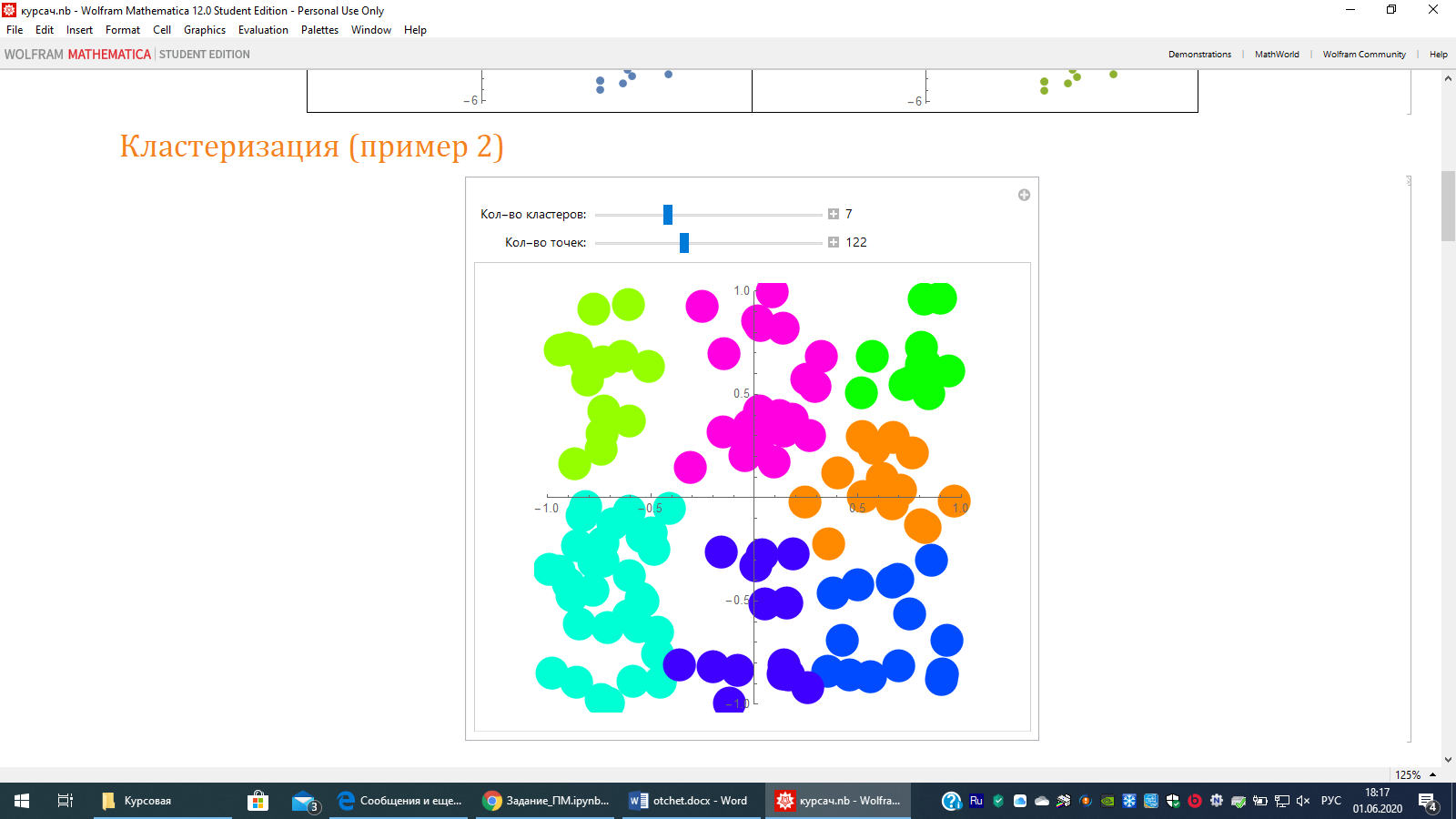


Рисунок 4 – Пример кластеризации

Объект - элементарная группа данных, с которой оперирует алгоритмы кластеризации.

Каждому объекту отождествляется вектор характеристик:

Компоненты являются отдельными характеристиками объекта. Количество характеристик определяет размерность пространства характеристик.

Множество, состоящее из всех векторов характеристик будем обозначать .

Кластер- подмножество «близких друг к другу» объектов из .

Расстояние между объектами и ***-***  результат применения выбранной метрики (или квази-метрики) в пространстве характеристик.

Далее описана общая схема кластеризации:

Кластеризация данных включает в себя следующие этапы:

* выделение характеристик;
* определение метрики;
* разбиение объектов на группы;
* представление результатов.

***Выделение вектора характеристик***

Для начала необходимо выбрать свойства, которые характеризуют наши объекты. Ими могут быть количественные характеристики (координаты, интервалы), качественные характеристики (цвет, статус, воинское звание) и т.д.

Затем стоит попробовать уменьшить размерность пространства характеристических векторов, то есть выделить наиболее важные свойства объектов. Уменьшение размерности ускоряет процесс кластеризации и в ряде случаев позволяет визуально оценивать ее результаты. Выделенные характеристики стоит нормализовать. Далее все объекты представляются в виде характеристических векторов. Мы будем полностью отождествлять объект с его характеристическим вектором.

***Выбор метрики***

Следующим этапом кластеризации является выбор метрики, по которой мы будем определять близость объектов.

Метрика выбирается в зависимости от:

* пространства, в котором расположены объекты;
* неявных характеристик кластеров.

Например, если все координаты объекта непрерывны и вещественны, а кластеры должны представлять собой нечто вроде гиперсфер, то используется классическая метрика Евклида.

На рисунке 5 представлен модуль, иллюстрирующий разбиение точек на кластеры в зависимости от выбранной пользователем метрики.

