Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа №6

за 7 семестр

По дисциплине: «КМиАД»

Выполнил:

Студент 4 курса

Группы ПО-4(2)

Тупик Д. Л.

Проверил:

Чичурин А. В.

2022

**Визуальный анализ данных**

**Выполнение визуального анализа данных**

Основной идеей визуального анализа данных является представление данных в некоторой визуальной форме, позволяющей человеку погрузиться в [данные](https://intellect.icu/informatsiya-znaniya-i-dannye-otnosheniya-mezhdu-nimi-otlichiya-i-skhodstva-preobrazovanie-priznaki-znanij-6173#term-dannye), работать с их визуальным представлением, понять их суть, сделать выводы и напрямую взаимодействовать с данными.

Визуальный анализ данных обычно выполняется в три этапа:

- беглый анализ — позволяет идентифицировать интересные шаблоны и сфокусироваться на одном или нескольких из них;

- увеличение и фильтрация — идентифицированные на предыдущем этапе шаблоны отфильтровываются и рассматриваются в большем масштабе;

- детализация по необходимости — если пользователю нужно получить дополнительную информацию, он может визуализировать более детальные данные.

**Характеристики средств визуализации данных**

- характер отображаемых данных, которые нужно визуализировать с помощью данного средства;

- методы визуализации и образы, в виде которых могут быть представлены данные;

- возможности взаимодействия с визуальными образами и методами для лучшего анализа данных.

**Методы визуализации**

Выделяют следующие типы методов визуализации:

* стандартные 2D/3D-образы — гистограммы, линейные графики и т. п.;
* геометрические преобразования — диаграмма разброса данных, параллельные координаты и т. п.;
* отображение иконок — линейчатые фигуры (needle icons) и звезды (star icons);
* методы, ориентированные на пикселы — рекурсивные шаблоны, циклические сегменты и т. п.;
* иерархические образы — древовидные карты и наложение измерений.

**Отображение иконок**

Подход, основанный на отображении иконок, предполагает каждому объекту данных ставить в соответствие некоторую иконку. При этом атрибуты объекта должны отображаться различными визуальными свойствами иконок. Иконки могут комбинироваться в матрицы или графики и, таким образом, предоставляют возможность анализировать все объекты в целом.

Использование иконок предполагает следующие методы визуализации:

* линейчатые фигуры;
* "лица Чернова";
* цветные иконки;
* глифы 1 и др.

**Иерархические образы**

Иерархические образы используются для отображения иерархий и отношений в данных. Они применяются в следующих методах:

* иерархические оси;
* наложение измерений;
* деревья.

Основная идея метода наложения измерений заключается во вставке одной координатной системы в другую. Иными словами, два атрибута формируют внешнюю систему координат, два других атрибута формируют другую систему координат, встроенную в предыдущую, и т. д. Этот процесс может быть повторен несколько раз.

Для визуализации данных используют два основных вида древовидных структур:

* древовидные карты;
* канонические деревья.

Система Mathematica предоставляет современную функциональность для анализа и синтезирования графов и сетей. При работе с графами доступны следующие возможности:

* Поддержка ориентированных, неориентированных и взвешенных графов.
* Непосредственная поддержка случайных графовых распределений.
* Обширная коллекция операций на графах и операций по изменению графов.
* Поддержка теоретико-множественных операций на графах и операций, заданных булевыми выражениями.
* Выделение графовых элементов и подграфов с помощью языка шаблонов системы Mathematica.
* Всеобъемлющий набор предикатов для тестирования свойств графов.
* Эффективное тестирование изоморфности графов.
* Локальные и глобальные структурные свойства, включая компоненты накрытия и сопоставления.
* Более 15 метрик и мер центральности для квалифицирования графов и сетей.
* Эффективные функции нахождения кратчайшего пути, циклов и маршрутов.
* Мультипарадигмальный подход к графовому программированию с использованием матричной, оптимизационной или булевой подсистем.
* Общие алгоритмы поиска в глубину и в ширину, обладающие гибким программным интерфейсом.
* Поддержка произвольных свойств элементов графа.

