Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

**Лабораторная работа №6**

По дисциплине

«Компьютерное моделирование и анализ данных»

за 7-й семестр

Выполнил:

студент 4 курса

группы ПО-5

Крощук В.В.

Проверил:

Чичурин А. В.

Брест, 2022

**Извлечение знаний из Web — Web Mining**

1) Проблемы анализа информации из Web.

Всемирная сеть сейчас содержит огромное количество информации, знаний. Однако это многообразие данных скрывает в себе проблемы, которые могут возникнуть не только при анализе, но и при поиске необходимой информации в сети 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 потребителя или индивидуального пользователя связана с предоставлением пользователю информации, которая оказалась бы ему интересна. Для этого требуется настройка и персонализация поисковой системы под конкретного потребителя или пользователя.
* Проблема персонализации информации связана с типом и представлением информации в зависимости от смысла, вкладываемого пользователем.

2) Этапы Web Mining.

Web Mining – технология, использующая методы Data Mining для исследования и извлечения информации из Web-документов и сервисов. В Web Mining можно выделить следующие этапы:

* входной этап (input stage) – получение "сырых" данных из источников (логи серверов, тексты электронных документов);
* этап препроцессинга (preprocessing stage) – данные представляются в форме, необходимой для успешного построения той или иной модели;
* этап извлечения шаблонов (pattern discovery stage);
* этап анализа шаблонов (pattern analysis stage) – интерпретация полученных результатов.

Конкретные процедуры каждого этапа зависят от поставленной задачи.

3) Web Mining и другие интернет-технологии.

Web Mining, являясь инструментом для обработки и анализа Web-ресурсов, рассматривается в одном ряду с такими интернет-технологиями, как получение информации (Information Retrieval — IR) и извлечение информации (Information Extraction — IE). Однако, имея с ними много общего, Web Mining имеет также существенные отличия.

Технология IR заключается в получении документов из Web-среды, релевантных запросу пользователей. При этом очень часто полученные результаты включают в себя как релевантные, так и нерелевантные документы. Для решения этой задачи строятся поисковые индексы. Для их построения используются различные методы, включающие в себя моделирование, классификацию и кластеризацию документов, фильтрацию и др. При этом для классификации и кластеризации документов могут быть использованы методы Data Mining (точнее Text Mining).

IE основное внимание уделяет структуре текстового документа, пытаясь извлечь из него ключевые понятия. С одной стороны, для исследования структуры и извлечения основных понятий из текста могут быть использованы методы Text Mining. В этом случае Web Mining является частью IE. С другой стороны, структуризация документов методами IE позволяет сохранять его в реляционной базе данных, что, в свою очередь, позволяет применить к нему методы Data Mining. Таким образом, IE может рассматриваться как технология препроцессинга на одном из этапов Web Mining.

4) Извлечение Web-контента в процессе информационного поиска.

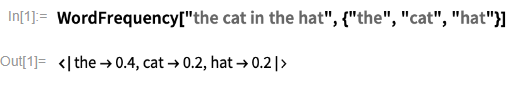
Методы извлечения Web-контента в процессе информационного поиска во многом зависят от типа анализируемых документов. Различают два основных типа:

* неструктурированный - все текстовые документы, не имеющие определенной структуры.
* слабоструктурированный - документы, имеющие структуру, позволяющую вхождение в структурный элемент неструктурированного текста. (HTML, XML и др.)

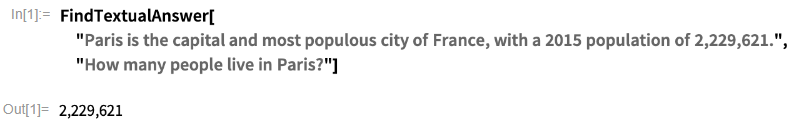
Большинство методов анализа неструктурированного текста использует представление текстового документа в виде множества или вектора слов. При этом в такие представления помещаются отдельные слова без учета их расположения, связи с другими словами, контекста и других лингвистических особенностей. Каждому слову во множестве ставится в соответствие некоторое свойство, которое может иметь или логический тип, отражающий наличие или отсутствие слова в тексте, или числовое значение, отражающее частоту появления слова в тексте. Последующая обработка может быть связана с удалением пунктуации, нечастых слов, стоп-слов и др.

Кроме большого размера модели, векторное представление документов имеет еще один существенный недостаток: оно не обрабатывает синонимы — документы считаются семантически далекими друг от друга, если в них нет одинаковых слов. Данный недостаток устраняется методом скрытой семантической индексации (Latent Semantic Indexing — LSI). В результате документы одной тематики, но которые не используют одинаковые термины, размещаются в одном разделе.

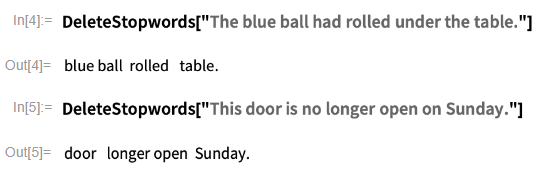
Wolfram предоставляет возможности по определению частоты появления слов в тексте:



Попытка найти ответы на вопросы в тексте:



Удаление стоп-слов:



5) Слабоструктурированные документы.