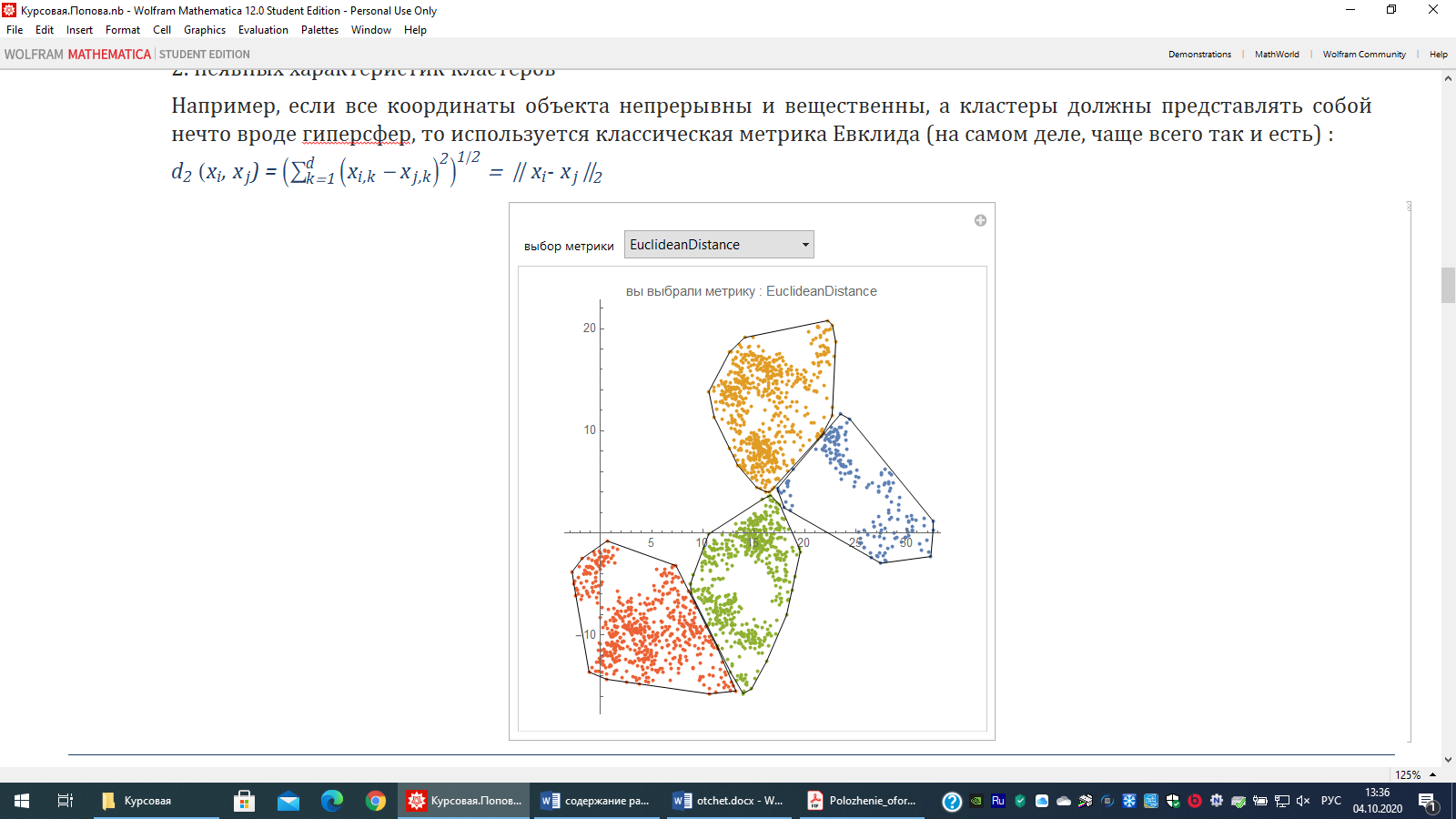


Рисунок 5 – Интерактивный модуль для демонстрации различных метрик

# **АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ**

Глава 2 содержит пять подразделов: классификация, иерархические алгоритмы и 3 детально разобранных алгоритма. (Рисунок 6)

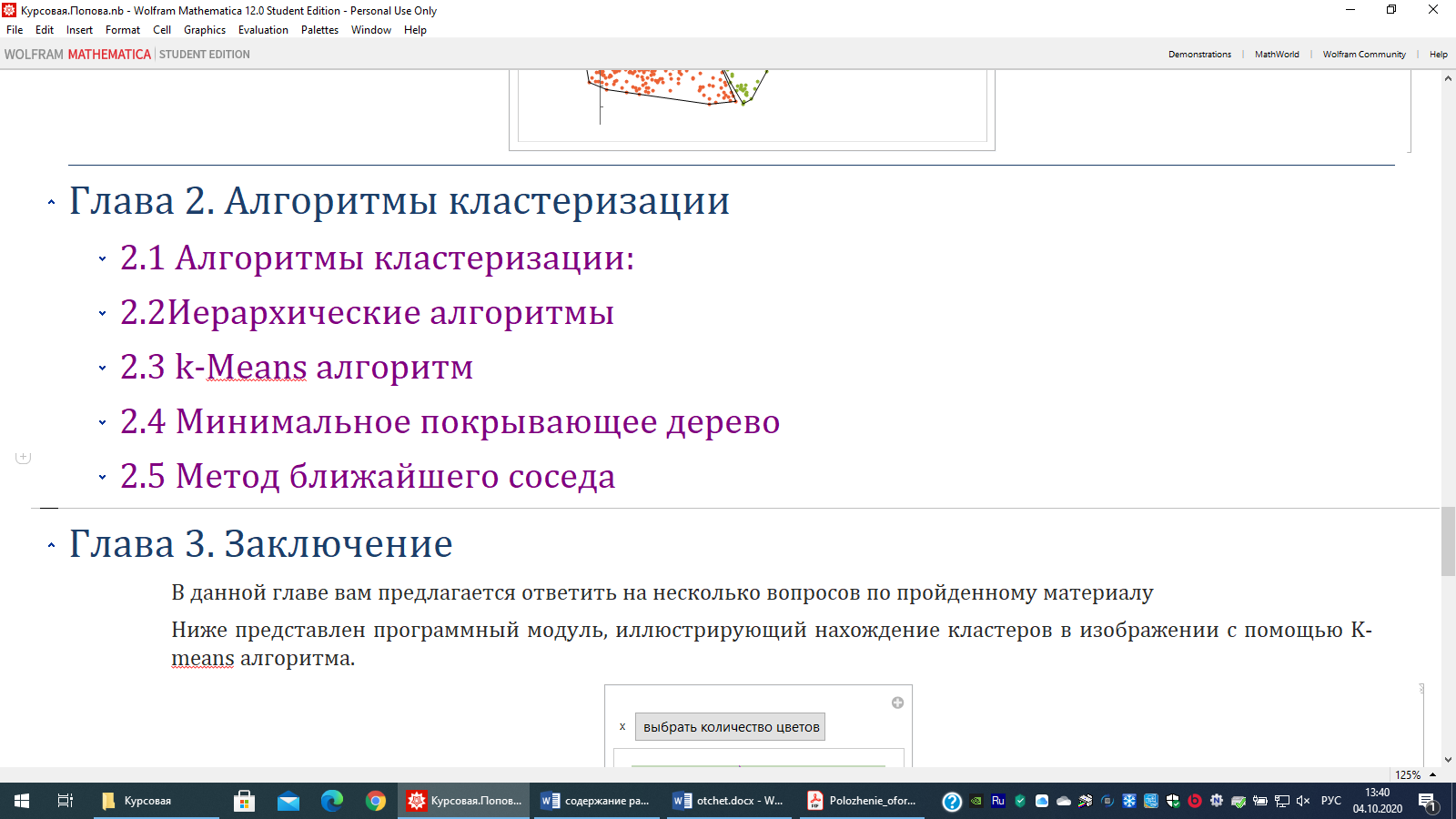


Рисунок 6 - Содержание главы 2

Существует множество различных алгоритмов, например:

* иерархические алгоритмы;
* k - means алгоритм;
* минимальное покрывающее дерево;
* метод ближайшего соседа;
* алгоритмы нечеткой кластеризации;
* применение нейронных сетей;
* генетические алгоритмы;
* метод закалки.

***Классификация алгоритмов***

Алгоритмы кластеризации делят на:

• строящие «снизу – вверх» и «сверху – вниз»;

• монотетические и политетические;

• непересекающиеся и нечеткие;

• детерминированные и стохастические;

• потоковые (online) и не потоковые;

• зависящие и не зависящие от начального разбиения;

• зависящие и не зависящие от порядка рассмотрения объектов.

На рисунке 7 изображена схема классификации алгоритмов кластеризации.

