ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Инструменты обработки данных

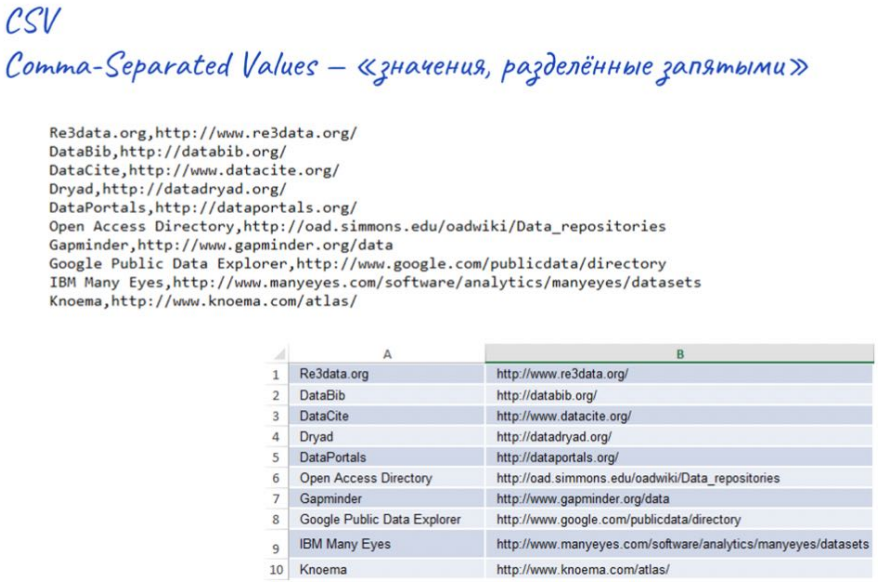
Высшая школа цифровой культуры
Университет ИТМО
dc@itmo.ru

1 Инструменты обработки данных

Когда мы работаем с наборами данных, то нужно простое и удобное средство для его хранения и обработки, так как это довольно трудно делать в уме.

Для обработки данных нужно подобрать подходящее ПО – это будет зависеть от объема данных и от задач по обработке данных, которые мы решаем. Если данных совсем немного, то их можно записать в файле самого простого формата, точнее, совсем без форматирования, создав документ, например, при помощи программы Блокнот. Такие файлы называются файлами ASCII, и сохраняют только символы, введенные туда без каких-либо указаний по оформлению.

Если данные представляют собой неструктурированный текст, то для их обработки созданы специальные текстовые процессоры, например Word. Если же данные состоят из однородных элементов, имеющих явную структуру, то лучшим средство хранения будут электронные таблицы. Часто для



	A	B
1	Re3data.org	http://www.re3data.org/
2	DataBib	http://databib.org/
3	DataCite	http://www.datacite.org/
4	Dryad	http://datadryad.org/
5	DataPortals	http://dataportals.org/
6	Open Access Directory	http://oad.simmons.edu/oadwiki/Data_repositories
7	Gapminder	http://www.gapminder.org/data
8	Google Public Data Explorer	http://www.google.com/publicdata/directory
9	IBM Many Eyes	http://www.manyeyes.com/software/analytics/manyeyes/datasets
10	Knoema	http://www.knoema.com/atlas/

Рис. 1: Comma-Separated Values

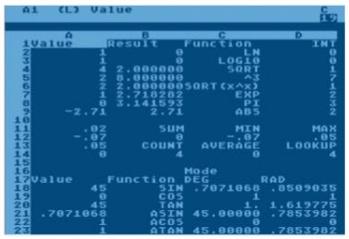
сохранения табличных данных в виде текста используется формат CSV – Comma-Separated Values – значения, разделённые запятыми). Строка таблицы соответствует строке текста, которая содержит одно или несколько полей, разделенных запятыми или другим разделителем, например, точкой с запятой или табуляцией. Многие приложения, которые работают с форматом CSV позволяют выбирать символ разделителя.

Для хранения таких наборов данных были придуманы электронные таблицы. Они помогают нам вычислять значения, упорядочивать и фильтровать

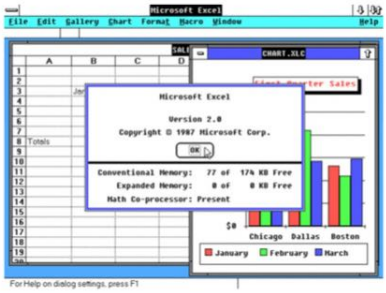
данные, делать разные преобразования, группировать, анализировать и графически представлять различные виды данных.

