***ФГБОУ ВО «Кубанский государственный аграрный университет имени И.Т. Трубилина»,***

***Российская Федерация***

**182 Практическое занятие №2. По дисциплине Теория информации, данные и знания. 2020-11-05**

Заголовок: Практическое занятие по анализу данных: Подготовка CSV/Excel и разбор студенческих работ

Резюме текста:

На видео зафиксировано практическое онлайн-занятие №2 (или №5, есть путаница) для группы ИТ20-03 по дисциплине "Теория информации, данные, знания", проводимое профессором Луценко Евгением Вениаминовичем 5 ноября 2020 года.

Основные темы и действия:

Организационные моменты: Обозначены дата, время, группа и дисциплина. Профессор отмечает высокую загруженность, из-за которой проверка работ по почте затруднена, и предлагает разбирать работы студентов непосредственно на занятии.

Телефонный разговор: Профессор прерывается на звонок, обсуждая с кем-то процедуру подачи научной статьи в журнал (декабрьский выпуск, так как ноябрьский утвержден), необходимость прислать статью на проверку и отсутствие необходимости в рецензии для внутренних публикаций.

Разбор работы студента Марка:

Проблема: Марк прислал данные, но возникли проблемы с файлом Excel, который некорректно конвертировался или отображался (смешение чисел и дат, некорректные символы).

Решение: Профессор показывает, как работать с CSV-файлом. Исходный файл Excel оказался непригодным из-за смешанных форматов и некорректного преобразования. Профессор демонстрирует очистку CSV-файла: замену разделителя (точка с запятой на запятую), обработку пропусков (замена пустых ячеек на нули), обеспечение единообразия числовых форматов (одинаковое количество знаков после запятой). Данные готовятся для загрузки в систему АСК-Анализ. Настраиваются параметры анализа: указываются описательные и классификационная шкалы, количество интервалов для градации.

Разбор работы студента Шамиля:

Проблема: Шамиль также столкнулся с ошибкой при обработке файла, вероятно, при конвертации CSV в DBF с помощью встроенного конвертера системы. Ошибка связана с некорректным форматом CSV (разделитель - точка с запятой) или неверно указанным путем к файлу.

Решение: Профессор рекомендует исправить разделитель в CSV-файле (заменить точку с запятой на запятую) и попробовать снова. В качестве альтернативы предлагается использовать онлайн-конвертер (Convertio), если встроенный не справляется. Обсуждается важность правильной подготовки текстовых данных (например, слитное написание названий штатов/регионов, чтобы они воспринимались как единый признак).

Общие рекомендации и публикация:

Подчеркивается важность корректной подготовки данных для анализа.

Обсуждается проблема отсутствия у студентов корпоративной почты, необходимой для регистрации на ResearchGate. Предлагается вариант публикации через старосту (если он зарегистрирован) или через профессора (включая его в соавторы).

Итог: Занятие посвящено практическим аспектам подготовки данных для анализа в системе АСК-Анализ, разбору типичных проблем с форматами файлов (Excel, CSV) и их решению, а также настройке параметров моделирования.

Детальная расшифровка текста:

1. Введение и организационные моменты

Здравствуйте, ребята.

Здравствуйте.

Здравствуйте.

Да, здравствуйте, здравствуйте.

Ну что, у нас сейчас э-э 5 ноября 2020 года. Четвертая пара, которая идет с 13:50 до 15:20. И практическое занятие номер два с группой ИТ20-03 по дисциплине Теория информации, данные, знания.

Занятие ведет профессор Луценко Евгений Вениаминович.

2. Обсуждение практических заданий и проверка работ

Вот. Теперь смотрим, что ж у нас там по практическим занятиям.

Ну, в общем, сейчас у нас, ребята, самое актуальное – это разработка собственного приложения. И мне приходят много писем по этому поводу, чтобы я проверил. А у меня нет возможности, потому что я… У меня вот сегодня, скажем, семь пар подряд. Еще надо там другую работу выполнять. То есть получается, что довольно-таки большая нагрузка, я просто не успеваю, если честно.

Сейчас, значит, 20:03. Сейчас я вас отмечу всех. Сам-то я отметился уже.

Вот. Практическое занятие номер пять, кстати, а не номер три. Номер пять. На четвертой паре.

Поэтому давайте сейчас, у кого, кто мне присылал письма, сейчас в чат шлете мне…

3. Телефонный разговор (Подача статьи)

(Телефонный звонок)

Да, привет, Виталий.

Здравствуйте, Евгений Вениаминович.

Ага.

Скажите, а вы в университете сейчас или?

Нет, я ж не хожу. Я ж дистанционно веду занятия.

Вы дистанционно?

И сейчас тоже, кстати, веду занятия.