Извлечение Web-контента из слабоструктурированных документов использует более развитые средства представления текста. Это 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 гиперссылки между документами.

Как и в случае с неструктурированными документами, к полученным представлениям применяются общие методы Data Mining. Область применения методов довольно широка:

1) гипертекстовая классификация;

2) классификации и кластеризации;

3) изучение отношений между Web-документами;

4) извлечение шаблонов или правила;

5) поиск шаблонов и слабоструктурированных данных.

6) Извлечение Web-контента для формирования баз данных.

Язык Wolfram поддерживает огромное количество форматов данных, которые можно читать и обрабатывать, что делает его удобным инструментом для сбора и трансформации данных; очень просто связываться с различными базами данных, в том числе специализированными; легко и эффективно загружаются всевозможные данные из Интернета. Поскольку в исходных данных имеются ошибки, для их устранения можно применить процедуры первоначальной специальной обработки, например, на базе алгоритмов нейросетей.

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:

1) моделирование и формирование запросов к Web;

2) извлечение информации и интеграция;

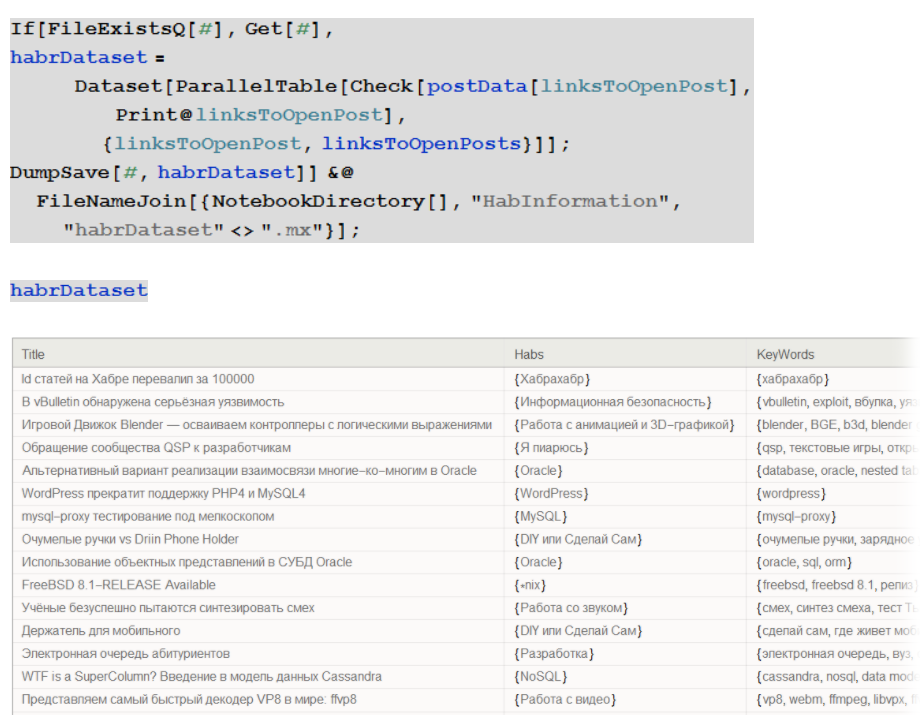
3) создание и реструктуризация Web-сайта.

Обычно методы извлечения Web-контента пытаются выявить структуру Web-документа или преобразовать его для сохранения в базе данных таким образом, чтобы улучшить информационное управление и cделать возможным запрос к нему. С точки зрения размещения Web-контента в базе данных целью является построение модели данных и объединение их таким образом, чтобы поиск мог выполняться не только по ключевым словам, но и по запросам, более приближенным к естественному языку. Этого можно достичь построением схемы Web-документов, формированием хранилища, базы знаний или виртуальной базы данных.

Из-за различий представления Web-документов большинство из используемых методов отличаются от классических методов Data Mining, которые действуют на плоских данных и требуют дополнительной модификации. Исключение составляют методы ILP, которые могут работать с отношениями и графическим представлением данных. Для того чтобы действовать на реляционных или графических данных используют вероятностные алгоритмы для обнаружения схем и для конструкции MLDB (многослойной базы данных), применяют измененную версию алгоритмов построения ассоциативных правил и модифицированную версию логики правил ассоциации.

Необходимо отметить, что существует также большая область, связанная с анализом мультимедийной информации, которая в большом объеме присутствует в Web-документах. Методы, позволяющие извлекать полезные знания из мультимедийной информации, объединяются в технологию Multimedia Mining.

Формирование базы данных постов Хабрахабр:



7) Исследование использования Web-ресурсов.

Процесс исследования использования Web-ресурсов обычно включает в себя только три фазы:

1) препроцессинг.

2) извлечение шаблонов.

3) анализ шаблонов.

Выделяют следующие типы данных, применяемые в задаче исследования использования Web-ресурсов:

1) использование — данные, описывающие использование страниц (IP-адресса, ссылки, дата и время доступа к ним;

2) пользовательские профили — демографическая информацию (пол, возраст, социальное положение и др.) о пользователе, а также регистрационную информацию.

