**ФГБОУ ВО «Кубанский государственный аграрный университет имени**

**И.Т. Трубилина»,**

**Российская Федерация**

**Колесников Роман Юрьевич, ПИ2102**

**roman563412@gmail.com**

**Лабораторная работа №11. По дисциплине "Интеллектуальные информационные системы и технологии".**

**Заголовок**

Анализ и обработка данных для системы Эйдос: Лабораторная работа и решение проблем

**Резюме**

1. Контекст Сессии

Видеозапись фиксирует начало онлайн-лабораторного занятия по дисциплине "Интеллектуальные информационные системы и технологии" для группы ИТ-18-23, проводимого 25 ноября 2020 года с 8:00 до 9:30. Занятие ведут профессора Луценко Е.В. и Рашинов Г.А. Тема занятия – продолжение разработки собственных приложений в системе Эйдос (лабораторная работа №11).

2. Основная Тема: Разработка Приложений и Оценка Усвоения

Профессор Рашинов отмечает, что занятие посвящено разработке студентами собственных приложений. В связи с отсутствием вопросов от студентов, он планирует провести опрос для оценки усвоения материала, так как приближается сессия. Приоритет отдается вопросам студентов по их проектам.

3. Анализ Данных Студентки (Анастасия)

Студентка Анастасия прислала ссылку на данные для анализа (связанные с бронированием отелей, спросом, количеством мужчин/женщин). Профессор Рашинов начинает анализ полученного файла (hotel\_bookings.xlsx, ~20 МБ, конвертированный из CSV ~16 МБ).

4. Выявленные Проблемы с Данными и Предложенные Решения

Размер файла: Файл размером 20 МБ слишком велик для загрузки и обработки в системе Эйдос (ограничение FTP, длительное время обработки). Идеальный размер – до 5 МБ, максимум – 10 МБ.

Формат данных:

Наличие некорректных символов (нижние подчеркивания) в названиях колонок мешает онлайн-конвертерам.

Несколько колонок содержат даты в абсолютном формате (год, месяц, дата прибытия), которые бессмысленны для анализа в Эйдос и должны быть удалены. Колонки с относительными датами (номер недели, день месяца, длительность пребывания) полезны.

Колонка "agent" содержит числовые ID и значения "NULL", что система воспринимает как текстовую шкалу, а не числовую. Значения "NULL" нужно заменить на 0.

Отсутствует целевая колонка с суточной ставкой (ADR), хотя она упоминалась как цель анализа.

Цель анализа: Предсказание получения отелем непропорционально большого количества особых запросов. Колонка "total\_of\_special\_requests" подходит как классификационная шкала (0-5).

Решение: Профессор предлагает удалить ненужные колонки (абсолютные даты, год), заменить "NULL" на 0 в колонке "agent", сократить объем данных (оставить ~40 000 записей из ~119 000), чтобы файл стал приемлемого размера (< 10 МБ). Он сам обработает файл и вышлет его Анастасии.

5. Вопросы по Системе Эйдос и Обновлениям

Обсуждается необходимость обновления системы Эйдос. Студентам рекомендуется использовать режим 6.2 для доступа к ссылке на обновление и заменить файлы системы. Отмечаются ограничения текущей реализации (написанной на Delphi/Object Pascal) по сравнению с потенциальной реализацией на C#, особенно в скорости обработки больших данных и визуализации.

6. Следующие Шаги и Заключение

Профессор Рашинов обработает данные Анастасии и отправит ей для дальнейшей работы. Студентам напоминают о необходимости задавать вопросы и готовиться к сессии. Подчеркивается важность правильной подготовки данных для анализа в системе Эйдос.

Детальная Расшифровка

1. Контекст Сессии

Введение и приветствие

(0:01) Доброе утро. Георгий Александрович, доброе утро, ребята.

(0:09) Доброе утро.

(0:14) Ну что? Сегодня у нас аж 25 ноября 2020 года. Первая пара, 8:00-9:30.

Параметры занятия

(0:27) Лабораторная работа номер 11 с группой ИТ-18-23 по дисциплине "Интеллектуальные информационные системы и технологии".

(0:42) Занятия ведут профессор Луценко Евгений Вениаминович и профессор Рашинов Георгий Александрович.

(0:51) Вот. А с кем вести, не очень понятно.

2. Основная Тема: Разработка Приложений и Оценка Усвоения

Тема занятия

(0:59) И у нас тема занятия сегодняшнего - это продолжение разработки вами собственных приложений.