Да? Я вам тогда попозже перезвоню.

А у меня позже тоже, у меня семь пар подряд.

А, даже так?

Да.

Да я хотел узнать там… Я ж так в электронный журнал и не подал тогда. Есть ли возможность?

Есть, только в декабрьский уже. Потому что ноябрьский утвержден. Ну присылай, присылай, да и все.

Мои действия какие тогда?

Присылай, присылай статью.

Статью, да? По правилам?

Ну, конечно.

Я по правилам журнала, куда?

А там в контактах моя почта есть.

Непосредственно вам, да?

Да, да.

Ага. А там дальше вы как по…

А дальше, а дальше я все там буду уже делать, все как принято у нас, да.

Там же еще рецензию надо?

Нет, не надо. Это когда внешняя.

Не надо, да?

И когда внешняя тоже не надо, вообще не надо.

Всё, я понял. Хорошо.

В общем, присылай, если что-то не так, я тебе скажу тогда.

Угу.

Ну, исправим, исправим.

Я понял. Спасибо.

Давай, ага.

4. Разбор работы студента Марка (Проблема с Excel/CSV)

Так, ребят, значит, кто-нибудь из вас присылал мне письмо, на которое я ответил, что это на занятиях будем смотреть?

Да, мне присылали.

Ну вот. Вот как раз вот это занятие и наступило.

Я вам просто сейчас еще раз до занятия отправил, чтобы было легче найти.

Да. Ну хорошо, сейчас посмотрю.

Марк, да? Вот. Да. Угу. Ну что ж, давай. Ты даже сам в Excel преобразовал, да?

Просто мне друг сказал то, что лучше в двух файлах сразу скидывать.

Ну, да. Ну бывает, иногда не очень хорошо преобразуется. Сейчас посмотрим.

4.1. Проблема с файлом Excel

Ссылка открывается некорректно. Вот, вот источник данных.

Так, значит, сейчас я должен вам экран открыть. Совместный доступ. Я сейчас вел занятия, ну не буду, ладно, говорить про это. Так. И смотрим.

Значит, исходные данные помещаем в папочку indata внутри. Сейчас я путь на нее пришлю вам. Так. Как вы преобразование в Excel осуществляли? Да?

4.2. Работа с CSV и конвертером

Вот смотри, Марк. Видишь, вот смотри, что тут получилось. У тебя вот получились числа, ZN – это числа, да? А потом написано 12 мая. Потом написано RM числа.

Да, кстати, вот проблема в этом, то что когда открываете Excel на компьютере, тогда вот так отображается. А если в таблицах в Google, тогда они по-другому отображаются.

Вот. Значит, нам такой файл не подходит экселевский.

4.3. Очистка данных (замена разделителей, обработка пропусков)

Значит, что мы делаем? Смотри. Мы берем его, удаляем, он нам не нужен. Вот. А здесь у нас числа вот эти, да, все? Угу. Значит, давай теперь вот что сделаем. Иногда бывают некорректные экселевские файлы. Ну они-то корректны, это ж формат такой широко распространенный. И есть там, как говорится, разные варианты. Вот. Но дело в том, что некоторые варианты очень сложные. Ну, например, там есть числа с плавающей запятой. Формально это не числа, то есть там буковка Е есть, понимаешь? Это не не цифровой символ. Такой условный. Вот. Ну плюс-минус то еще ладно. А когда буковка Е, ну если есть буква Е, тогда надо вообще по правилам преобразования чисел с плавающей запятой это преобразовывать число. Ну можно, конечно, сделать конвертер, который будет это делать, но это надо делать же сидеть.

Значит, есть в системе есть конвертер. CSV DBF конвертер. Опа!

Какой это режим был? 2 2 3.0?

2 3 2 10. Ну дело в том, что 2 3 2 10 он сейчас дал ошибку мне. Хм. Ну, ладно, давай тогда 2 3 2 2.

Значит, здесь тоже есть CSV конвертер, но только он не знает, сколько там колонок в CSV файле. А почему здесь такая получилась ситуация? Ну-ка. Это подумаю. Так. Group by file input data CSV. А что там его разве нету? Есть. Там у меня файл CSV просто. А? Ну вот просил. Вот.

Сейчас выделил первую строчку блоком и посмотрим ее в вордовском файле эту строчку. Что я делаю? Вот. Вставляем эту строчку. Ага. Значит, здесь проблема, знаешь в чем? Смотри. У тебя эти поля отделены друг от друга не запятой, а точкой запятой. Ну это необычное дело. Сейчас посмотрим. Что можно сделать? Знак абзаца заменить все. Вот. И посмотрим, что у нас значит, если перевести на русский. Ничего не поймешь, что это значит. Ну давай сейчас посмотрим, что там у тебя в этом описании.