Первая цифровая электронная таблица – VisiCalc была выпущена в 1979 году. С течением времени цифровые таблицы стали одним из самых популярных видов использования компьютеров. И самым популярным программным обеспечением для работы с электронными таблицами за последние 30 лет является Microsoft Excel. Первая версия Excel была разработана компанией Microsoft в 1985 году. MS Excel является частью пакета MS Office, в который входят текстовый редактор Word, PPT, Outlook, Access и прочие полезные для жизни программы. Таблица Excel могут состоять из нескольких листов.

Первая цифровая электронная таблица - VisiCalc - 1979 год

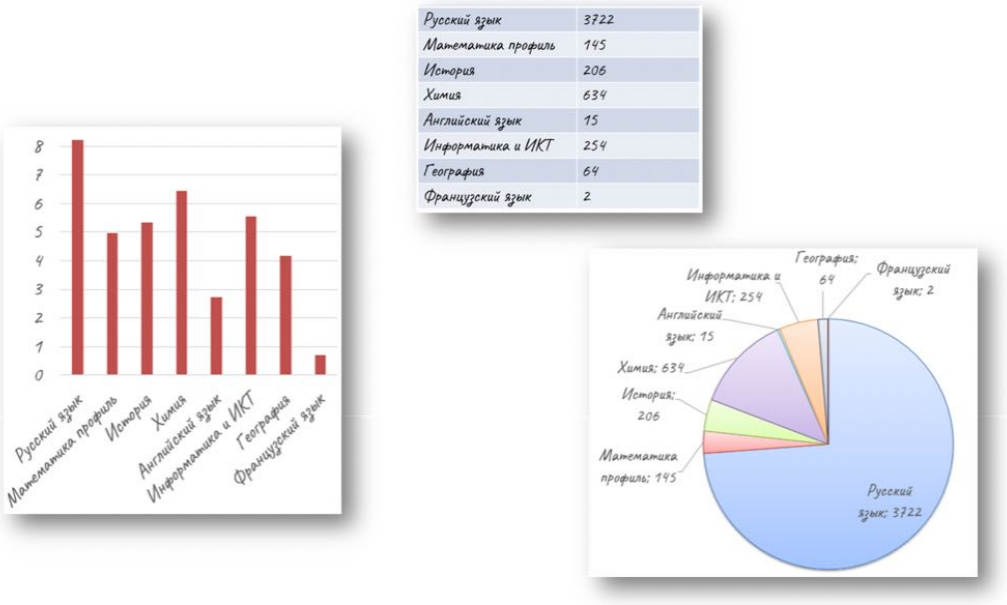


Первая версия Microsoft Excel - 1985 год



Общее количество строк и столбцов на листе в версии 2016 года ограничено 1 048 576 строками и 16 384 столбцами. На первый взгляд, это количество кажется очень большим, но, во-первых, количество данных может быть больше, а во-вторых, с таким большим количеством данных на листе работать очень неудобно.

Если данных становится очень много, они имеют сложную структуру, данные должны надежно сохраняться, и при этом работать с ними нужно множеству пользователей, то для их хранения и обработки понадобятся совсем другие средства – например, системы управления базами данных. В зависимости от наличия или отсутствия четкой структуры в данных, подойдут либо реляционные СУБД, например, Oracle или PostgreSQL, которые замечательно обрабатывают структурированные данные, либо NoSQL хранилища, которые специально разрабатывались работы с неструктурированными или слабоструктурированными данными. Важный шаг в понимании данных дает визуализация. Наглядное представление информации использовать часто



намного удобнее, чем просто ряды цифр. Достаточно большой спектр возможностей для визуализации данных дает Excel, Google-таблицы и другие электронные таблицы.

Если для данных нужна аналитика по неизвестным заранее параметрам или сложная визуализация, например, пространственных или многомерных данных, то для этого есть специализированные пакеты. В качестве примера можно привести Tableau. Tableau специализируется на анализе данных через их визуализацию. Tableau можно назвать системой интерактивной аналитики, позволяющая в быстро проводить глубокий и разносторонний анализ больших массивов информации, используя данные из самых разных источников – это могут быть CSV- или просто текстовые файлы, PDF, файлы баз данных, расположенные на внутренних устройствах хранения или в облачных хранилищах. У Tableau есть несколько платных продуктов. Для работы с открытыми данными можно порекомендовать Tableau Online – это облачная платформа с веб-интерфейсом, которую можно использовать бесплатно, но при условии, что все решения будут храниться на общем сервере и будут опубликованы в открытом доступе. Следующий уровень исследования данных – data mining, или интеллектуальный анализ данных. Такой анализ данных представляет собой систематический и последовательный процесс выявления и обнаружения скрытых закономерностей в больших наборах данных. Для этого разработаны различные алгоритмы машинного обучения.