Данные об использовании Web собираются в различных источниках, которые можно разделить на следующие основные группы:

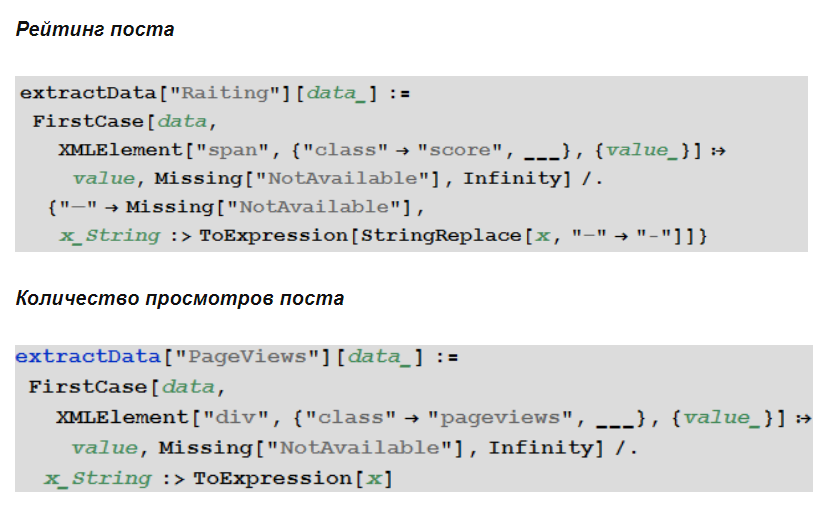
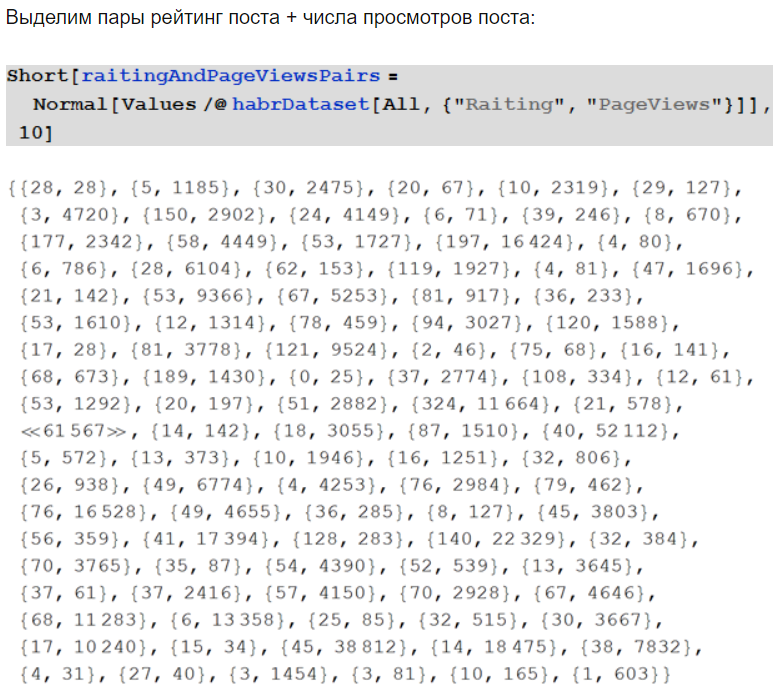
* серверы. Информация для анализа может извлекаться из логов, трафика сервера, куки-файлов, запросов пользователей и др.
* клиенты. Информации о поведении пользователей, от просматриваемых страниц до щелчков мыши.
* прокси-серверы. Промежуточный уровень между клиентом и сервером, используются для кэширования часто запрашиваемых страниц. Могут хранить действительные запросы от нескольких клиентов к различным сайтам, которые не дошли до серверов.

Ни один из видов источников не хранит в себе полной информации, а следовательно, не может рассматриваться как единственный поставщик данных для анализа. Для повышения качества анализа поведения пользователей в нем должна участвовать информация от всех видов источников данных.

8) Этап препроцессинга.

Для решения задачи исследования использования Web-ресурсов на этапе препроцессинга в массиве анализируемых данных должны быть выделены основные элементы, необходимые для дальнейшего анализа, что значительно усложняет неполнота данных, получаемых с одного источника. Кроме этого часто выполняется фильтрация по контексту. Такая задача может возникнуть, например, если аналитика интересует использование Web-ресурсов по определенной тематике. Для такой фильтрации применяют методы анализа Web-контента.

Функции извлечения конкретных данных из символьного XML представления поста (рейтинг и количество просмотров в паре):

** **

9) 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.

Методы статистики часто используются для анализа посещения сайтов и трафиков. Многие инструменты анализа трафика позволяют получить такие характеристики, как наиболее часто посещаемые страницы, среднее время посещения страниц или средняя длина пути перемещения по страницам. Подобные отчеты могут содержать низкоуровневые ошибки, связанные с невозможностью определить неавторизированные точки входа или недействительные URI (унифицированный идентификатор ресурса). Такой вид получаемых знаний может быть весьма полезным для улучшения производительности систем, повышения безопасности систем, решения задач модификации сайтов и обеспечения поддержки для решения маркетинговых задач.