В результате, значит, потребление алкоголя студентами. А ты не единственный, кто это пишет, а? На эту тему.

Я другую ссылку скидываю. Я смотрел у друга просто, как тема называется. У меня с этими… с предсказанием цен на недвижимость. Сейчас, давайте я сейчас найду и вам скину.

Ну это ж часто потребление алкоголя у меня ссылка открылась. Ты ж по твоему письму я кликнул на ссылку.

Я, наверное, просто скопировал…

Подожди, подожди, сейчас, сейчас другую, сейчас другую. Нет доступа. Так, а здесь почему нет доступа? Ну скидывай ссылочку тогда. Угу.

Набор данных о недвижимости. Здравствуйте, здравствуйте. То есть у нас классификационная шкала будет средняя стоимость частных домов в тысячах долларов. Тысячах долларов.

Так, смотрим. У нас здесь действительно точки запятой все. Ну я предлагаю, знаете что сделать? Взять вот так вот его выделить блоком и поменять точки запятой на запятые. Да, я не вижу тут запятых внутри, поэтому это как бы внутри значения, там запятых нет, там просто вот точка запятой. И все. И на этом, собственно говоря, должны все проблемы сняться с преобразованием. А остальное, вот эти точки, которые там в самих этих…

Они не будут играть значения?

Нет, они они будут преобразованы в запятые самим конвертером. А, ну естественно. Потому что я же его перекомпилировал. Поэтому он заменен или поврежден. Ну ладно. Ну сейчас мы сделаем. Поставим тот, который был. Я же пытаюсь это все совершенствовать по мере возможности. Можно просто его запустить. Потом я нужный файл… Input data CSV. А там что, нету что ли его? Ну ладно. Сейчас мы его этот файл скопируем сюда. Заново систему всю развернем. То есть там тот конвертер, который в самой инсталляции. Так, input data CSV. А почему она думает, что его там нету? Может, его там нету? Не докопировал.

Когда такая ситуация возникает, непонятная, что с ним делать, тогда давайте я вам покажу онлайн-конвертер хороший, который, в общем, позволяет решать такие проблемы. Вот. Значит, мы запускаем тогда вот так делаем: CSV. У меня с этими файлами бывают проблемы, ребята, бывают. И вот онлайн вариант этот вот. Есть Convertio.co. Вот этот конвертер. Значит, я его сейчас вот переделываю систему и в описание, в чат, в Help я включаю информацию об этих делах вот. О том, что могут быть такие вопросы, там объясняю это все. Вот. И смотрим теперь этот файлик. Конвертировать. Это хороший конвертер, но у него одно ограничение есть: не больше 100 мегабайт. То есть он больше 100 мегабайт файлы не конвертирует. Но это такое, не очень не очень жесткое ограничение, которое как бы ничего так. И вот здесь уже по-нормальному. Видите, здесь уже дат нету, ребята. Видите? Вот видно, да, что дат уже нет. Этот файл нормальный в этом смысле. Но есть пробелы. Видите, вот здесь кое-где есть пробелы. Это классификационная шкала.

Теперь, в системе обязательно нужно колоночку вот здесь в начале вставлять. То есть не шкалу, а просто колоночку, которая будет номер строки означать или… Вообще она предназначена для того, чтобы информацию о строке содержать. Содержит информацию о строке.

Теперь смотрите, ребята, вот эти вот пробелы, их нужно заменить на нули. Это можно сделать вот так вот. Прямо берем, выделяем, Ctrl+H, пустое место на ноль. И заменяем. Пять таких случаев было во всей таблице. Почему? Потому что иначе это будет рассматриваться как текстовая колоночка. А так, в остальном здесь уже все нормально. Смотрите, видите, здесь уже запятые стоят. То есть это числовые колонки. И у нас, когда вводятся данные в систему, то там есть место, где задается, как за что, как интерпретировать нули и пробелы. Как значащие, как информацию реально измеренную об объекте наблюдения, или это у нас отсутствие данных. И вот в данном случае это будет отсутствие данных.

4.4. Настройка параметров модели

Теперь смотрим, какие колоночки считаем. Со второй по тринадцатую. Здесь у нас формула. Это у нас описательные шкалы, а пятнадцатая – классификационная шкала. Пятнадцатая, пятнадцатая. А со второй по четырнадцатую – описательные шкалы. Всё.