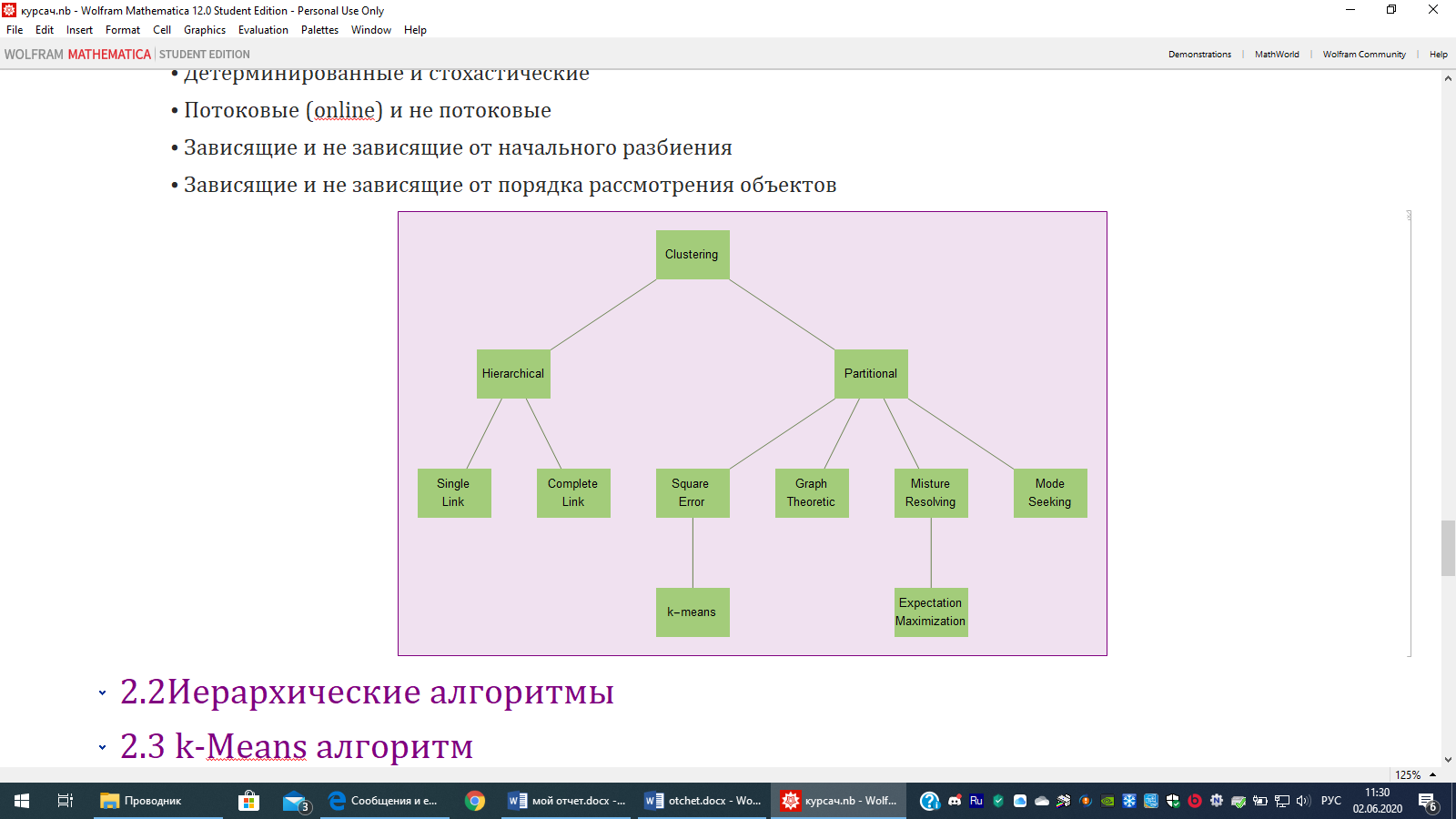


Рисунок 7 – Схема классификации

## **2.1 Иерархические алгоритмы**

Результатом работы иерархических алгоритмов является дендограмма (иерархия), позволяющая разбить исходное множество объектов на любое число кластеров.

Два наиболее популярных алгоритма, оба строят разбиение «снизу-вверх»:

1. Single-link - на каждом шаге объединяет два кластера с наименьшим расстоянием между двумя любыми представителями;

2. Complete-link, на каждом шаге объединяет два кластера с наименьшим расстоянием между двумя наиболее удаленными представителями.

На рисунке 8 изображены дендограммы, которые являются результатом работы Single-link и Complete-link алгоритмов.

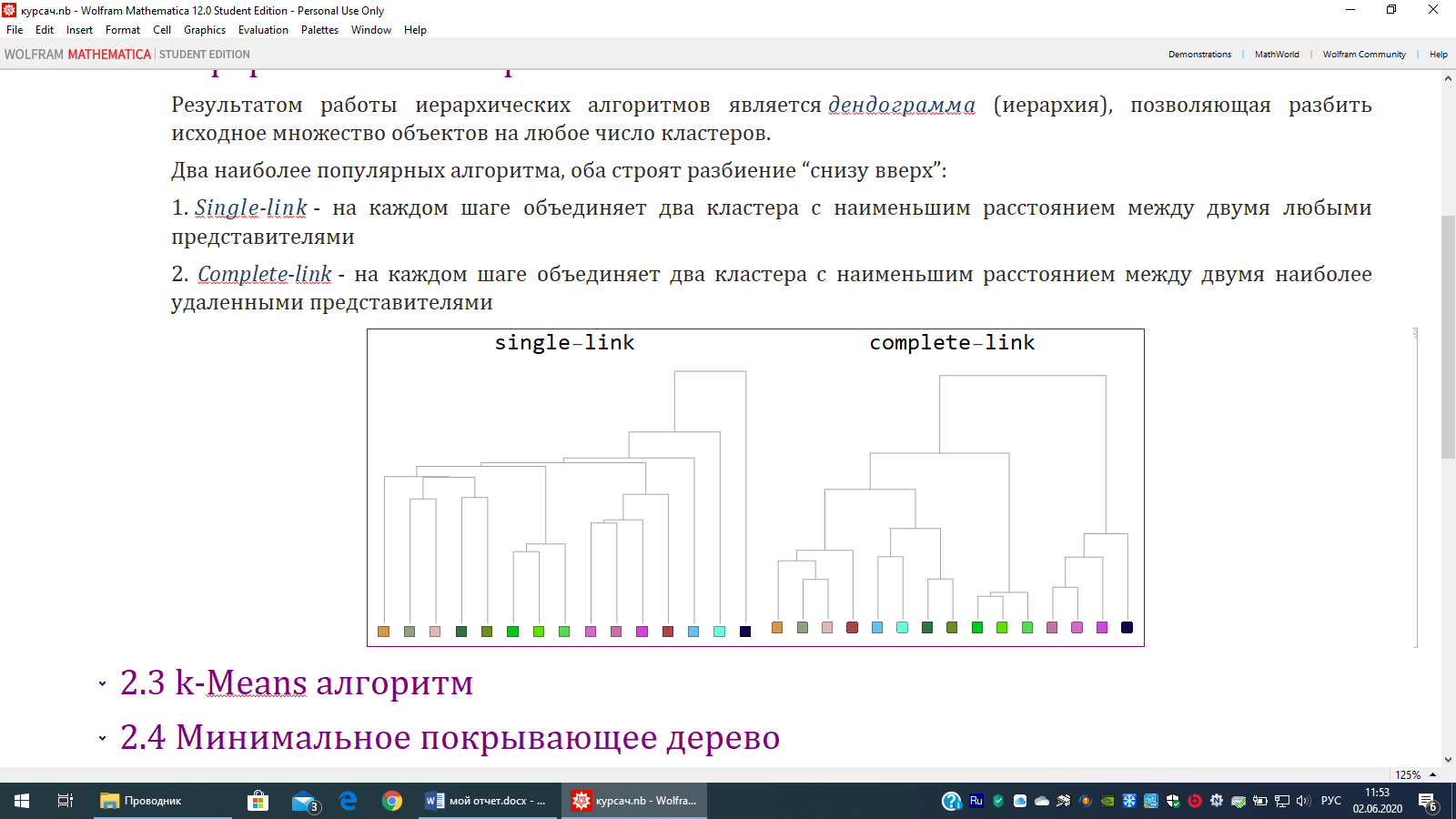


Рисунок 8 – Пример различий Single-link и Complete-link алгоритмов

## **2.2 K–Means алгоритм**

Данный алгоритм состоит из следующих шагов:

1. Случайно выбрать k точек, являющихся начальными "центрами масс" кластеров (любые k из n объектов, или вообще k случайных точек).

2. Отнести каждый объект к кластеру с ближайшим "центром масс".

3. Пересчитать "центры масс" кластеров согласно текущему членству.

4. Если критерий остановки алгоритма не удовлетворен, вернуться к шагу 2.

В качестве критерия остановки обычно выбирают одни из двух:

1. Отсутствие перехода объектов из кластера в кластер на шаге 2.

2.Минимальное изменение среднеквадратичной ошибки.

\*Алгоритм чувствителен к начальному выбору "центров масс".

Далее пользователю представлен программный модуль (Рисунок 9), который пошагово разбирает пример работы алгоритма k-Means на заданном множестве точек.