Опрос студентов

(1:15) И заодно я сейчас начинаю вас опрашивать, как вы усвоили материал. Потому что вы ничего не спрашиваете, а потом, когда вас опрашиваешь, то выясняется, что вы ничего вообще и не знаете. Ну, некоторые знают, если честно.

(1:37) Естественно, какие у вас могут возникать вопросы, если вы вообще непонятно где там, на Марсе витаете в облаках.

Приоритет вопросам студентов

(2:20) Поэтому, поскольку у нас такая тема, то приоритет отдаётся вашим вопросам. То есть, если у вас у кого-то есть вопросы какие-то, то мы поможем.

План при отсутствии вопросов

(2:36) Если же вопросов нет, то я начинаю тогда, скажем так, промежуточное тестирование. Уже приближается сессия. Наверное, вы в курсе, что осталось не очень много занятий уже. Вот. И поэтому нужно уже подводить какие-то итоги.

3. Анализ Данных Студентки (Анастасия)

Получение данных от студентки

(3:13) Ну и сначала вопрос такой: кто-нибудь вам из вас посылал мне письма с этими исходными данными, там, описаниями или нет?

(3:20) А, я вам вчера прислала вечером ссылку, чтобы вы посмотрели. Мы на прошлом занятии с вами разбирали, и вы сказали, что там некорректные данные на выводе.

(3:31) Что-что? Корректно, некорректно?

(3:33) Некорректные данные, то, что мы мне присылали по, сейчас скажу точно, спрос на... по людям там было, сколько человек, мужчины и женщины. Вы сказали, что там некорректные данные, попросили посмотреть, и если нет, то выбрать другую тему. Я вам другую тему прислала.

(3:52) Угу. Сейчас посмотрим.

Первичная оценка файла

(4:14) Ну, прежде всего, что я могу тебе сказать, Анастасия? Тут, видишь, здесь данных мегабайт. Мегабайт - это перебор. Это будет долго обрабатываться.

(4:34) Так, ну сейчас посмотрим мегабайт чего.

4. Выявленные Проблемы с Данными и Предложенные Решения

Цель анализа данных

(5:05) Здесь, э-э, какие-то не особые запросы. И они хотят предсказать, получит ли отель непропорционально большое количество особых запросов. Тут у них особые запросы.

Проблема с форматом дат

(6:29) Вот здесь вот, видишь, Анастасия, здесь есть много колонок, связанных с датами. Вот год приезда, дата, месяц, дата прибытия, номер недели, день... Вот номер недели, номер, число дней - это всё можно обрабатывать. А вот, э-э, дата этих, надо поудалять сразу из исходных данных. Вот, э-э, длительность пребывания - это одно дело, а когда вот он прибыл, вот с конкретная дата, она бессмысленна. Вот здесь три колонки таких. А вот, э-э, продолжительность пребывания, если вот он прибыл там... А здесь есть номер недели, там, номер дня в году - это имеет смысл. То есть сезонные характеристики. А когда просто дата - это надо всё преобразовывать.

Анализ колонки "Особые запросы"

(7:22) Ну и что мы можем отсюда узнать? Где здесь эти вот, э-э, особые, особые заказы эти, бронирования? Всего специальных запросов. Вот это вот может быть классификационной шкалой.

Отсутствие целевой колонки (ADR)

(8:10) Значит, когда, когда лучше всего бронировать номер в отеле или продолжительность пребывания, чтобы получить лучшую суточную ставку? Так, где тут суточная ставка? Привет. Соотношение между датой... А здесь нету суточной ставки.

Проблема с онлайн-конвертером

(8:41) Ладно, сейчас посмотрим. (Пытается открыть файл)

(9:14) Вот здесь есть... (Файл не открывается корректно)

(10:00) Значит, я сейчас буду конвертировать этот файл с помощью онлайн-конвертера.

(11:40) (Конвертер не работает)

(18:06) Однако, плёный конвертер, который мне очень нравится, вообще-то. Ну ладно тебе.

(20:18) Из-за того, что там нижние подчёркивания, он не будет переводить.

(20:22) Это, это я понимаю. Да я, это я догадываюсь.

Повторный анализ колонок и данных

(18:27) Так, значит, я что? Вот это вот конкретную дату, думаю, надо её убирать. Хотя вот месяц можно... Ну, месяц... Год, год, наверное, не нужно, а вот месяц можно оставить. Это неделя в году, да? День месяца. Длительность ночей.

(21:15) Всего специальных запросов.

(21:33) Вот это вот может быть классификационной шкалой.

(21:47) Вот здесь есть запросы специальные. А там, на Кагле, почему-то не было видно их, что они есть.