Ну теперь, ребята, я еще хочу вам одну вещь сказать. Обратите внимание, здесь есть числа целые, есть с одним знаком после запятой, а есть с двумя знаками после запятой, в одной и той же колоночке, видите? Значит, если где-нибудь будете публиковать статьи или какую-то работу научную проводить, то я вам убедительно рекомендую вам, чтобы в колоночке было одинаковое число знаков. Вот, допустим, здесь вот у нас сколько? Пять знаков. Значит, везде должно быть пять знаков. Как это сделать? Берем формат ячейки, числовой, пять знаков. Везде, во всей колоночке. Вот здесь вот у нас есть один знак, да, вот число с одним знаком после запятой. Значит, все числа должны быть с одним знаком после запятой. И нули, и другие числа. Вот. Здесь вот у нас с тремя знаками. Значит, должны быть все числа с тремя знаками после запятой. Вот я вам очень советую это придерживаться этого правила. Причина какая? Смотрите. Число знаков после запятой – это точность измерения величины. И вот если у вас где-то один знак, где-то два знака, а где-то ни одного знака, да? То что это значит? Что у вас точность измерения величин до целых. То есть вот здесь вот стоимость, видишь, написано 24, например, да? Мы знаем, что это в тысячах долларов, да? Ну значит, ровно 24.000 долларов. А вот это 21,6 тысяч долларов. Тут возникает вопрос: только написано 24? Или просто неизвестно, какая точность до десятых? Или известно, но там действительно их нету, понимаете? То есть вот этот вопрос возникает, и он снимается тем, что вы делаете везде знак после запятой. Вот 24.0 – это означает, что там именно ноль, понимаете? Что там именно вот 24.000. А если там ничего нет, то там неизвестно, что, понимаете? Вот смысл какой. То есть тогда там неизвестно просто. Ну так вот заменяем. Вообще по-хорошему надо во всех колоночках вот так заменить. Ну сейчас для расчетов это не принципиально, но просто нужно это знать, что это правило хорошего тона. Вот здесь у нас есть с двумя знаками после запятой. Значит, надо с двумя и делать. Вот. Вот сейчас мы перевели в порядок табличку. А вот здесь вот у нас с одним знаком, и с двумя тоже есть. Значит, давайте с двумя везде укажем. Вот сейчас мы сделали все правильно, как положено. И еще вот здесь вот тоже. А здесь три знака. Вот теперь совсем другое дело. Теперь нормальная табличка.

5. Разбор работы студента Шамиля (Ошибка конвертации)

5.1. Проблема с файлом и режимом конвертации

Вот здесь вот у нас мы этого не делали. А вот здесь вот у нас…

А можно еще насчет ResearchGate сказать? У нас же так и нет корпоративной почты.

Ну, значит, придется мне размещать. Вот. А мне это…

Да, хорошо. Просто я позвонил куратору, ну, написал ей. Она сказала то, что все-все хорошо будет, но так до сих пор еще ее нет.

Ну, надо напоминать. Потому что это вообще, если так между нами, это непорядок, понимаете, так мягко сказать.

Ну здесь берите пять интервалов, а получится получше эти вот модели, получше получаются. Видите, да? Видно, разница больше между истинными и ложными. Вот остальное все так же берете.

Значит, и модель какую брать надо? Наверное… Тут разница не очень большая между INF3 733, здесь 739, 878, вот здесь 879. В общем, берете модель INF3 все-таки, потому что она получается обычно получше в выходных формах. Все четко совершенно, видите, да? То есть истинных решений значительно больше при всех уровнях сходства, чем ложных. Но начиная с 60% вообще нет ложных решений. Это хороший результат, убедительный. То есть хорошо выглядят все эти формы. Вот. То есть, Марк, короче, ты берешь вот эту вот э-э все, кроме вот здесь вот указываешь не 10 вот здесь вот, а пять указываешь. Пять, пять. И нажимаешь кнопочку "Пересчитать шкалы и градации", она тебе пересчитывает это все. И потом дальше можешь уже все делать.

Понятно?

Алло.

Тебе понятно? Вот здесь вот в этой форме, видишь, здесь R10. Задаешь там пять вот здесь. И нажимаешь "Пересчитать шкалы и градации". Она тебе пересчитывает это все. И потом дальше можешь уже все делать.

Угу, понял.

Вот. У тебя тема утверждена.

Вот. Ну и вопрос возникает, а кто еще хочет посмотреть, чтобы мы утвердили его тему, посмотрели на его тему? Только один Марк присылал письмо мне? Больше никто не присылал из вас, из присутствующих? По-моему, Шамиль присылал.

Я вам писал за ошибку, что у меня выбивалось.

А, ты чем? 3 3 1 3 1?

Ну я тебе помог или нет?

Нет, вы сказали разберемся на занятии.

А, и не разобрался. Ну так давай тогда на занятиях и разберемся.

У вас же есть мой файл? Вы мне его скидывали.

Я могу скинуть.