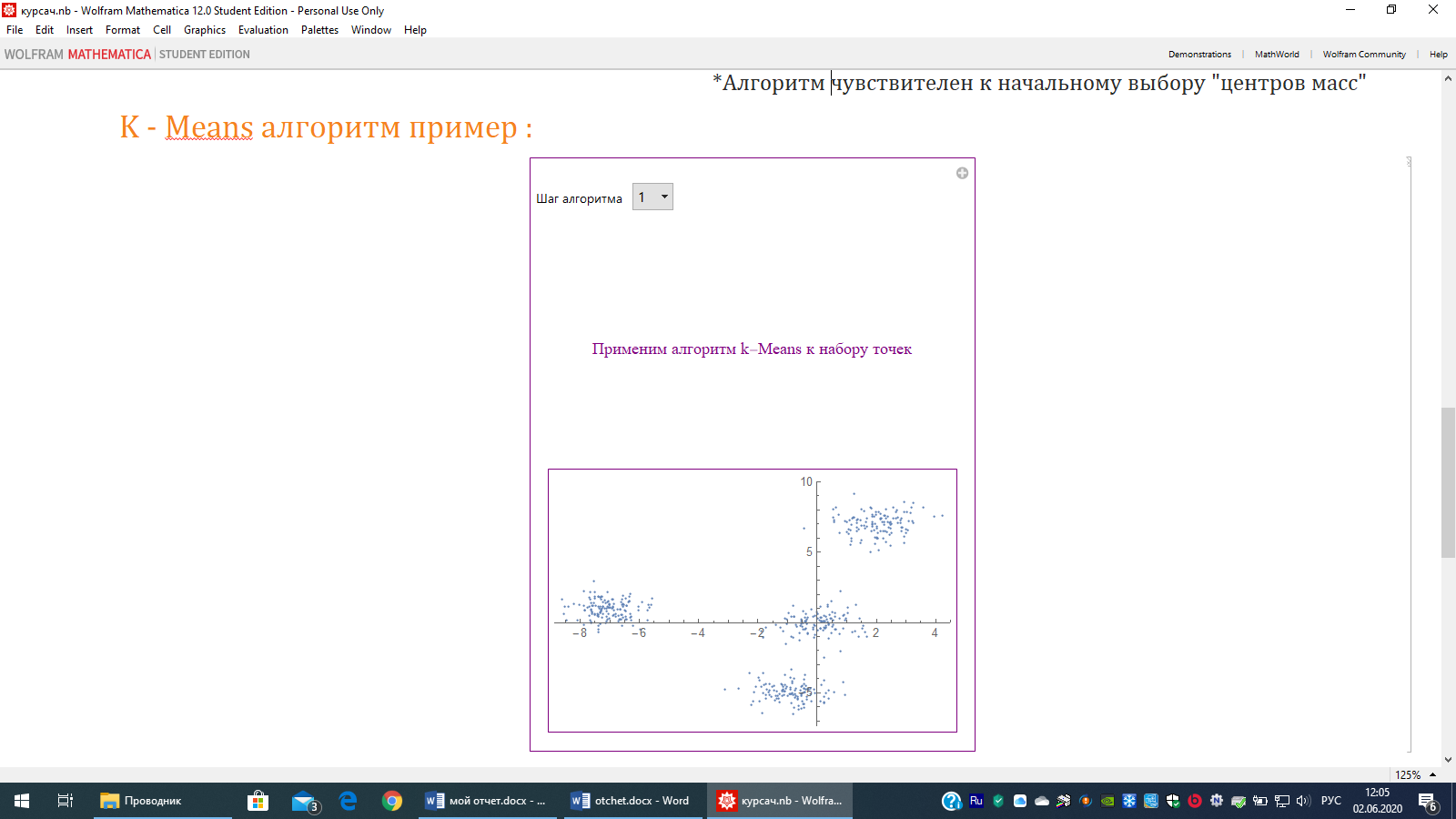


Рисунок 9 - Модуль для демонстрации алгоритма k-Means

С помощью Manipulate и Initialization, в который помещены переменные text1 и risunok1, мы совмещаем текст выбранного пользователем шага алгоритма и его визуализацию, которая находится в переменной risunok1. (Рисунок 10)

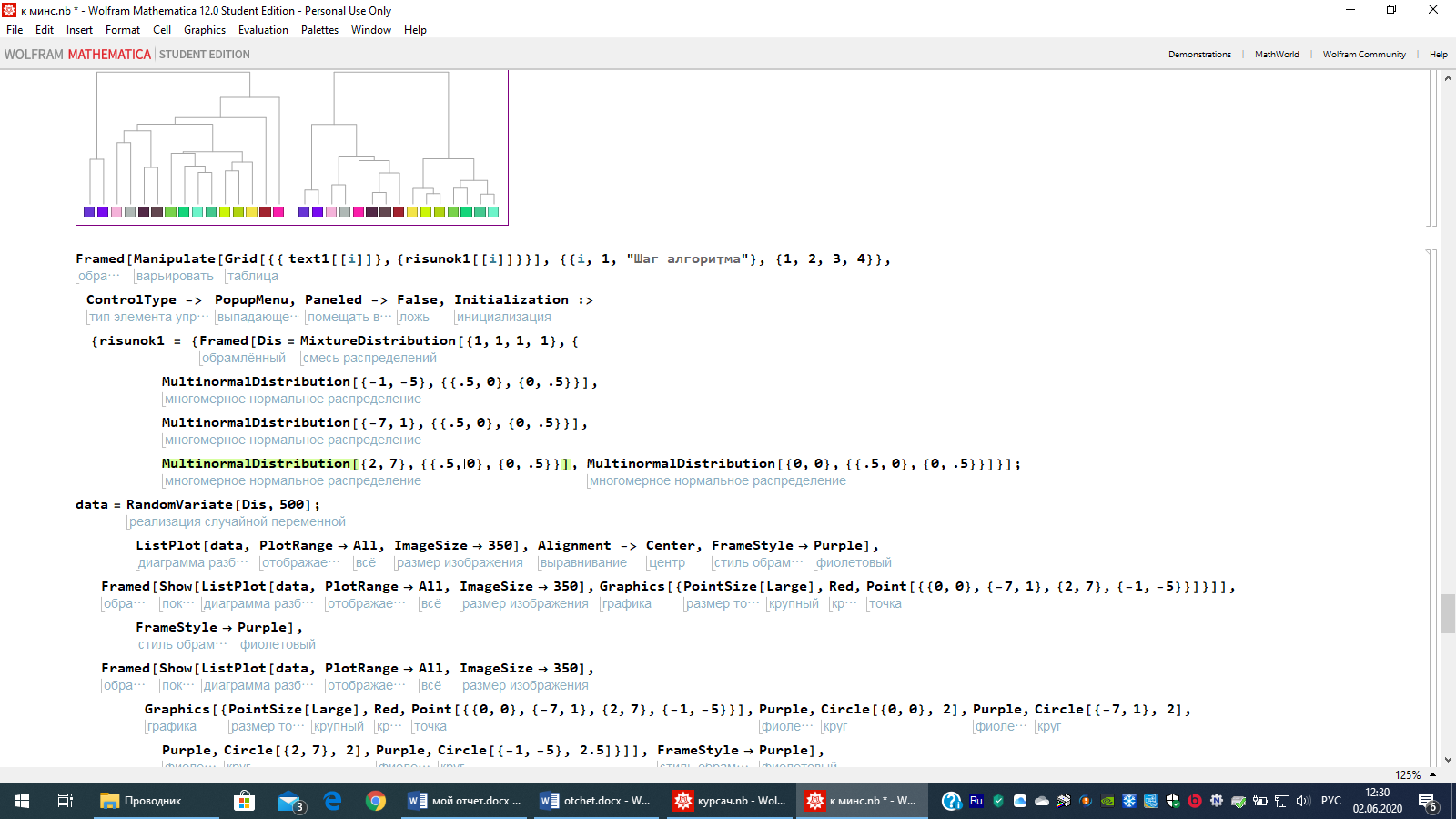


Рисунок 10 – Пример работы алгоритма k-Means с помощью Manipulate

Далее наглядно представлена работа алгоритма k-Means, который я реализовала, в зависимости от количества кластеров. (Рисунок 11)