Проблема с колонкой "Agent"

(22:13) (Находит колонку "Agent") Да, ну здесь... Вот, видите, она почему надо было это самое, конвертировать XLSX? Потому что здесь, видите, ограничение сработало на число этих записей.

(28:00) (Прокручивает данные)

(35:51) Вот тут, ребята, смотрите, вот это вот у нас числовая шкала, которая она не поняла, что она числовая. Агент. Видите? Почему? Потому что там есть символы, скорее, то есть не цифровые символы.

(36:14) Ну сейчас мы это увидим. (Ищет нечисловые значения)

(38:05) Вот здесь вот надо нули убирать. Вот так я. Так, ну ладно. Короче, там, где данных нет, ребята, вот у нас там написано null. Вот мы там ноль и напишем. Тогда эти шкалы станут числовыми. Вот, стало числовой. Вот.

Итог по данным Анастасии и решение

(24:31) (Профессор обращается к Анастасии)

(25:06) Ну вот сразу тебе скажу, Анастасия, что экселевский файл получился больше, чем помещается в... Значит, здесь что? 119.000. Чем, ну, в этом старом Экселе он не помещается, но беда в другом. Он занимает больше 10 МБ. Всё, он не может быть записан в FTP на FTP сервер системы Эйдос, к сожалению.

(29:13) Поэтому мы будем просто определять вот, э-э, специальные эти бронирования. Ну, для этого надо что сделать? Сначала добавим колоночку вот сюда. (Добавляет колонку) Вот. Потому что это не классификационная шкала. То есть, ну, то есть, вернее, это не шкала вообще. левая колоночка. У нас такой нет колоночки, которая не является шкалой. Поэтому мы её сделали. Вот. И, пожалуй, надо вот этот колоночку статус перенести в другое место, чтоб последняя была колонка классификационная. Она может быть либо первой быть, либо последней. Вот. Ну, нам удобнее последней сделать её. Всё. Теперь смотрим ячейки считаем. Колонки. Короче говоря, со второй по тридцатую - это у нас описательная шкала, а тридцать первая - это у нас классификационная шкала.

(31:46) Ну и так вот.

(32:00) 20 МБ. Его в облаке, в Эйдос-облаке не разместите. Ну, можно попробовать, знаете, что сделать? Взять какое-то подмножество, чтоб этот файл исходных данных, чтоб он был меньше 10 МБ. Ну, скажем, мегабайт 5, может быть, вот так вот сделать. То есть посмотреть, какой он размер, и сколько надо там уменьшить. Мы же знаем же, видим это. Можно догадаться. Здесь он 20 МБ, то есть надо в два раза уменьшать его, грубо говоря. Ну лучше больше, чем в два раза.

(33:27) Вот. У нас тут 119.000 записей. Вот. Ну если мы оставим, скажем, 50, ну тогда он будет работать.

(34:45) Хорошая задача. Не хотелось бы удалять. Ну придётся, наверное, всё-таки где-то треть оставить. Ну это вы сами сделаете тогда уже.

(42:30) Ну что, Анастасия, видишь, да, как получается, когда... Скажу, что, конечно, здесь на самом деле не 1 МБ данных. Ты поняла уже, да? Там 16 МБ CSV файл сам. Это шутка. То есть это не соответствует действительности. А экселевский файл получился 20 МБ. Что там, ну никак не 1 МБ. Тогда имеется в виду, что архив 1 МБ, это может быть текстового файла.

(43:30) (Профессор резюмирует план) Я что? Вот это вот конкретную дату, думаю, надо её убирать. Год, год, наверное, не нужно, а вот месяц можно оставить. Вот здесь вот надо нули убирать (в колонке Agent). Короче, там, где данных нет, ребята, вот у нас там написано null. Вот мы там ноль и напишем. Тогда эти шкалы станут числовыми. Вот. Ну, можно попробовать, знаете, что сделать? Взять какое-то подмножество, чтоб этот файл исходных данных, чтоб он был меньше 10 МБ. Ну, скажем, мегабайт 5, может быть, вот так вот сделать. То есть посмотреть, какой он размер, и сколько надо там уменьшить.

(43:30) Я тебе пошлю, когда досчитается это всё, вот это я сделаю то, что сказал. Минимизирую систему, то есть удалю из неё лабораторные работы, лингвистические базы данных, языковые базы данных, базу лемматизации. Ну, то есть минимизирую инсталляцию, заархивирую и пошлю тебе. И ты сможешь на своём компьютере установить на диске C это всё в папочке EidosX. И эта модель у тебя будет работать. То есть тебе не придётся вот это всё считать самой повторно.