Ну, ну есть, у меня надо искать. Ну сейчас я посмотрю. Может…

Скинуть вам?

Может, тоже написали, что на занятиях отправлю все. Давайте, отправляйте мне. Да, лучше отправляйте еще раз. Никита прислал. Вот. А Иван ничего не прислал, просто письмо прислал, да? Письмо без данных ты прислал.

5.2. Рекомендации по подготовке данных (уникальные текстовые названия, ингредиенты)

Так, смотрим. Никита. Скачиваем все это. Попробуем сейчас сделать. Вот хорошо преобразовался файлик, да? Так, на вид вроде нормально все. Никита, вы меня слышите, Никит?

Да, он написал.

У него микрофон выключен.

А. Мне лучше в чат не писать, потому что у меня… А у тебя микрофона нет? Классно. А сейчас кто сказал, что микрофона нет?

Я, я.

У тебя появился микрофон, что ли?

Нет, э-э, Шамиль.

А. Короче так, вот здесь вот первая колонка должно быть наименование. Вот я его пишу. Там любое: общ, там, номер, там, что-нибудь. Потом вертикально расположил и сжал по размеру, видите, все эти колоночки. Потом сделал такой вот вид у него, получше. Значит, э-э, первая колонка не классификационная шкала, не является шкалой вообще никакой, это просто номер строчки. Теперь давайте, говорите мне, какой смысл этой задачи и где тут классификационная шкала, что мы здесь хотим определить? Что мы хотим определить, решая эту задачу? Что там на что влияет? Какой смысл этой задачи? Имя артиста, страна или что это такое? А это что? Что это за задача? Так, смотрим. Рейтинг песен. Так, ну и где тут у нас рейтинг? Давай посмотрим. Популярность. Значит, ребята, вот эта популярность – это последняя колонка, это будет классификационная шкала. То есть мы вот эту шкалу последнюю рассматриваем как то, что нам нужно определить. Определяем мы на основе чего? Вот на основе всего этого. Имени там артиста, там и так далее, да? Значит, вторая колоночка и вот по по двенадцатую – это у нас получаются описательные шкалы. А тринадцатая – это у нас классификационная шкала, да? 13, 14, 15 – классификационная шкала.

А вы зайти в Excel? У вас вроде бы столбец State сдвинутый, да?

Сдвинуто вот колонка наименование. Первую строку она как наименование рассматривает.