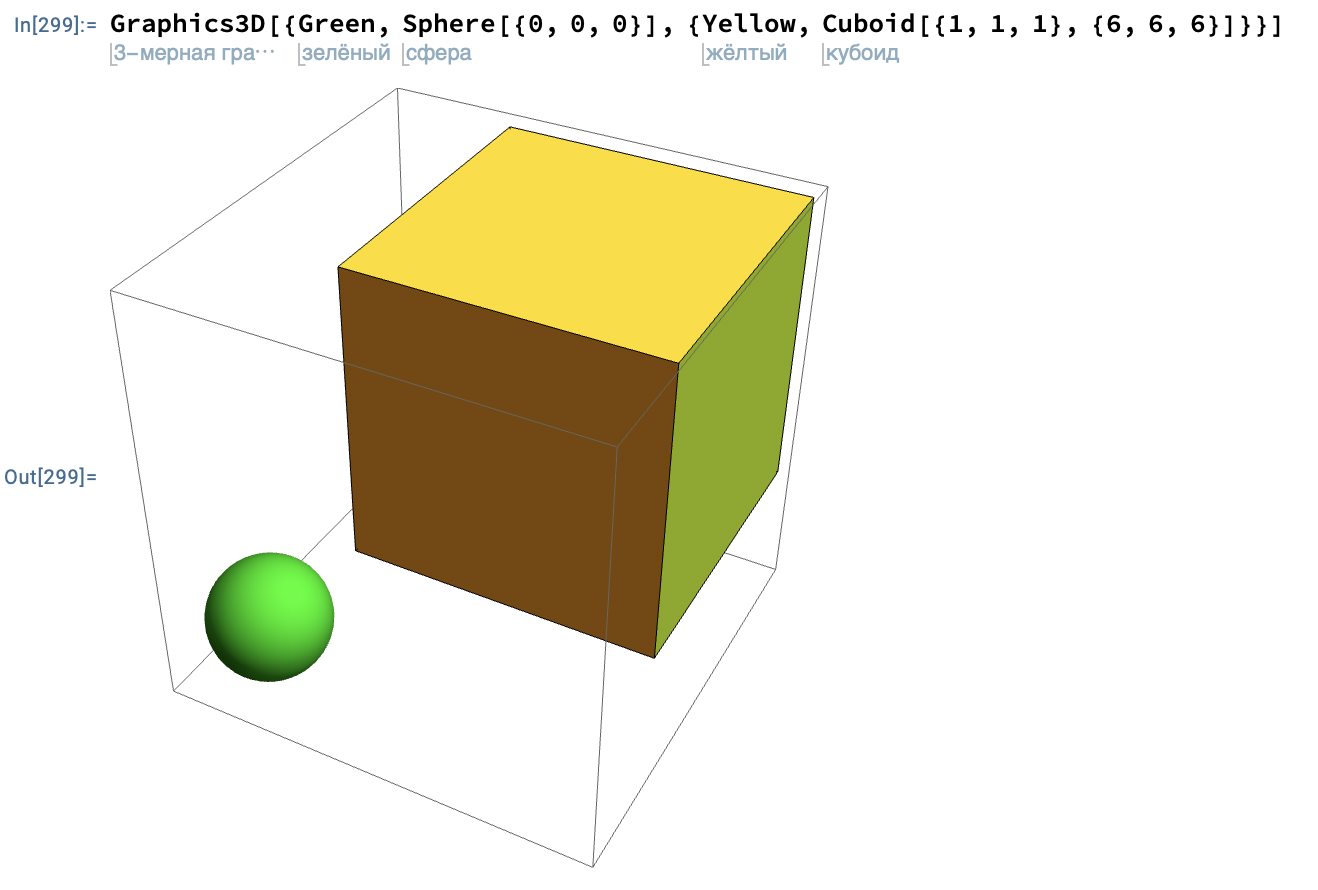
**Graphics3D**

Mathematica включает в себя графические примитивы, при помощи которых Вы можете создавать двумерную и трехмерную графику любой сложности.