Методы кластеризации в области исследования использования Web-ресурсов применяются как для кластеризации пользователей, так и для кластеризации страниц. Кластеризация пользователей позволяет группировать пользователей с похожим поведением просмотра страниц. Такие знания полезны для того, чтобы сделать выводы о демографии пользователей и выполнении маркетинговой сегментации рынка в электронной коммерции или обеспечении пользователей персональным Web-контентом. Кластеризация страниц позволяет выявить группы страниц с близким по смыслу содержимым. Эта информация полезна для поисковых машин и персональных ассистентов. В обоих случаях пользователям могут предлагаться гиперссылки в соответствии с их запросом и историей запрашиваемой информации.

Методы обнаружения шаблонов в последовательностях применяются для выявления межсессионных шаблонов, в которых элементы следуют друг за другом в упорядоченном по времени множестве сессий и эпизодов. Данный подход может помочь в прогнозировании структуры будущих посещений, что в свою очередь помогает в правильном размещении рекламы, нацеленной на конкретную пользовательскую аудиторию.

10) Этап анализа шаблонов и их применение

Последним этапом в исследовании использования Web-ресурсов является анализ извлеченных шаблонов. Цель анализа - отфильтровать наиболее интересные шаблоны и отбросить ничего не значащие. Методология анализа во многом зависит от области применения. Более общей формой анализа шаблонов является механизм запроса знаний, такой как SQL. Методы визуализации, такие как раскрашивание или графическое изображение шаблонов, могут выделять характерные шаблоны или тренды в данных. Контент или информация о структуре может быть использована для фильтрации страниц, используемых определенным образом, содержащих информацию определенного типа или имеющих определенную структуру гиперссылок.

Выделяют следующие области применения систем анализа использования Web-ресурсов:

* *Персонализация (Personalization)* — обеспечивает для каждого пользователя индивидуальный подход и позволяет давать наиболее эффективные рекомендации пользователям в достижении их целей.
* *Улучшение систем (System Improvement)* — анализ использования Web-ресурсов позволяет рассматривать изменения трафика, обращения к страницам, поведение пользователей. Результаты анализа могут быть применены для разработки политики кэширования, балансировки нагрузки и распределения данных, что позволяет повысить производительность систем. Выявление закономерностей поведения пользователей позволяет выявлять атаки на сайты и тем самым повышать их безопасность.
* *Модификация сайтов (Site Modification)* — анализ использования Web-ресурсов обеспечивает дизайнера сайта обратной связью от пользователей и информацией, необходимой для принятия решения об изменении структуры и его содержания.
* *Бизнес-интеллект (Business Intelligence)* — выполняет анализ информации об использовании данных Web-сайтов и маркетинговой информации. Выделяют четыре задачи, решаемых в этой области: привлечение новых пользователей, удержание пользователей, проведение перекрестных продаж и определение отказа пользователя от Web-сайта.

**Визуальный анализ данных**

1. Выполнение визуального анализа данных.

Анализ данных представляет собой процесс извлечения, представления и моделирования информации, взятой из первоисточника. Визуализация данных очень важна в различных сферах деятельности, так как многие теряются в обилии окружающей информации и на её восприятие и понимание тратится очень много времени.

Система Wolfram Mathematica содержит много различных инструментов и настроек для содержательной и зрелищной визуализации. Модули графики обеспечивают высококачественное статическое или динамическое представление функций, данных, диаграмм, изображений или аннотаций, определяя баланс между скоростью вычислений, вывода и визуальным соответствием виду объекта.

В контексте визуализации данных Wolfram – достаточно мощный инструмент, который предполагает:

* ***Графическое представление структурированных и неструктурированных данных.***Символьная архитектура и динамический интерфейс языка позволяют применять гибкие подходы к визуализации информации. В самом языке присутствует множество автоматизированных функций, которые позволяют загружать данные и тут же строить их графическое представление в динамическом режиме – от простых точечных диаграмм до графов и сетей.
* ***Редактирование и контроль результатов визуализации в интерактивном режиме.*** Инструментарий Wolfram предусматривает множество возможностей контроля над каждым этапом визуализации – от разработки и построения и до окончательного внешнего вида. Унифицированная символьная графическая архитектура языка позволяет программисту одновременно работать по двум направлениям – настраивать генерацию визуального представления в программном коде и интерактивно редактировать готовый результат.
* ***Импорт и экспорт данных*.**Wolfram поддерживает и автоматически обрабатывает данные практически любого известного формата; все они интегрированы в среду программирования в виде символьного обозначения. Для каждого формата данных связь между его представлениями внутри программы и вне ее может быть построена на любом уровне детализации – для этого в языке предусмотрен механизм работы с элементами данных.

1. Характеристики средств визуализации данных.

Mathematica содержит полный набор функций для отображения структурированных и неструктурированных данных в 2d и 3d. Встроены функции для вывода контурных графиков, графиков из точек, линий и поверхностей, векторных графиков и графиков линий тока, гистограмм, двумерных и трехмерных секторных и столбчатых диаграмм, пузырьковых диаграмм, графиков для специализированных областей (финансы, статистика, теория графов и т.д.)