Вот. Теперь смотри, значит, три градации числовых описательных шкал. Она две таких обнаружила. Ну и две есть вот эти числовые вот время там что-то там, да? Вот. А остальные текстовые, это вот просто вот эти слова. Вот. А классификационная у нас все слова, там числовых нету. Вот. Поэтому мы смотрим, что там у нас получилось. Значит, эти регионы, эти вот длинные. А здесь у нас, где числовые, там написано по 75 примеров на диапазон числовой. А эти все уникальные текстовые значения. И все. Значит, выходим на… И тебе я не помню, оставил это или нет. Выходим на создание модели. Вот с этими параметрами. На довольно быстро это сделала. Вот, 2 секунды. И потом мы смотрим, что же у нас получилось. Вот регионы она как сформулировала. Вот это регион, где нет где нет информации о том, какой регион. Там, кстати, можно пошутить, написать "неизвестный регион", например. Ну, можно ничего такого не делать. Вот. Теперь ингредиенты. Она не не разделила их на ингредиенты. Это неправильно. Значит, э-э, чтобы она их разделила на ингредиенты, нужно использовать другую другие параметры входные. Вот здесь нужно поставить равные интервалы, иначе она не может их разделить, слова анализировать. Вот это вот надо сюда вообще вставить даже информацию об этом. 2 3 2 2. Если… Вот так вот, когда все это рассказываю, показываю, очень много всяких таких нюансиков замечаю, которые я не могу исправить по времени просто. То есть это такое дело, что там много всего. Вот. И потом выходим. Сейчас она эти ингредиенты разделит. Вот, видишь, здесь нету уже столько примеров, там вот так получится у нас. Вот. И отправляю тебе на почту. Название я присылаю тебе на почту. Вот. И дальше мы смотрим достоверность. Вот, получилось неплохая достоверность, даже хорошая. Вот это вот 0,958 – это вообще замечательная достоверность по критерию F-критерию Ван Ресбергена. И не менее высокая достоверность и по L1-критерию, и по L2 тоже очень высокая. Ну даже, я вам скажу так, что даже нет необходимости там что-то еще выдумывать. Когда настолько достоверная модель, то там как бы все ясно. Там есть прямая зависимость вот этих вот рейтинга от этих показателей. Только ну тут какая прямая зависимость? Это вот эти уникальные текстовые названия. Ну, даже говорить не о чем. Видите, у вас тут ложных решений очень мало, и они при очень низких уровнях сходства. А дальше там все абсолютно, так сказать, спокойно идентифицируется. Вообще даже и как бы вариантов нет. Нет сомнений. Что у вас все получилось, Никита. Так что вы можете эту задачу описывать дальше. А шаблон описания вы знаете, где взять, да? Знаете? Где взять шаблон описания? Вот тут вот рейтинг есть авторов по числу статей, по числу обращений к статьям. И вот любую статью открываете вот, первую попавшуюся. И вот, пожалуйста, ребят, смотрите, вот это вот берете вот, вытаскиваете. Только надо вытаскивать не из PDF-файла, а из вордовского. Вот я показываю, смотрите, вы сейчас видите, что я показываю, ребят? Вордовский файл показываю, вот, DOC.zip. Он скачивается этот файл. Вот так вот берете, э-э, и в шаблоне описания тоже типа такого же есть файла. Вот, там Янина сделала ссылки на вордовские файлы вот здесь. Это ссылки на вордовские файлы. Вот это, вот это вот. Вот здесь вот есть ссылка на PDF-файл, там, ну там есть и вордовский файл. Вот. Что-то так она как-то не открывается. А, вот. Вот смотрите, ребят, видите? Вот статья новая, только что вышедшая. А вот строчечка. Вот так блоком выделяете, вставляете в свою статью. А вообще-то в шаблоне там там моя фамилия, скорее всего, уже есть, то есть там вообще ничего делать не надо. Свою добавляете просто. Как вы считаете, там хотите спереди добавлять, хотите сзади добавлять после меня. Вот, это ваше проблема решать эти вопросы. Вот. Так что эта модель утверждена. Эта модель у нас Никиты. Я послал ее. Она вообще получилась замечательная. Видишь, здесь совершенно четко видно, что истинные все решения положительные, все практически истинные. Чуть-чуть есть там с очень низким уровнем сходства несколько ложных положительных решений, а так все истинные. Ссылочки я послал, а сами сами файлы сейчас пошлю еще. Что-то я их не прикрепил. Вот. Всё. Делаешь вот такую модель, все это полностью, как в статье описываешь, будет прекрасная у тебя работа. Шамиль, Шамиль. Шамиль. Так. А тут что у нас такое? Напомни мне, Шамиль, что здесь, о чем речь? Данные о 50 популярных блюдах в Индии. А. Здесь что у нас там, что-то не получалось, да? Нет, там все нормально, просто… Я так понимаю, там все уже нормально, потому что вы не заходили в тот режим, в котором у меня ошибка. Ну я не сказал тебе, что утверждена тема, нет? Просто сказал, что рассмотрим на занятии, да? Нет, вы мне ее и дали этот файл. Да? Да. Значит, смотри, он все равно, здесь все равно нужно кое-что сделать. Вот здесь нужно первую строчечку внести. Зачем? Ну, если мы хотим, допустим, чтобы у нас э-э наименование… Что здесь такое? Кошмар какой-то. Вот. Ну здесь вот у нас э-э наименование блюда, наверное. А здесь из чего оно состоит, да? Вот. А здесь какое-то вегетарианское, там диетическое. Какие блюда, в каком регионе, с каких ингредиентов делают? Вот так что-то типа этого, да? Значит, если мы хотим, чтобы название этого блюда было тоже э-э признаком блюда, признаком региона, то тогда надо здесь добавить колоночку, и тогда это будет у нас э-э тоже э-э информация о том, какой регион, короче. И здесь