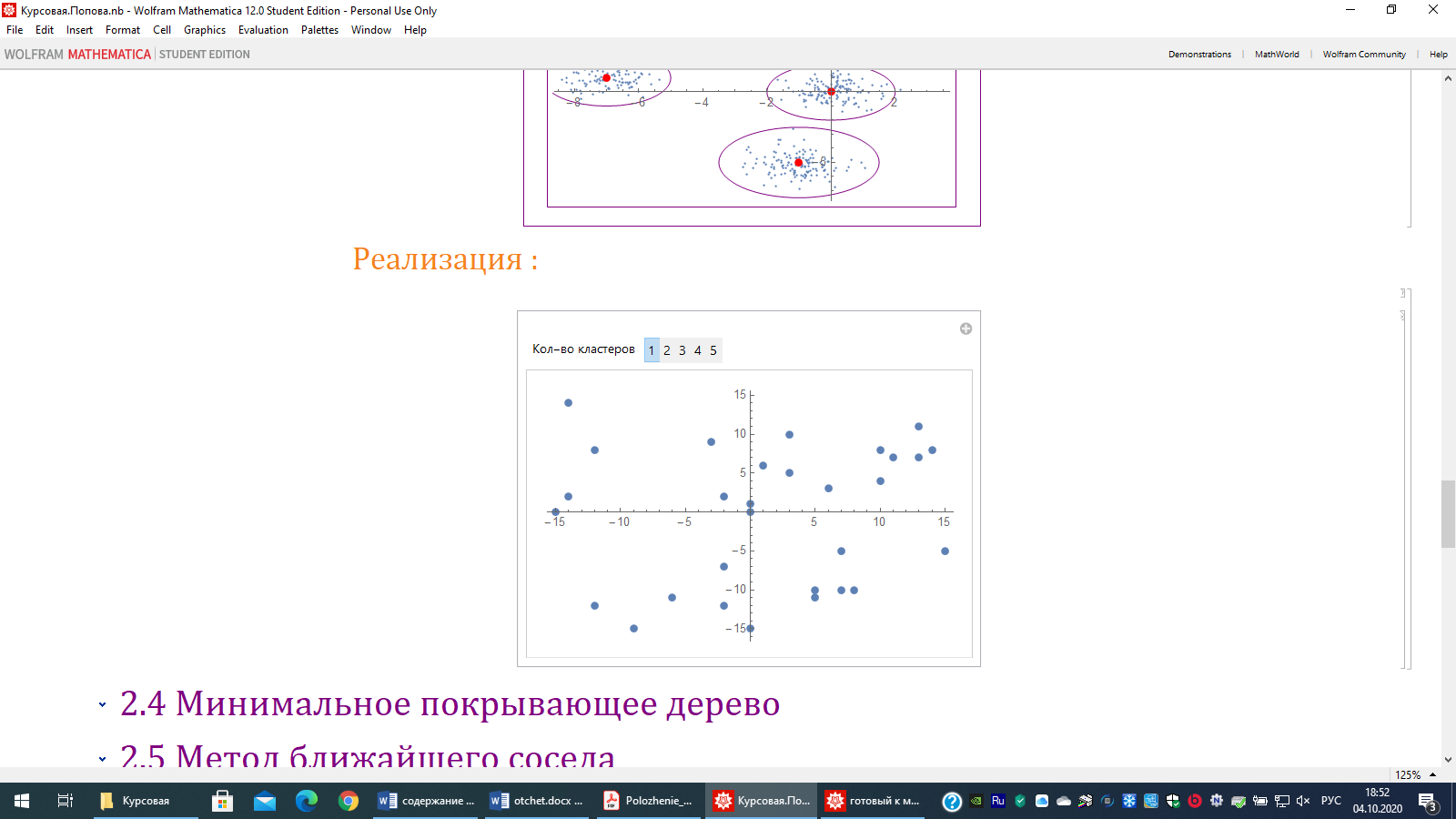


Рисунок 11 – Визуализация реализованного k-Means алгоритма

## **2.3 Алгоритм минимально покрывающего дерева**

Данный метод производит иерархическую кластеризацию «сверху – вниз».

Сначала все объекты помещаются в один кластер. Затем на каждом шаге один из кластеров разбивается на два, так чтобы расстояние между ними было максимальным.

Далее пользователю предложен для освоения предложен программный модуль, который пошагово демонстрирует работу алгоритма.

Данный Manipulate сделан по аналогии с Manipulate для алгоритма k-Means. Также создаются списки текста и картинок, куда помещаются все последовательные описания работы алгоритма и его визуализация. (Рисунок 11)

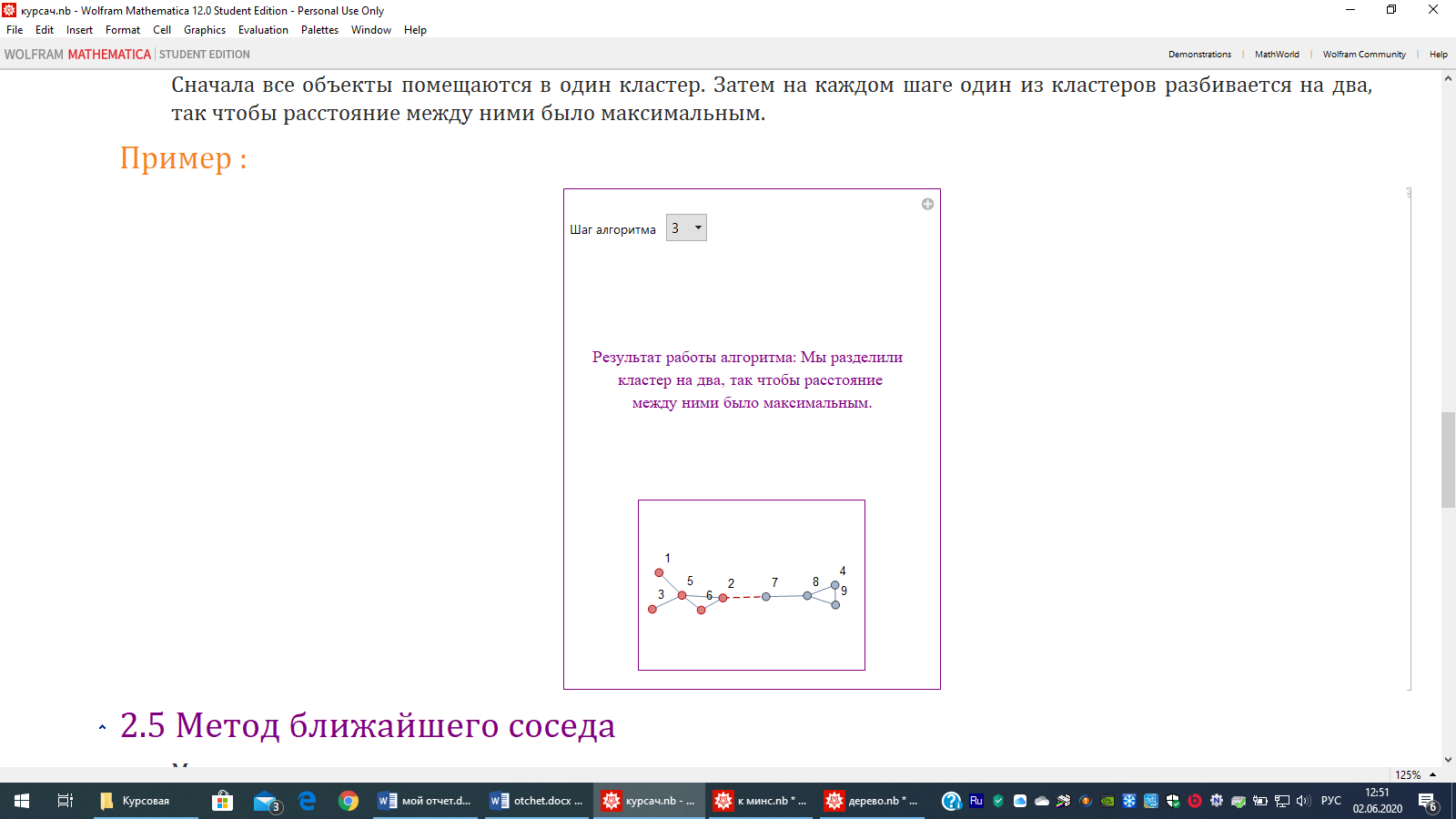


Рисунок 11 - Программный модуль для демонстрации работы алгоритма минимально покрывающего дерева

Далее наглядно представлена работа алгоритма минимально покрывающего дерева, который я реализовала, в зависимости от количества кластеров. (Рисунок 12)