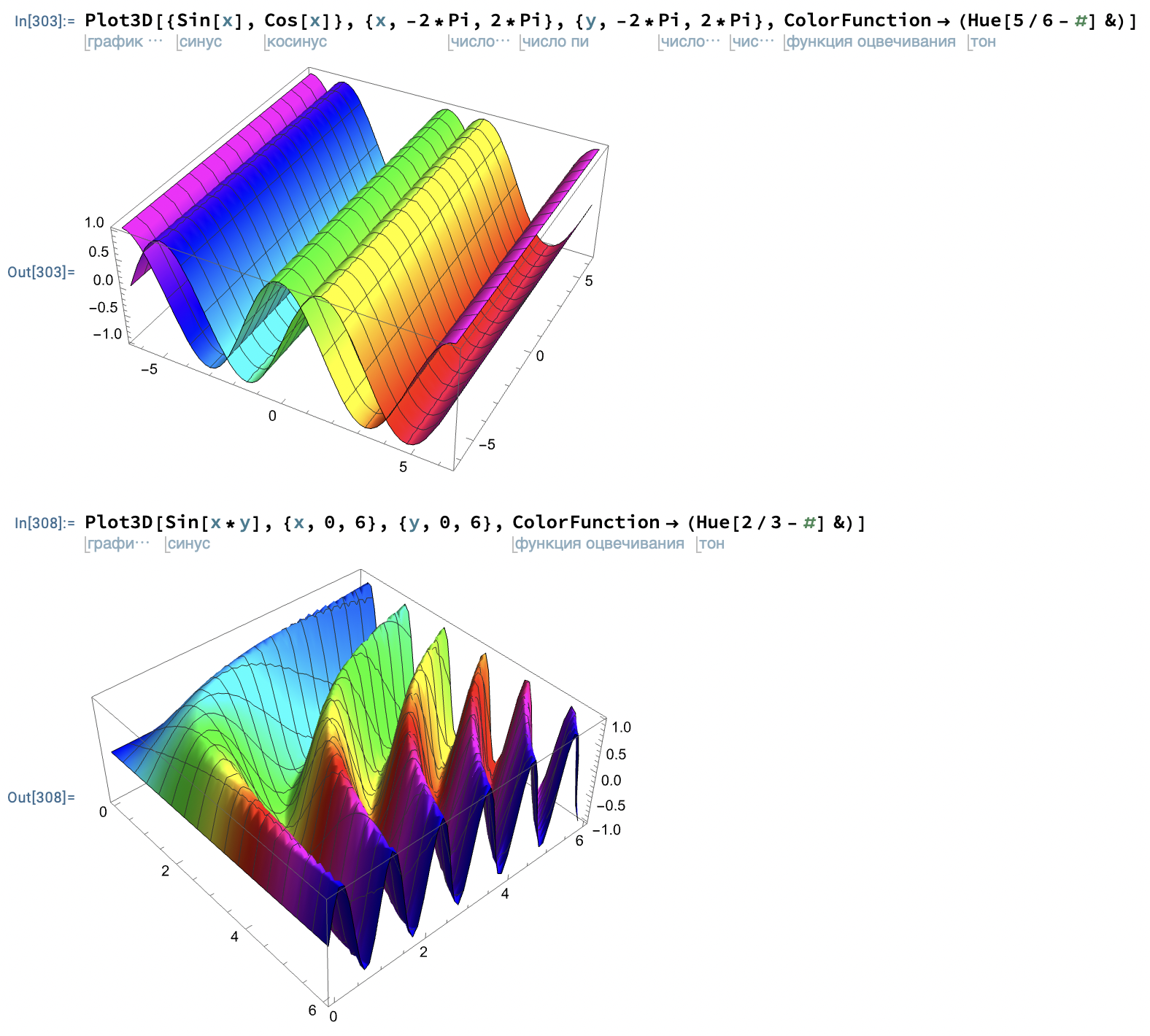
Функция Graphics3D преобразует трехмерные примитивы в трехмерные

графические объекты, добавляя в структуру примитива стили (цвет, толщину и

т.д.).