При формировании иллюстраций Wolfram Mathematica автоматизирует с подбором оптимальных вариантов параметры построения графика, обеспечивает корректное отображение деталей поведения функций на всех участках, выбор масштаба. Используемые по умолчанию настройки графики тщательно подобраны, что наилучшим образом подходит для подавляющего большинства случаев. Дополнительно, Mathematica позволяет делать настройки визуализации в соответствии с потребностями пользователя, система предоставляет сотни опций, позволяющих контролировать каждый аспект отображения, включая размер рисунка, метки осей, отображение сетки, разбиение, заливку и т.д. Обычно, пользователю достаточно использовать лишь несколько нестандартных опций за один раз.

Любое средство визуализации может быть классифицировано по параметрам:

* по виду данных, с которым оно работает;
* по визуальным образам, которые оно может предоставлять;
* по возможностям взаимодействия с этими визуальными образами.

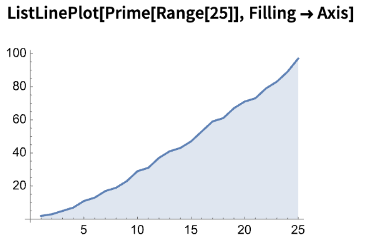
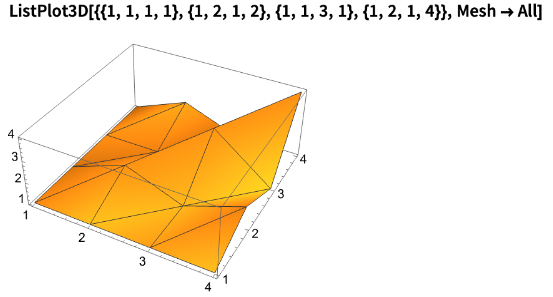
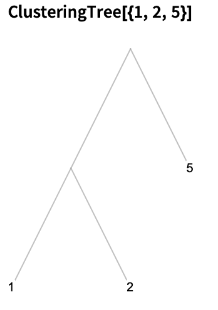
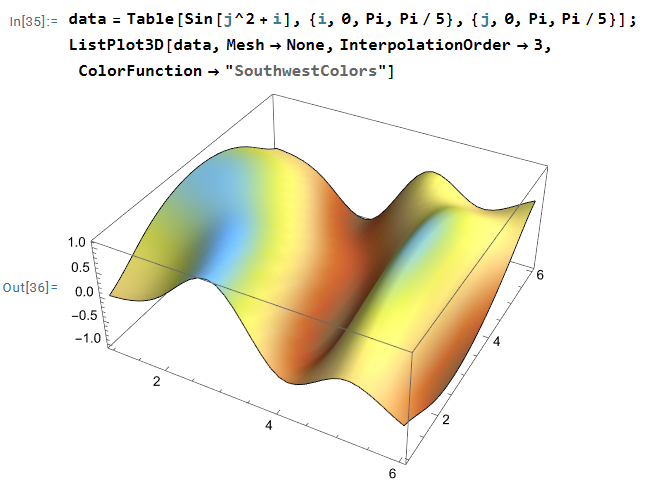
Очевидно, что одно средство визуализации может поддерживать разные виды данных, разные визуальные образы и разные способы взаимодействия с образами.

1. Методы визуализации.

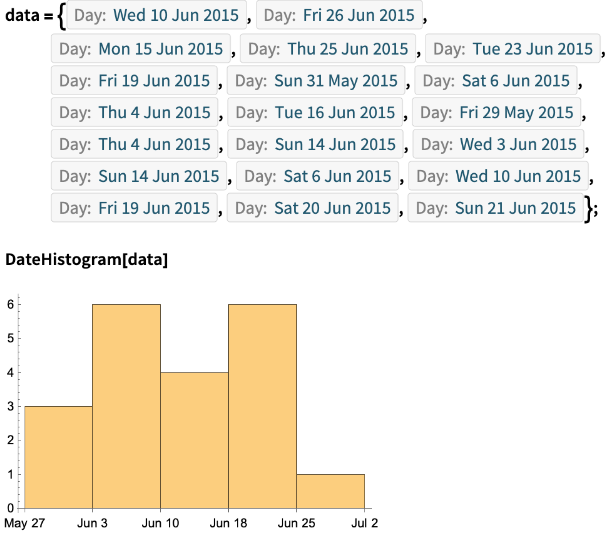
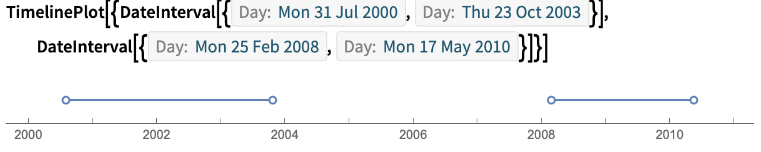
Существует достаточно много методов визуализации, но все они имеют как достоинства, так и недостатки. Основная идея комбинирования заключается в объединении различных методов визуализации для преодоления недостатков одного из них. Различные проекции рассеивания точек, например, могут быть скомбинированы с методами окрашивания и компоновки точек во всех проекциях, что позволяет определить зависимости и корреляции в данных. Такой подход может быть использован для любых методов визуализации. Таким образом, комбинирование нескольких методов визуализации обеспечивает бóльшую информативность, чем в общем независимое использование методов.