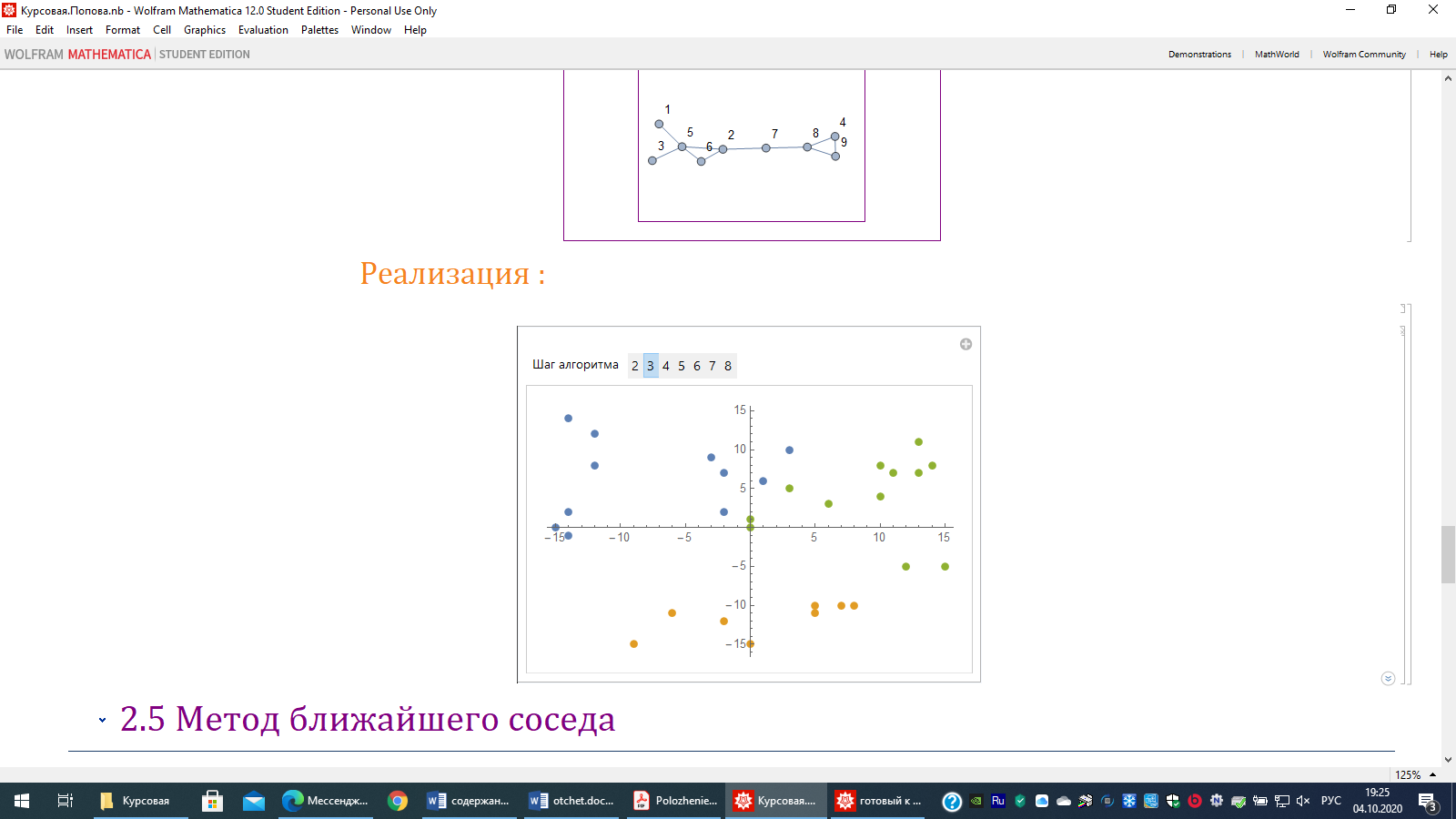


Рисунок 12 - Визуализация реализованного алгоритма минимально покрывающего дерева

## **2.4 Алгоритм – метод ближайшего соседа**

Множество методов иерархического кластерного анализа различается не только используемыми мерами сходства и различия, но и алгоритмами классификации. Из них наиболее распространен метод ближайшего соседа. Этот метод известен также под названием метод одиночной связи.

Пусть требуется провести классификацию заданного множества объектов методом ближайшего соседа. Расстояние между двумя классами определяется как расстояние между ближайшими их представителями.

Перед началом работы алгоритма рассчитывается матрица расстояний между объектами. На каждом шаге в матрице расстояний ищется минимальное значение, соответствующее расстоянию между двумя наиболее близкими кластерами. Найденные кластеры объединяются, образуя новый кластер. Эта процедура повторяется до тех пор, пока не будут объединены все кластеры.

Далее пользователю представлен программный модуль (Рисунок 13), который пошагово разбирает пример работы алгоритма метод ближайшего соседа на заданном множестве точек.

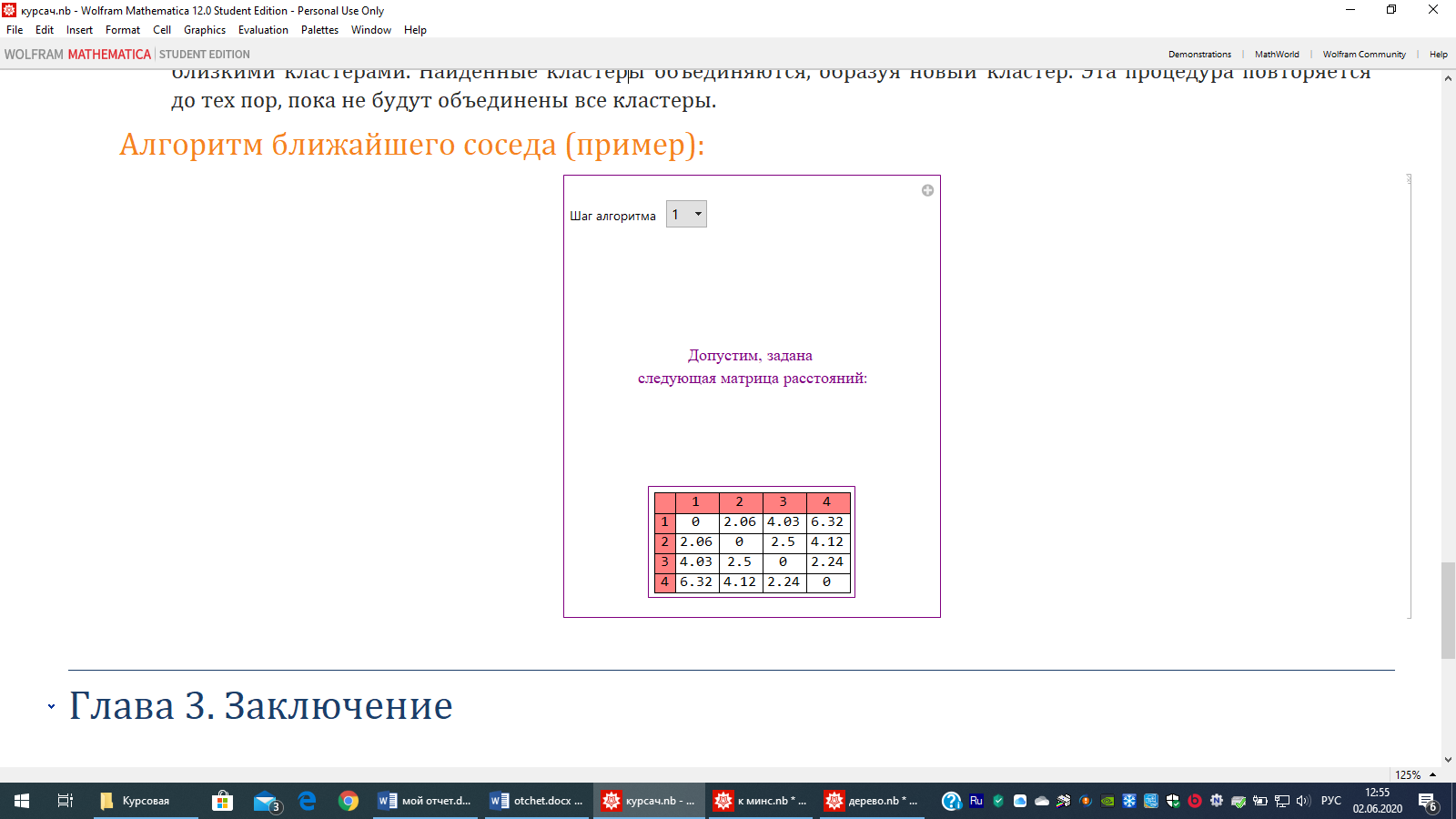


Рисунок 13 - Программный модуль для демонстрации работы алгоритма ближайшего соседа

Далее наглядно представлена работа алгоритма минимально покрывающего дерева, который я реализовала, в зависимости от количества кластеров. (Рисунок 14)