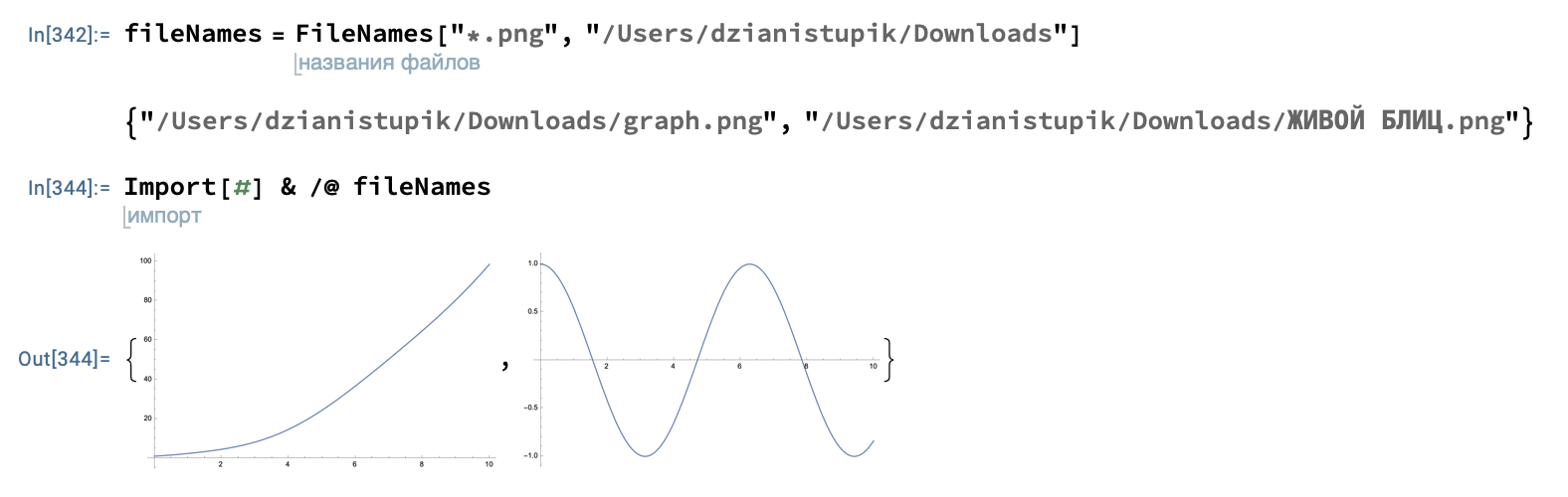


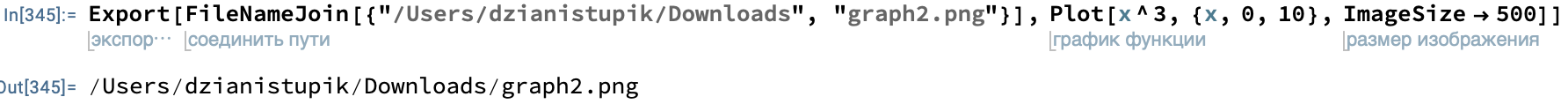
**Plot3D**

Mathematica позволяет создавать не только статичную, но и анимированную графику. Первый аргумент – выражение f, график которого должен быть построен. Второй и третий аргументы определяют границы изменения переменных x и y задаваемые в виде списков. Обычно функция вычисляется на сетке 15 на 15 и каждый кусок закрашивается согласно светоотражающий модели.

**Import, Export**

Импорт и экспорт может обрабатывать не только табличные данные, но и данные, соответствующие графике, звукам, выражениям и даже целым документам. Импорт и экспорт часто могут вывести соответствующий формат для данных, просто посмотрев на расширение имени файла для файла, в котором хранятся данные.