Wolfram предоставляет огромное количество возможностей визуализации. Среди них можно выделить визуализацию:

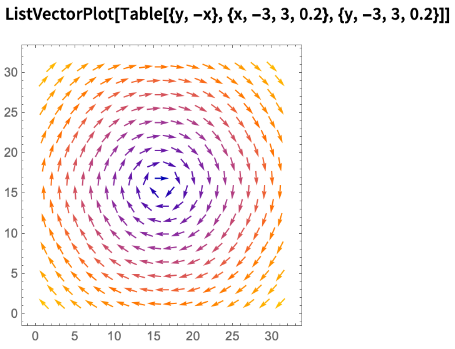
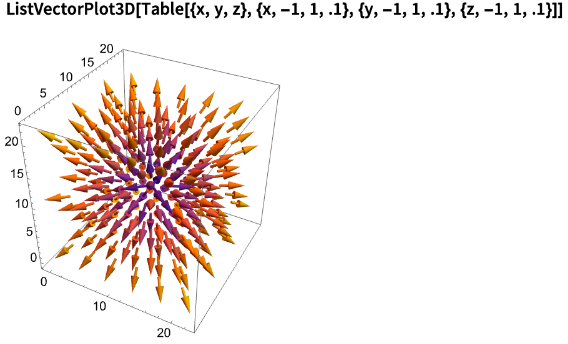
1. Данных в иерархическом и неиерархическом виде

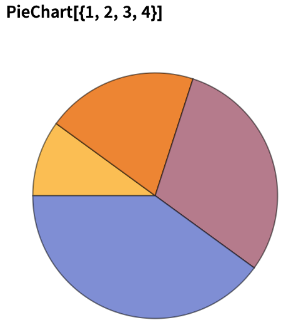
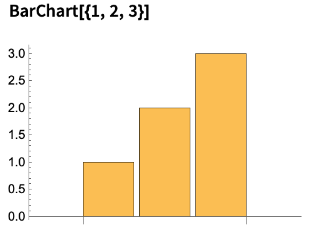
1. Даты и времени

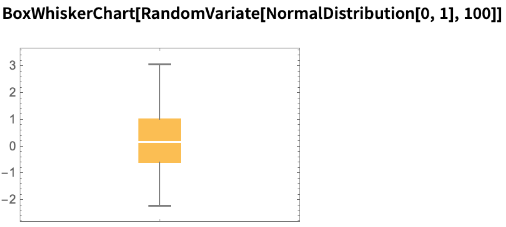
1. Векторов

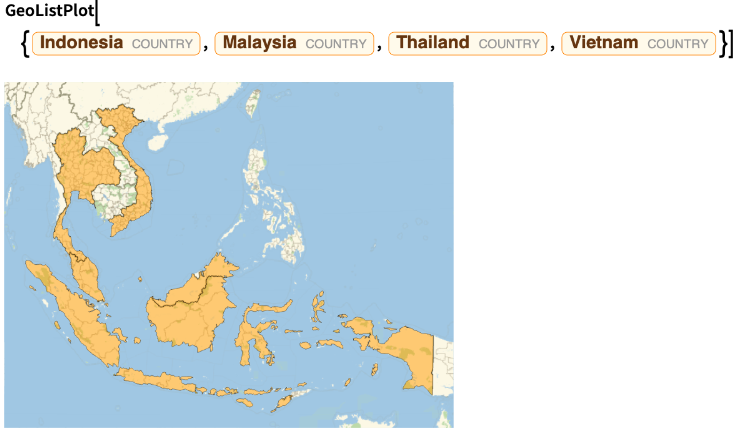
1. Диаграмм

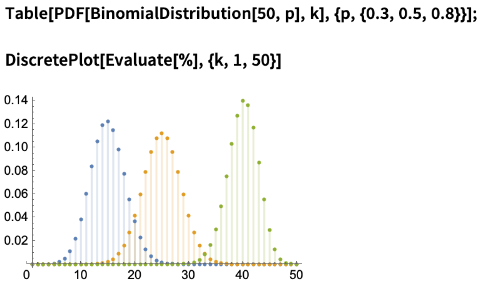
1. Статистических данных



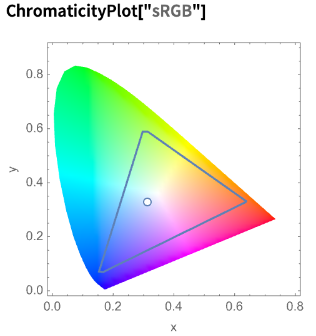
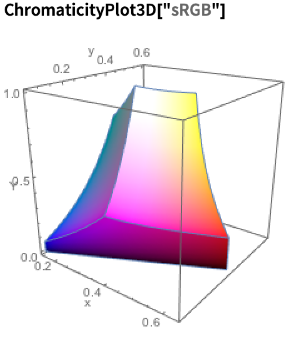
1. Географических положений