Чтобы применять их, вовсе не обязательно быть программистом. Можно просто воспользоваться специальным программным обеспечением, среди которых Microsoft Azure, Rapid Miner, Weka. Эти программы позволяют производить очистку и подготовку данных, обнаруживать закономерности и аномалии в данных, строить прогнозы и анализировать тексты. Все эти программ-

ные средства обладают удобным графическим интерфейсом. Можно просто загрузить туда свой набор данных, выбрать нужный алгоритм обработки, применить его – и задача решена.

Но вот какие задачи можно ставить в анализе данных, и какие алгоритмы нужно применять для их решения, и как интерпретировать результат – это большой вопрос. Чтобы в нем разобраться, нужно иметь много знаний. Для этого и предназначен наш курс. Но путь предстоит долгий.

Изучив алгоритмы анализа и обработки данных, может появиться желание применить их при помощи средств программирования. Наилучшим выбором тогда будет язык Python, высокоуровневый язык программирования общего назначения, который довольно легко освоить. Python широко применяется в образовательной сфере, для научных вычислений, больших данных и машинного обучения, в веб- и интернет-разработке, графике, GUI, играх и других направлениях.

Для программирования на Python разработано большое число библиотек, что позволяет легко собрать прикладную программу, подключая функции нужных библиотек.

Но наша задача сейчас – первичная обработка данных, поэтому в качестве инструменты мы сосредоточимся на электронных таблицах.

Обычные электронные таблицы привязаны к одному компьютеру, что затрудняет обмен данными. Кроме того, если ваш файл был случайно удален или потерян из-за сбоя компьютера, восстановить информацию было практически невозможно. Сейчас появилось множество облачных хранилищ данных, к которым можно обращаться с различных устройство и хранить там все необходимые файлы, предоставляя доступ к ним при необходимости другим людям.

В 2006 году Google вывела электронные таблицы в Интернет с помощью пакета Google Docs. Теперь в Google Sheets, или Google таблицах можно

создавать электронные таблицы, работать с ними сразу нескольким пользователям в режиме онлайн и обрабатывать данные с любого подключенного к Интернету устройства.

Google Sheets выглядит и функционирует так же, как и любой другой инструмент для работы с электронными таблицами, но поскольку это онлайн-приложение, он предлагает гораздо больше, чем большинство инструментов для работы с электронными таблицами. Например:

- веб-таблицы можно использовать где угодно, их невозможно забыть дома, как обычный файл на компьютере;
- он работает с любого устройства, с мобильными приложениями для iOS и Android вместе с основным веб-приложением;
- Google Sheets бесплатен и включает в себя Google Drive, Документы Word и Слайды PowerPoint для совместной работы и обмена файлами, документами и презентациями в Интернете;
- он включает в себя почти все те же функции электронных таблиц – если вы знаете, как использовать Excel, вы легко сможете справиться и с Google Sheets;
- вы можете загружать дополнения, создавать свои собственные и писать собственный код;
- он находится в сети, поэтому вы можете автоматически собирать данные с помощью вашей электронной таблицы и делать практически все, что захотите, даже если ваша таблица не открыта.

Справедливости ради надо сказать, что есть еще средства для удаленной работы с электронными таблицами – например, Excel Online. Для этого Вам понадобится учетная запись Microsoft. (Ссылку можно найти в полезных ссылках в ИСУ). Также надо зарегистрироваться в outlook.live.com, чтобы иметь доступ к Microsoft Online.

С Excel Online при использовании веб-браузера для создания, просмотра и редактирования книг хранения в OneDrive или Dropbox. Если ваша организация или колледжа подписка Office 365, начните использование Excel Online, Создание и Сохранение книги в библиотеках на вашем сайте.

Office Online сочетает в себе самые популярные функции Office и возможности совместного редактирования в реальном времени, чтобы вы могли и на учебе, и дома работать в команде над общими документами, презентациями и таблицами.

Кроме того, Office Online взаимодействует с установленными на компьютере приложениями Office – вы можете выбирать удобный способ работы. Используйте Office Online для динамического сотрудничества и совместного редактирования в режиме реального времени. А если у вас уже есть Office, продолжайте работать с полнофункциональными приложениями Word, PowerPoint и Excel, установленными на вашем компьютере Mac или под управлением Windows.

Мы рекомендуем Вам в рамках этого курса использовать Google Sheets, и поэтому рассмотрим их функционал более подробно, но Вы также можете выполнять упражнения в обычном MS Excel или Excel Online, если они Вам больше нравятся.