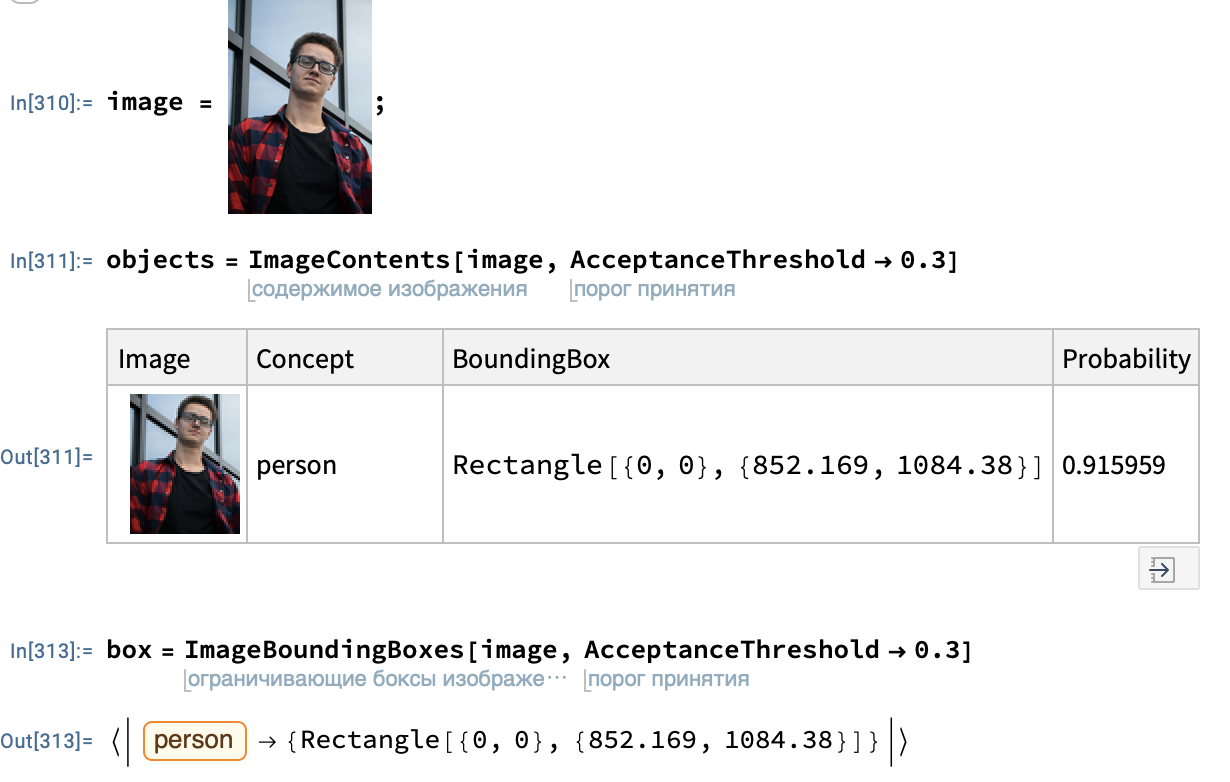
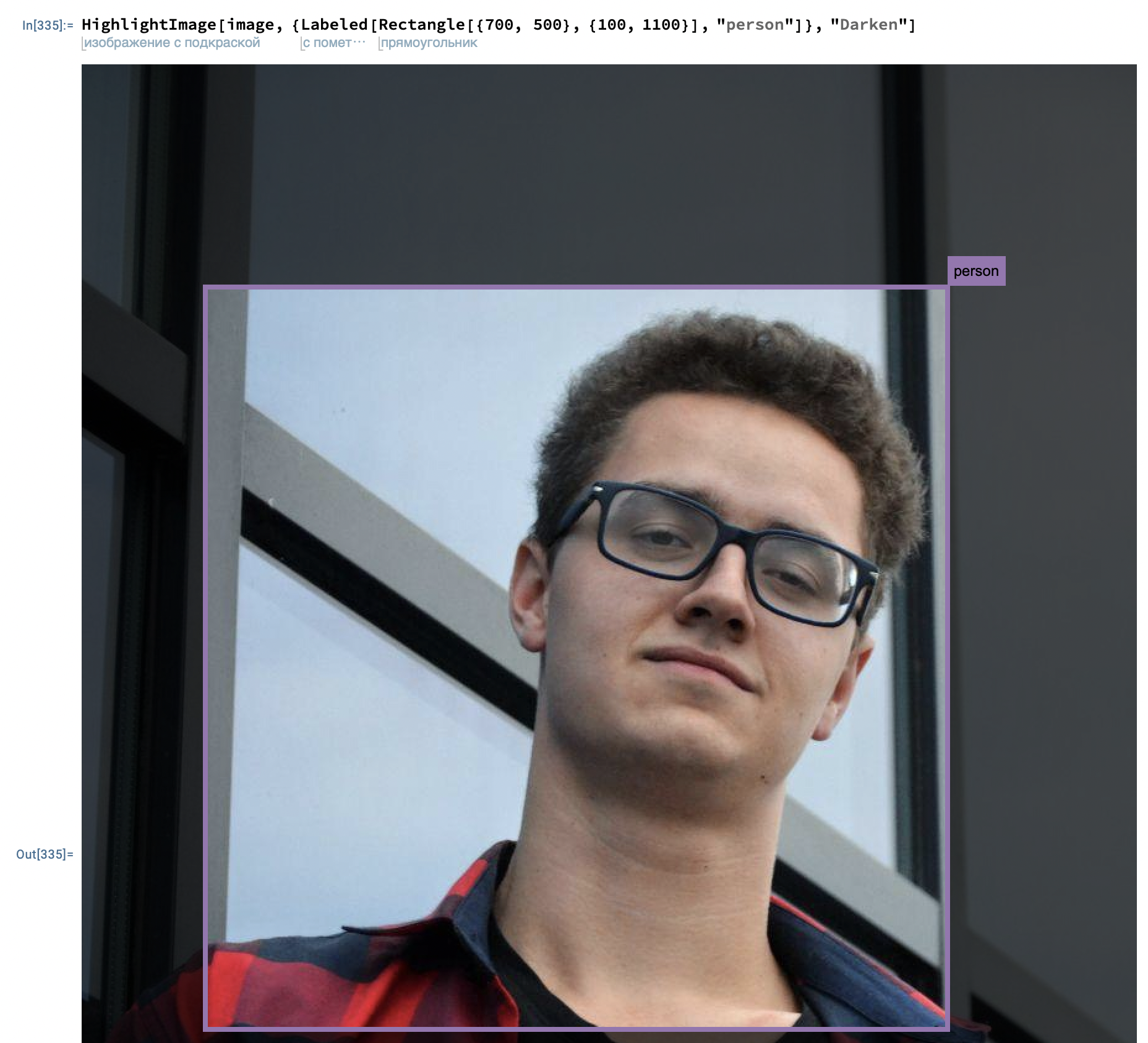
Импортирует данные из источника, возвращая их представление Wolfram Language.