1. Дискретных функций

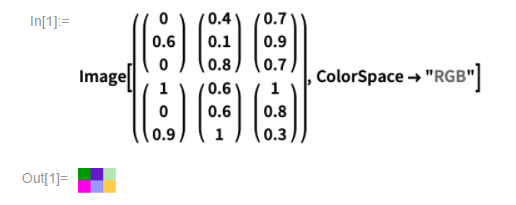
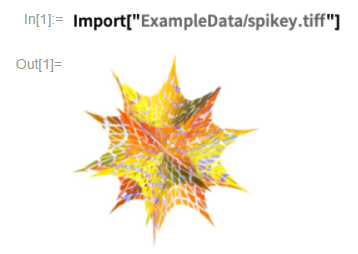


1. Цветового пространства

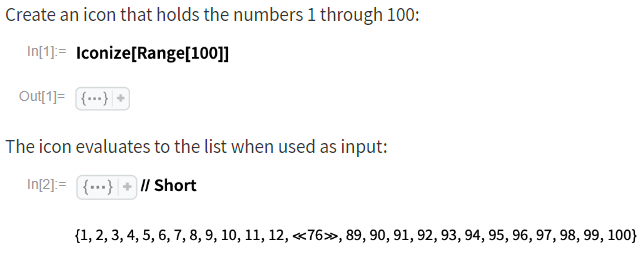
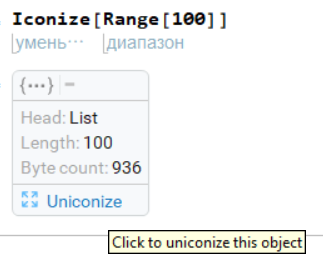
 

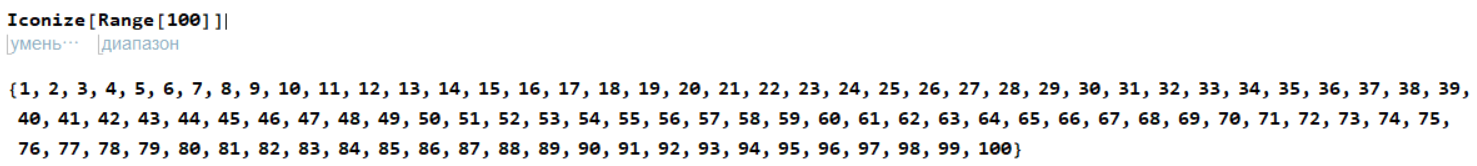
1. Отображение иконок.

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. Wolfram позволяет рисовать иконки из пикселей прямо в коде, импортировать файлы с изображениями, а также использовать их при построении трёхмерных графиков.

Также в Wolfram Mathematica предоставляется способ «отображения в иконку». Он позволяет хранить код и данные, форматированные в форме иконки. После чего данные можно привести и к обычному виду:

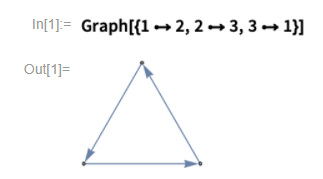
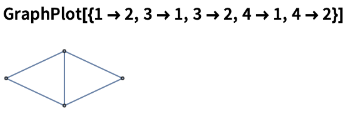


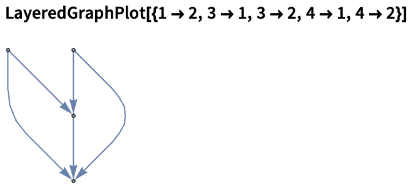
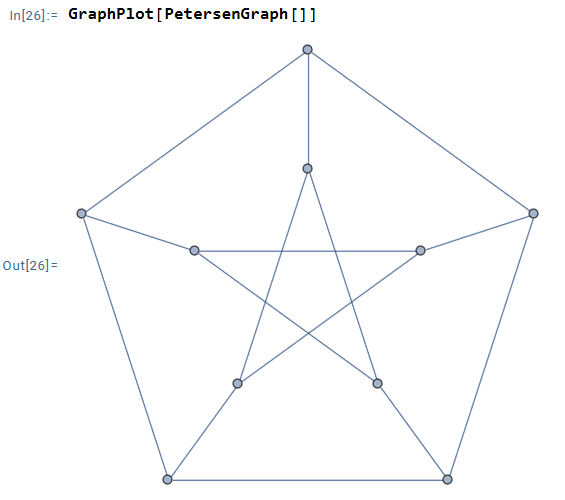
1. Иерархические образы.

Иерархические образы используются для отображения иерархий и отношений в данных. Wolfram позволяет работать с такими типами иерархических данных как:

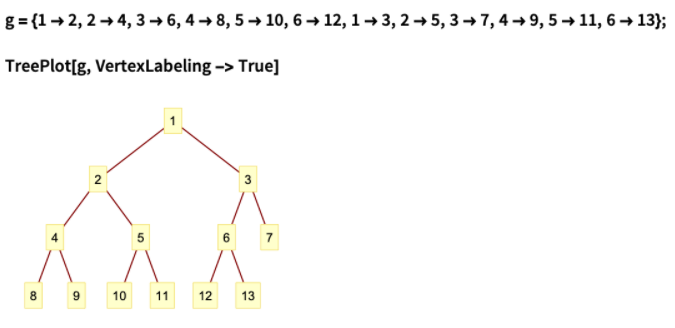
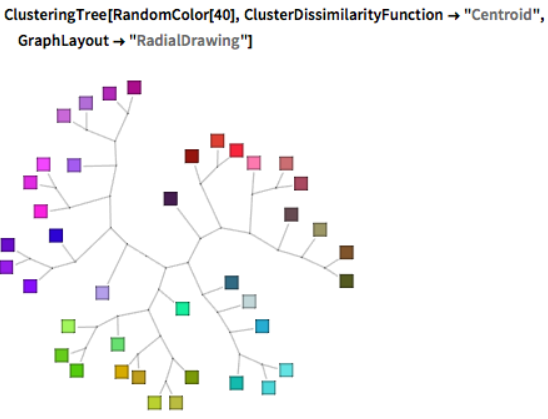
* Графы
* Деревья