(41:20) Хорошо, спасибо.

(41:24) Вот. Ну это будет уже после занятия, в перерыве, наверное, скорее всего. То есть мы сейчас не увидим, какая достоверность модели во время занятия.

5. Вопросы по Системе Эйдос и Обновлениям

Проблема с запуском Eidos

(9:47) Ребят, у вас кто-нибудь запускает Start Eidos файлик? Или исполнимый модуль?

Рекомендация по обновлению

(9:57) Короче, вам надо, видимо, обновить систему, но вместе, значит, с файлом Start Eidos надо обновить. И для этого надо режим 6.2 запустить. Понятно. Режим 6.2. А здесь шестой пунктик в режиме 6.2. И там есть ссылочка на обновление. Скачать обновление и заменить все файлы.

6. Следующие Шаги и Заключение

Обработка данных профессором (см. раздел 4)

Ожидание завершения расчетов

(44:08) Ну ладно. Это уже получше будет. Хоть поместится. Вот теперь смотрим внимательно, какие параметры ввода. 2 3 2 2. Нули делать значащими. Равные интервалы. Должно побыстрее всё это тогда работать. А значащие, потому что эти вот кодирования там идёт особый, не особый, там сколько этих особых? Ноль тоже играет роль при этом. Ну сейчас такая вот более-менее разумная получается задача. 7 МБ файл исходный. Шкалы будут числовые, то есть размерность модели значительно меньше будет, чем получилось вот тогда у нас в прошлый раз. Быстрее это всё будет преобразовываться, кодироваться. Вот, видите, 10 градаций классификационных шкал и 687... Ну это многовато, конечно, описательных шкал. 655. Так. 687. Ну это роли не играет, это не принципиально.

(46:30) Выход отсюда один, ребята, C# использовать для реализации системы. Она будет гораздо быстрее работать тогда, чем вот на этом языке для бизнеса. Ну, так получилось, что я на нём оказался, застрял. Потому что когда-то с него начинал, наизусть его знаю, и другие мне осваивать долго и переделывать долго. Очень большой объём работы. Поэтому я и продолжил на нём тоже. Хотя я знал, какие у него недостатки, но что делать? То есть ту систему, которая сейчас есть, её можно как прототип рассматривать какой-то какой-то более хорошей будущей системы, которую не я буду делать.

(48:12) Тут ещё я сообщения выводил, чтоб хоть понятно, что оно что-то делает там. Я тестировал на больших данных, и вот вот я как раз и обнаруживал, что какая-то пауза непонятная. Надо что-то вывести какую-то информацию. Вот эти 10 раз в секунду, они независимы для всех значений, которые здесь отображаются: прогнозов времени, прошедшего времени, оставшегося. Вот это всё, сама полосочка, этот прогресс-бар. Всё это отображается 10 раз в секунду, и одно от другого совершенно независимо. Самой визуализации.

(51:37) Ну побыстрее, конечно. Не 4 часа. Ну, может, там минут 15, я не знаю, получится. Ну всё равно.

(52:00) (Расчеты продолжаются)

(53:36) Вот. Вот это уже получше. И сейчас мы проверим, тут у нас числовые, не числовые. Вот они у нас нормально здесь воспринимаются. Значит, и система Эйдос нормально тоже воспримет их как числовые. А они не воспринимались как числовые по той причине, что там пустые строки потом идут текстовые.

(54:28) Так. 2 3 2 2. Нули считать значащими. Специальные какие-то заказы, равные интервалы. Ну сейчас такая вот более-менее разумная получается задача.

(55:21) Вот. Видите, 10 э-э градаций классификационных шкал и 687... Ну это многовато, конечно. 655. Так. 687. Ну это роли не играет, это не принципиально.

(56:30) Прогноз времени уменьшается имени исполнения. Думаю, что 10:16 закончит. В это время будет уменьшаться сейчас потихонечку.

(57:55) В общем, нужно скопировать обязательно в другой лист. Вот, вот это уже получше. И сейчас вы проверим, тут у нас числовые, не числовые. Вот они у нас нормально здесь воспринимаются. Значит, и система Эйдос нормально тоже воспримет их как числовые.