Экспортирует данные в файл, преобразуя их в формат, соответствующий расширению файла.Изображение

**ImageContents, ImageBoundingBoxes**

Дает набор данных идентифицированных сущностей на изображении.

**HighlightImage**

Дает сопоставление списков ограничительных рамок для каждой идентифицированной категории объектов на изображении.

**Вывод:**

Основной идеей визуального анализа данных является представление данных в некоторой визуальной форме, позволяющей человеку погрузиться в данные, работать с их визуальным представлением, понять их суть, сделать выводы и напрямую взаимодействовать с ними.

Визуальный анализ данных обычно выполняется в три этапа: беглый анализ, увеличение и фильтрация, детализация по необходимости.

Выделяют три основные характеристики средств визуализации: характер отображаемых данных, методы визуализации, возможности взаимодействия с визуальными образами.

Выделяют следующие виды данных, с которыми могут работать средства визуализации: одномерные, двумерные и многомерные данные, тексты и гипертексты, иерархические и связанные данные, алгоритмы и программы.

Выделяют следующие основные типы методов визуализации: стандартные 2D/3D-образы, геометрические преобразования, отображение иконок, ориентированные на пикселы методы, иерархические образы.

Для анализа визуальных образов часто используют следующие возможности взаимодействия: динамическое проецирование, интерактивная фильтрация, масштабирование образов, интерактивное искажение, интерактивное комбинирование.

Основная идея методов геометрических преобразований — визуализировать преобразования и проекции данных в декартовом и недекартовом геометрических пространствах.

Подход, основанный на отображении иконок, предполагает каждому объекту данных ставить в соответствие некоторую иконку, при этом атрибуты объекта должны отображаться с помощью различных визуальных свойств иконок.

Методы, ориентированные на пикселы, используют для представления каждого элемента данных цветные пикселы.