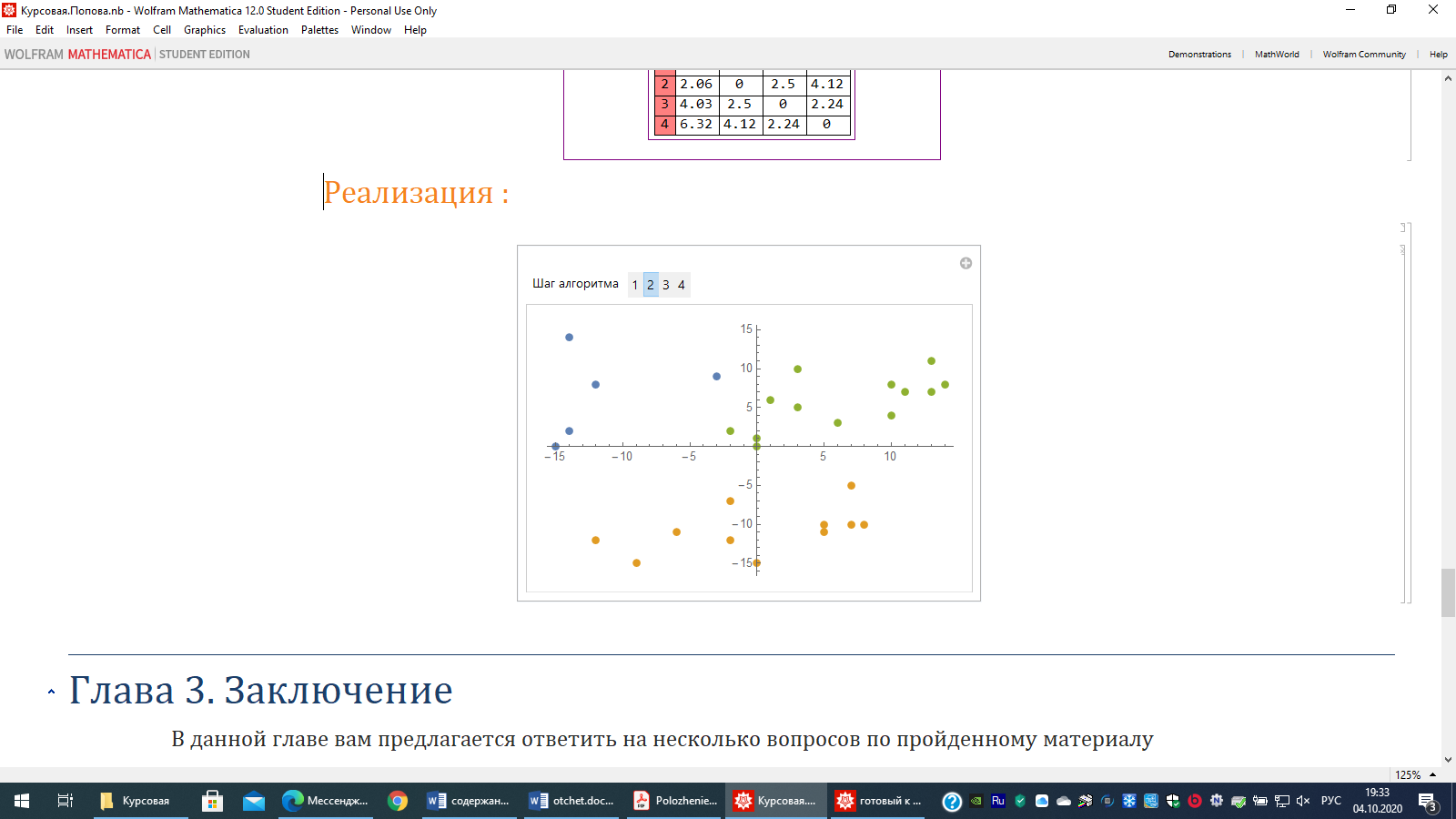


Рисунок 14 – Визуализация реализованного алгоритма метода ближайшего соседа

## **2.5** **Тестирование**

В этой главе пользователю представлен программный модуль применения алгоритма k-Means для изображения, а также небольшой тест для самопроверки.

Пользователь может нажать на кнопку «выбрать количество цветов» и изображение розы разделится на случайное количество кластеров от 2 до 50. (Рисунок 15)

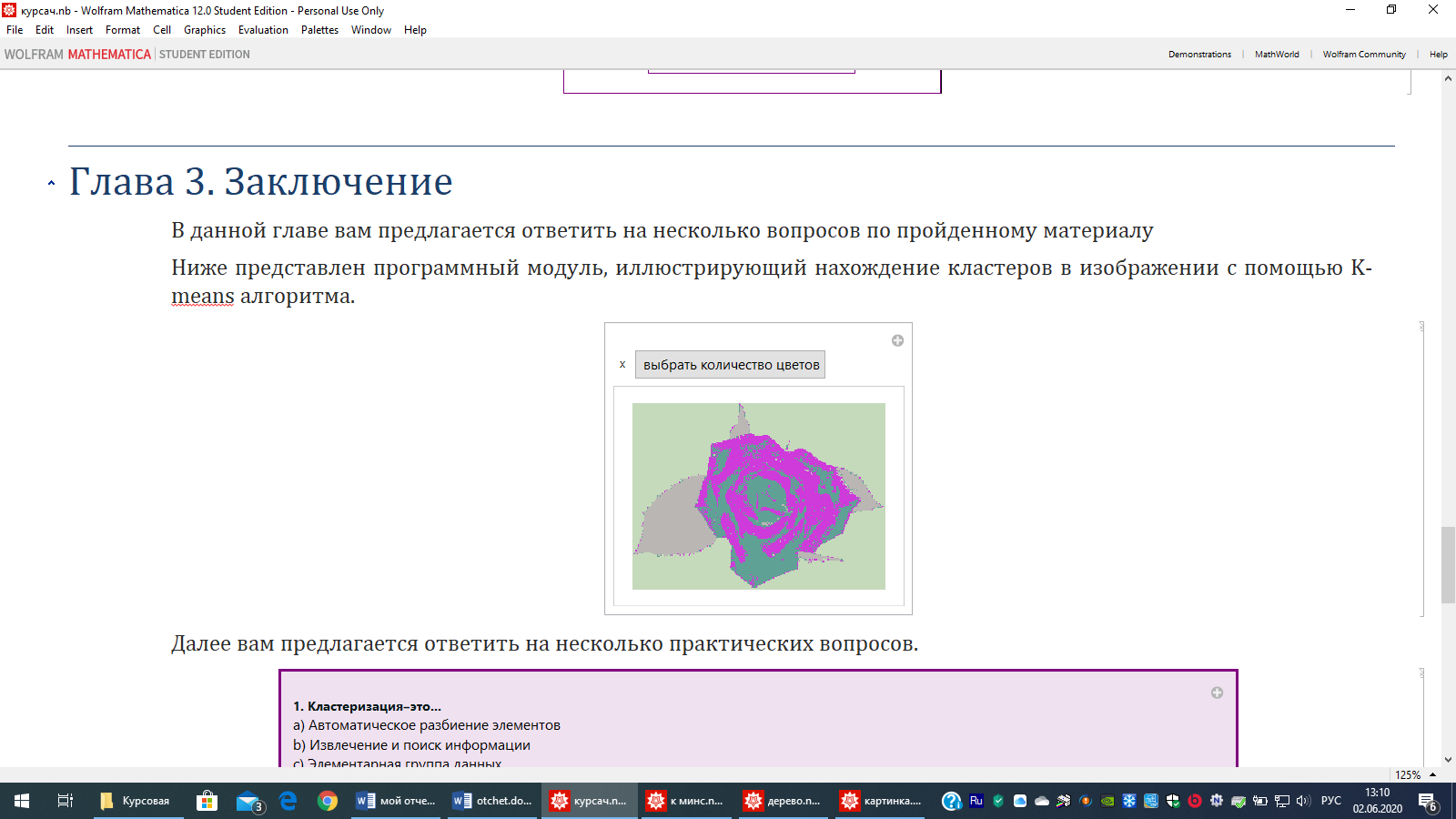


Рисунок 15 – Пример кластеризации изображения

Теперь рассмотрим тест для самопроверки (Рисунок 16 и 17). Тест содержит 5 вопросов, в которых пользователю предлагается выбрать один вариант ответа из 4.