Примеры построения графов:

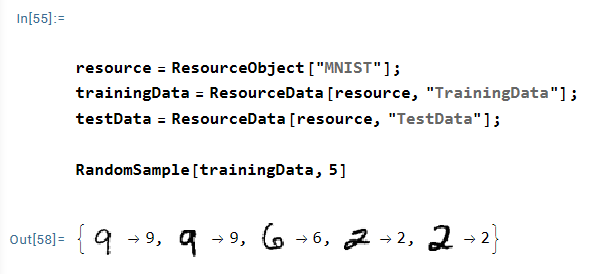


Деревья являются частными случаями графов, следовательно, для деревьев доступны такие же операции, как и для графов. Пример построения дерева:

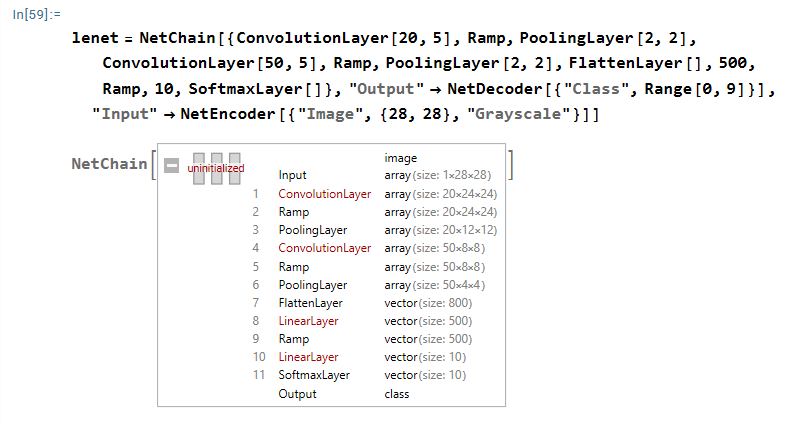
 

Пример нейронной сети:

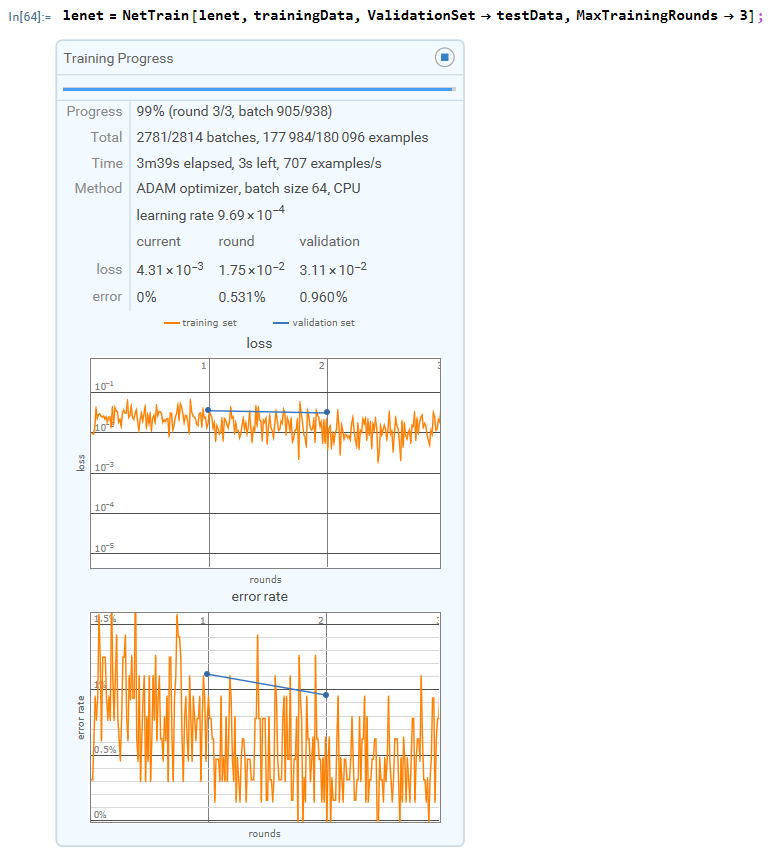
Начнем с получения данных режима обучения и контрольных данных.



Задаем свёрточную нейронную сеть, которая принимает входные данные в виде полутоновых изображений размером 28×28.



Проведем 3 раунда тренировки нейронной сети.



Оценим обученную сеть непосредственно на изображениях, случайно отобранных из контрольного набора.