Иерархические образы используются для отображения иерархий и отношений в данных.

**Извлечение знаний из Web — Web Mining**

**Проблемы анализа информации из Web**

Всемирная сеть сейчас содержит огромное количество информации, знаний. Пользователи на различных условиях могут просматривать всевозможные документы, аудио- и видеофайлы. Однако это многообразие данных скрывает в себе проблемы, которые могут возникнуть не только при анализе, но и при поиске необходимой информации в Интернет.

* Проблема поиска нужной информации связана с тем, что пользователь не всегда сразу может найти необходимые ему электронные ресурсы. Лишь небольшой процент ссылок среди предложенных поисковыми системами приводит к требуемым документам. Также труден поиск неиндексированной информации такими средствами.
* Проблема обнаружения новых знаний. Даже если найдено множество информации, для пользователя извлечение полезных знаний является довольно трудоемкой и непростой задачей. Сюда же можно и отнести сложности, связанные с осмыслением сведений, понятием тех идей, которые были вложены авторами.
* Проблема изучения потребителей связана с предоставлением пользователю информации, которая оказалась бы ему интересна. Это особенно актуально для электронных торговых порталов, которые могли бы "подсказывать" пользователю при выборе товара.

**Этапы Web Mining**

Web Mining — технология, использующая методы Data Mining для исследования и извлечения информации из Web-документов и сервисов. Выделяют следующие этапы применения Web Mining:

1. Поиск ресурсов — локализация неизвестных документов и сервисов в Web.

2. Извлечение информации — автоматическое извлечение определенной информации из найденных Web-ресурсов.

3. Обобщение — обнаружение общих шаблонов в отдельных и пересекающихся множествах сайтов.

4. Анализ — интерпретация найденных шаблонов.Web Mining и другие интернет-технологии

**WolframAlpha** - это смесь алгоритмов и использование искусственного интеллекта для поиска нужной информации.

**Извлечение Web-контента в процессе информационного поиска**

Методы извлечения Web-контента в процессе информационного поиска во многом зависят от типа анализируемых документов. Различают два основных типа: неструктурированные и почти структурированные. К неструктурированному типу относятся все текстовые документы, не имеющие определенной структуры. К почти структурированным относятся документы, имеющие структуру в целом, но позволяющую вхождение в структурный элемент неструктурированного текста. К таким документам относятся HTML, XML и др.

**Слабоструктурированные документы**

Извлечение Web-контента из слабоструктурированных документов использует более развитые средства представления текста. Это в первую очередь связано с тем, что в документах уже выделены некоторые структурные элементы. Практически все методы в этой области для представления документа используют HTML-структуры внутри документов. Некоторые методы используют также для представления гиперссылки между документами.

**Извлечение Web-контента для формирования баз данных**

Задача извлечения Web-контента для его размещения в базе данных относится к проблеме управления информацией и обработки запросов к ней. Существуют три класса задач, относящихся к этой проблеме:

* моделирование и формирование запросов к Web;
* извлечение информации и интеграция;
* создание и реструктуризация Web-сайта.

**Исследование использования Web-ресурсов**

Процесс исследования использования Web-ресурсов обычно включает в себя только три фазы:  
1. Препроцессиг  
2. Извлечение шаблонов  
3. Анализ шаблонов.

**Этап препроцессинга**

Для решения задачи исследования использования Web на этапе препроцессинга в массиве анализируемых данных должны быть выделены перечисленные сущности. Это значительно усложняет неполнота данных, получаемых с одного источника.