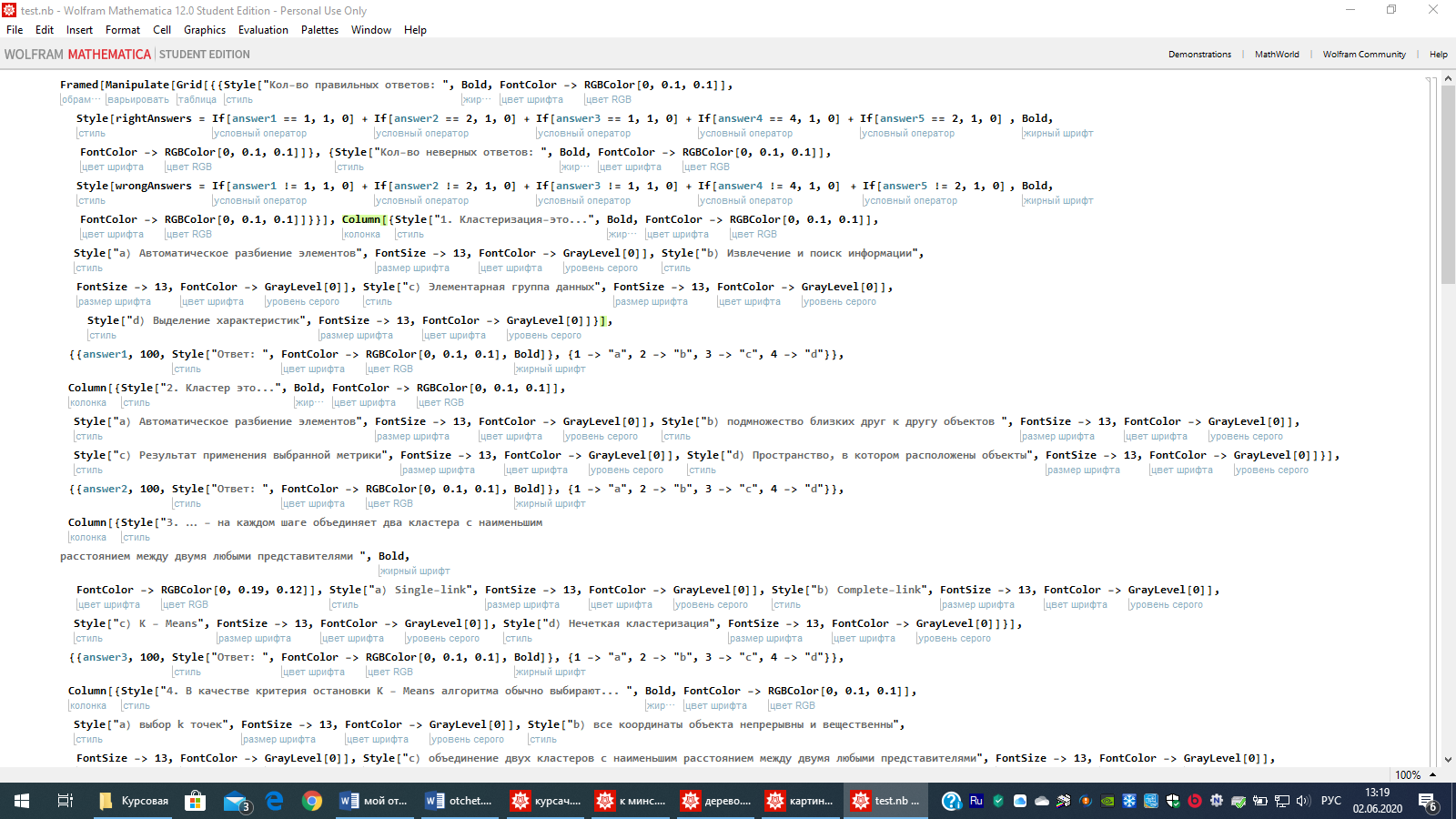


Рисунок 16 – Manipulate для создания теста

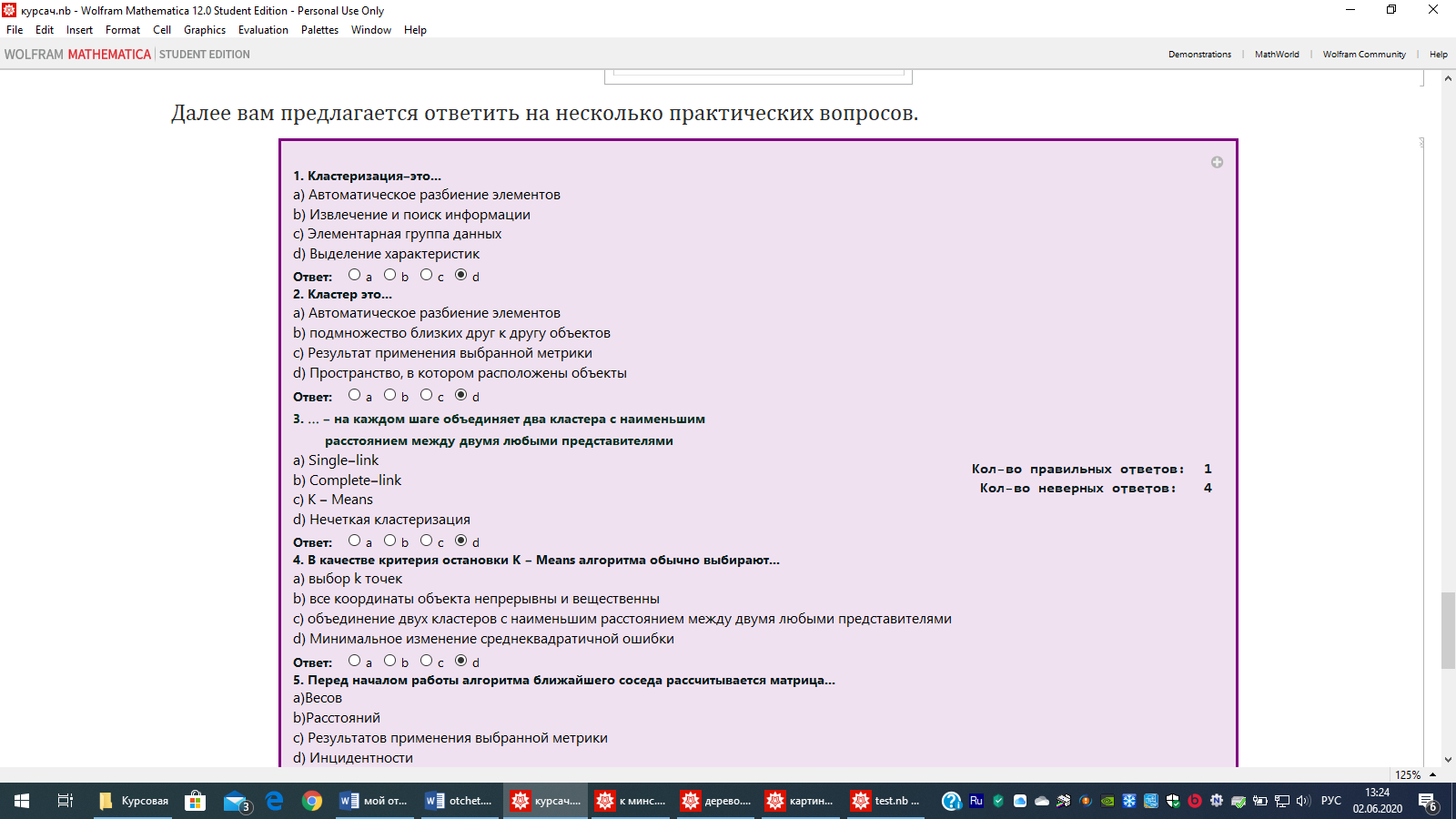


Рисунок 17 – Тест для закрепления информации

Как только пользователь меняет ответ на вопрос в тесте (Рисунок 17), «количество правильных/неправильных ответов» автоматически меняется.

# **ЗАКЛЮЧЕНИЕ**

В ходе работы были продемонстрированы навыки пользования компьютерной системой Wolfram Mathematica, например, основой для улучшения качества восприятия файла был раздел 𝐸𝑑𝑖𝑡 𝑆𝑡𝑦𝑙𝑒𝑠ℎ𝑒𝑒𝑡. Также были реализованы алгоритмы применены базовые знания использования функций визуализации для демонстрации работы алгоритмов кластеризации.

С помощью структурированного текста, интерактивных блоков и теста, которые были представлены в обучающем курсе «Кластеризация данных» и сделали его доступным для понимания, пользователь сможет ознакомиться с базовыми знаниями в этой теме и при желании продолжить ее изучение самостоятельно.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Tyon R.G. Cluster analysis. L.: Ann Arbor Edwards Bros, 1939, 139 p.
2. Компьютерная система Wolfram Mathematica [Электронный ресурс] / - Режим доступа: <http://www.wolfram.com/mathematica/>
3. Райзин Д. В. Классификация и кластеризация. - M.:Мир, 1980. - 390 с.