вот можно по словам анализировать. Какие слова? Где ж, видишь, они молоко там, что-то молоко, сахар там, вода, да? Роза. Вот. То есть вот это все она может вытаскивать и отдельные слова использовать как признаки того, что в каком регионе это принято, ну, как более более представлено, скажем так. И по названиям тоже так же она может слова вытаскивать. И, допустим, вот здесь названия-то разные, но первое слово одинаковое, например. Вот она может это слово считать признаком этих вот регионов, например, запад там, восток. А второе слово тоже будет где-то использоваться. То есть можно две сделать шкалы, которые будут слова анализировать. Вот. Ну поэтому то есть сами названия блюд тоже она будет анализировать. А можно оставить их просто названием строчки вот здесь вот. А вот это начиная с этой колонки признаки уже рассматривать. То есть по ингредиентам определять, какое будет, какой регион. Ну давай так вот и сделаем. Там со второй по восьмую. Нет, вот мы берем со второй, значит так. Вот. Вторая по восьмую – это у нас получится описательные шкалы. А девятая – это будет у нас одна классификационная шкала. А здесь получилось, смотри, в некоторых, у некоторых блюд неизвестно, какой регион. Она по этим характеристикам попытается все равно определить, какой регион для них. Здесь, да, потом смотри, она будет анализировать слова как признаки, вот будет рассматривать. А и чтобы она не разбивала вот эти слова, штат и регион, не разбивала на на слова, для этого мы взяли, заменили пробелы на нижнюю черточку здесь. Угу. Тогда она будет целиком вот так вот рассматривать там Уттар-Прадеш, например. Вот все, она целиком его будет брать. Ну и все. И теперь просто надо его вводить в систему и смотреть, что получится. По-моему, мы это что, я не помню, делали мы или нет? По-моему, делали даже. Делали, да. И что, и что-то сейчас, секундочку. Получается около 0,700. Сейчас посмотрим. 2 8. Вот здесь вот такая особенность, что если мы хотим слова рассматривать как признаки, у вас там… Нет, в классификационных шкалах. А в описательных, по описательным шкалам. По описательным шкалам, это не классификационная, там у нас целиком значение поля. А в описательных шкалах мы хотим слова как признаки рассматривать, причем любые, даже коротенькие. Вот это я сейчас тебе посылаю, это вот такая вот настройка этого режима, чтобы он правильно все анализировал у нас, чтобы он выявлял правильно эти слова как признаки. Теперь давайте посмотрим, что у нас получится. Заняло это 17 секунд. Всё. И смотрим на это. Лучшие 50 песен. АСК-анализ индийской кухни. В режиме 1.3 заходим и меняем название. Смотрим достоверность. Вот, получилось очень неплохая достоверность. 0,828 по F-критерию Ван Ресбергена – это очень высокий результат. Даже 925, 946 такие вот. Это высокие результаты, понимаете? Вот. То есть получились хорошие модели у тебя. Сейчас мы эту форму тебе пришлю. Вот. Ну, в общем, принимается тема у тебя. Ну, на модели используй вот эту вот INF3 сумма знаний. Вот эта модель. Она хоть и не совсем идеальная, но, в принципе, получилось тоже. Видишь, у тебя тоже, начиная с какого-то уровня, начиная с 60%, больше доля истинных решений. Ну, то есть есть зависимость года от того, какой язык программирования наиболее популярный. Вот. И что мы получаем в результате? Значит, вот смотрим. Сейчас я пошлю тебе это. Вот, смотри, мы заходим, допустим, 4 4 8 и смотрим, что у нас для какого-то года какой-то язык программирования является наиболее популярным. Вот тут, допустим, Ада, Visual Basic там и так далее. А, причем это максимальное значение, это минимальное. А для 2020 года Котлин, Свифт. То есть видно какие, причем это, по-моему, не соответствует. Джулия. Я я не знаю таких даже языков, если честно, самих названий даже не слышал. А теперь, ну некоторые знаю, конечно. Значит, вот смотрим теперь вот, допустим, язык э-э Visual Basic. Для каких, для каких лет он наиболее характерен? Вот такую вот информацию ты можешь извлечь. Смотри, он характерен для пятого, шестого, девятого, седьмого, четвертого, десятого, восьмого. То есть Visual Basic был популярен э-э в какие-то годах. Вот. Эти года. А потом вот он стал непопулярен в 2018, 19. Сравнительно. И можно все это увидеть, короче, наглядно. Вот. Короче, можно все это э-э писать статью, делать модель и смотреть. Вот популярность языка Авар, например. Она была э-э, значит, э-э высокая популярность была в 2005, в 2008 году. Средняя была, значит, э-э в восемнадцатом году, а потом понизилась. И вот так вот мы можем взять и все все языки посмотреть. Вот это вот 3.3 – это наиболее э-э популярно. Это средняя популярность, это низкая популярность. То есть низкая популярность была раньше. А вот Ада, у него все время растет популярность, видите? Восемь э-э постоянно. А у C++ падает популярность. Вот. А C# возрастает. Короче, все там у тебя получилось нормально. Можешь спокойно делать описание, все эти вот формы приведешь, сделаешь выводы. Вот. И, значит, если будут вопросы, спрашивай. Дельфи падает популярность, закономерно. Ну как падает? У него э-э низкая популярность была в 2020 году, а высокая популярность в 2004 году. То есть популярность уменьшается. Ну все, ребят, значит, на этом заканчиваем занятие. Вот. Всего самого-самого хорошего. До свидания. До свидания. Остальные готовьте данные, на следующих занятиях тоже будем аналогичные смотреть, делать выводы. Конец записи. До свидания.