(59:59) Вот. Ну, в принципе, она вот... Ну вы можете себе представить, что было бы там 120.000, ну было бы в три раза дольше все вот эти операции, в три раза дольше были бы. Ну, как долго, конечно, но приемлемо ещё. То есть ещё можно, скажем, посчитать. Вот вечером вот так запустить, утром прийти, будет всё посчитано. Ну, смысле, на исходной базе данных вот этой оригинальной. То есть вот 100.000 нам - это ещё как бы, ну, реально можно посчитать. А если там будет 4 млн, ребят, то это уже всё. Тогда вариант единственный такой: синтез модели отдельно сделать, а проверку достоверности сделать на каком-то количестве, которое посчитать следующим образом. То есть можно сделать проверку, допустим, достоверности на 100 случайных этих объектах. Заметить, сколько это заняло времени. Там будет, кстати, вот это время исполнения будет написано, что это заняло там 6 минут 42 секунды. Вот вы пишете, допустим, 100 объектов, проверка на достоверность на 100 объектах заняла 6 минут 40 2 секунды. Проверка на X объектах заняла 480 минут. То есть вы можете посчитать, сколько нужно, сколько можно задать объектов для проверки достоверности, чтобы она посчитала за 8 часов это. Понятно, да? И спокойно себе вечером запускаете там 35.000, например, там или 15.000 объектов выборки. Она довольно-таки неплохо оценит достоверность модели. Генератор случайных чисел там очень хороший. Я сделал его с внутренним источником энтропии. Раньше делал генератор фон Неймана с внешним источником энтропии, а сейчас сделал с внутренним. А в качестве источника энтропии я использую э-э несколько разрядов э-э времени, прошедших от начала прошедшего от начала суток в сотых долях секунды. Вот прошло там какое-то там число четырёхзначное. Вот я беру там или десятизначное, там девятизначное. Вот я беру четыре разряда из этой серединки этого числа. И это не само случайное число, а это коэффициент, на который умножается случайное число, посчитанное с помощью генератора встроенного в язык программирования генератора случайных чисел. Этот встроенный, он имеет э-э недостатки, то есть у него циклы есть, называемые. То есть через какое-то время он начинает повторяться, короче говоря. А когда вот такой приёмчик используешь, то он, в принципе, не может никогда повториться. Пусть он повторяется, допустим, но он повторяется в разное время. И раз он повторяется в разное время, то будет отличаться значение чисел. То есть цикл исчезает. А это время, оно меняется очень своеобразным образом. А ещё я не само время использую, а квадрат этого времени или куб, я сейчас не помню уже. Вот. Поэтому там числа меняются очень хорошо. Если само бы время, то там они могли бы расти, э-э рост наблюдался бы на определённых периодах. А так никакого роста, ничего, там совершенно равномерное распределение идеальное. А почему я это сделал? Потому что я делал, смотрел случайные модели. При увеличении объёма выборки случайная модель должна стремиться к нулю. Все все коэффициенты в этих вот матрицах модели, системно-когнитивной, должны стремиться к нулю. А они не стремятся, и даже при очень больших объёмах выборки, около 100.000 примеров, которые с помощью генератора случайных событий встроенного в язык программирования, да, вот. Вот, э-э, ну система спокойно находит там закономерности, понимаете? То есть там есть закономерности, она их находит. Как, э-э, не то, что получалась модель очень высокой достоверности, но видно, что она нашла закономерности, и видно какие закономерности, понимаете? То есть это не случайные числа, короче говоря. Вот. И вот когда я вот это сделал, тогда эта особенность исчезла. То есть она стала хорошо стремиться к нулю при увеличении объёма выборки. То есть стало работать хорошо в этом смысле. Ну я убедился, что случайные модели, они действительно являются случайными с этим генератором. Нигде в литературе не описаны генераторы случайных событий с внутренним источником энтропии, которые я предложил на основе обобщения генератора фон Неймана. То, что я сейчас вам рассказал, это вообще нигде не написано. Это только вот в системе Эйдос есть несколько строчек, которые это делают. Может, написать? Можно статью написать, провести исследования.

Завершение занятия

(88:35) Ну что? Анастасия, тогда я тебе пошлю, когда досчитается это всё.

(89:14) Вот, ребят, на этом наше занятие заканчивается. Вот. А Дима, ну тогда до следующего раза.

(89:22) Хорошо.

(89:23) Да.

(89:25) Ну можешь сам попробовать. Можешь и сам попробовать так же точно сделать всё. Вот. Ну, конечно, данные надо брать какие-то более разумные по размеру. Я написал там в инструкции, что не мегабайты, а килобайты. А мы сейчас взяли всё-таки 20 МБ исходные данные в экселевском файле. Всё, ребят, на этом занятие заканчивается. До свидания. Всего самого хорошего.

(89:51) Всего доброго. До свидания.

(89:52) До свидания.

(89:53) До свидания, Георгий Александрович.

(89:55) До свидания.

(89:56) До свидания.