**Этап извлечения шаблонов**

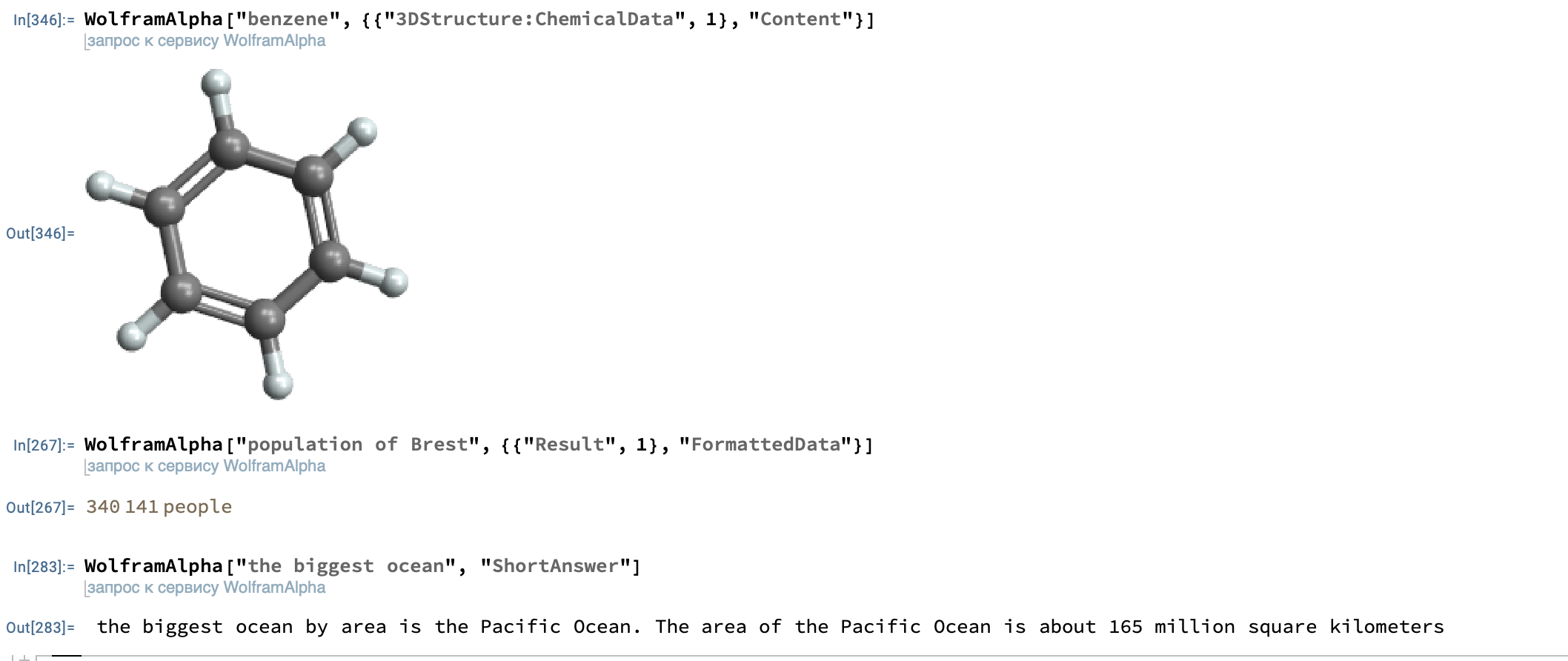
Для извлечения шаблонов из информации об использовании Web-ресурсов применяются различные методы как классической статистики, так и относящиеся к области Data Mining.

**Этап анализа шаблонов и их применение**

Последним этапом в исследовании использования Web-ресурсов является анализ извлеченных шаблонов. Целью анализа является отфильтровать наиболее интересные шаблоны и отбросить ничего не значащие шаблоны. Методология анализа во многом зависит от области применения, в которой он выполняется.

**WolframAlpha**

Oтправляет запрос в адрес Wolfram| Альфа и импортирует выходные данные.

****

**Вывод:**

- Web Mining включает в себя следующие этапы: поиск ресурсов, извлечение информации, обобщение и анализ.

- Различают следующие категории задач Web Mining: извлечение Web-контента, извлечение Web-структур и исследование использования Web-ресурсов.

- В решении задачи извлечения структуры Web используются подходы из области социальных сетей, библиометрики, ранжирования документов и т. п.

- В задаче исследования использования Web анализу подвергаются вторичные данные о взаимодействии пользователя с Web: протоколы работы, куки, авторизация и т. п.

- Существуют два основных подхода анализа использования Web-ресурсов: преобразование данных использования Web-сервера в реляционные таблицы до выполнения адаптированных методов Data Mining и использование информации из файла протокола непосредственно, применяя специальные методы предварительной обработки.