6. Общие рекомендации и публикация

Ребята, значит, теперь смотрите, э-э, если вы не можете зарегистрироваться, и к тому времени в ResearchGate не сможете зарегистрироваться, ну тогда придется меня писать туда соавтором. Вот. А еще есть один вариант, ребята, еще один есть вариант. Я придумал такой вариант. Значит, если кто-то из вас сможет зарегистрироваться, слышите, нет? Вот староста, к примеру. У старосты ж есть, есть же почта корпоративная? Староста зарегистрируется. Кто у нас староста? У нас староста 2003 группе Евгений Волков, да? Правильно? Да, правильно. Правильно. Значит, Волков, он просто сам Бог велел зарегистрироваться, потому что у него есть корпоративная почта. Всем старостам выдали. Вот. И он… Можете брать его в соавторы, а не меня. Сразу у него появятся баса статей. Он такой станет продвинутый сразу же, моментально. Вот. Я думаю, что это неплохой вариант, такой нормальный. Ну, ваше дело, короче. Но можете спокойно его взять. Или еще кого-нибудь, у кого там зарегистрирован. Я к этому отношусь абсолютно спокойно. Я вам даже советую. Понятно, да? Почему я спокойно отношусь? Я вам сказал, потому что я вас учу, даю вам все. А вы с этим делаете то, что считаете нужным. И я вот могу быть соавтором, ну, в случае необходимости, если у вас не получается там уже разместить, ну тогда я, чтобы разместить, мне нужно быть соавтором. А так я абсолютно на это не претендую. И вы можете взять кого угодно, кто там зарегистрирован, видите его в соавторы, и он размещает. Даже можно рядышком вместе сесть и разместить там. Если будут вопросы какие-то, вы меня спрашиваете. Я вам могу сказать, что когда начинаешь размещение, ну я, наверное, покажу вам. Вот когда к этому дело подойдет, что кто-нибудь один хотя бы сделает описание, мы сразу его разместим, сразу же. Я вам покажу, как там делается. Но там все понятно, в принципе. Указываешь PDF-файл, э-э, указываешь, что кто авторы. Тебе пишет. Ну надо все писать э-э по-английски, ребята. Это можно делать с помощью переводчика вполне. Вот Яндекс-переводчик считается наилучшим. Я спрашивал людей, которые свободно владеют английским, какой лучше переводчик, Google или Яндекс? Все говорят, что Яндекс лучше. Вот. Ну когда касается перевода вот с русского языка на английский, по крайней мере. Вот. И, значит, потом указывается авторы, потом указываете э-э лицензию. Там есть список лицензий. Берете самую первую из списка лицензий, самая такая простецкая. То есть она меньше всего ограничений накладывает. И потом там есть место, ребят, слушайте дальше внимательно, где спрашивается: присваивать DOI или нет? Digital Object Identification. Цифровой идентификатор объекта международный. Обязательно это место не пропускайте, ставьте там, что да, присваивать. Причем, если вы там как-то у вас не получится, то не не расстраивайтесь. Значит, потом можно, когда уже статья размещена в ResearchGate, можно войти в нее и потом там это сделать эту операцию выполнить. То есть он такой, знаете, этот ResearchGate, он прощает э-э недочеты. То есть если там что-то не заметил там, что-то не то нажал, то это все не проблема, понимаете, там. Вот я беру, допустим, любую публикацию, захожу. Вот публикация. Вот захожу, допустим, э-э управление знаниями. Вот здесь DOI, видите, есть? Да? Видно, да? А если бы не было, а тогда вот сюда открываешь, а здесь написано: присвоить, добавить DOI, вот здесь прямо будет написано. Ну здесь его сейчас нет, потому что оно есть это DOI присвоено. А если его нет, тогда прямо написано. Я нажимаю "Присвоить DOI", появляется, значит, она раз, тенк, присваивает вот сюда вот DOI. А этот DOI, он необходим для того, чтобы чтобы в РИНЦ разместить. Вот. Если в РИНЦ, значит, если его нет, то РИНЦ не не берет. А в ResearchGate присваивается DOI только препринтам, препринтам и еще есть где-то тут научные отчеты. Или технические отчеты. Видите? Нет, не это. Сейчас найду. Препринт, конференция, доклад, данные, книги, другие. Проекты. Нет. Я размещал. Данные, тоже не данные. Вот, ребят. Вот когда технический отчет размещаешь, первый пункт, тоже присваивается DOI. А остальным нет. Ну технический отчет – это, считай, научный отчет. То есть либо вот этот выбираете первый пункт, либо второй. Ну я всегда выбираю второй, когда вот типа статьи что-то. А когда там уже такое что-то приличное, побольше, по объему, тогда как отчет я делаю там, ну если листов 80-90, тогда вот уже больше похоже не на статью, а на отчет. Там книги, видите, здесь статьи, книги есть вот. Вот так вот.

7. Заключение

(Завершение занятия, прощание)