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

***И.Т. Трубилина»,***

***Российская Федерация***

***Колесников Роман Юрьевич, ПИ2102***

***roman563412@gmail.com***

**АСК-анализ текстов (локальная лаб.работа 3.02) и спектральный АСК-анализ изображений (интеллектуальное облачное Эйдос-приложение №277).**

**Заголовок**

Автоматизированный системно-когнитивный анализ текстов в Eidos (Лаб. 3.02) и обзор возможностей системы

**Резюме**

Лектор начинает занятие, проверяет запись и приветствует студентов. Планируется рассмотреть две лабораторные работы: обработку текстов (Лаб. 3.02) и обработку графики. Лектор напоминает, что ранее была подробно изучена Лаб. 3.03, на примере которой объяснялись базовые понятия системы Eidos.

Лабораторная работа 3.02: Обработка текстов

Цель: Демонстрация возможностей Eidos по обработке текстов, представленных в табличном виде (Excel).

Запуск и настройка Eidos:

Запуск системы Eidos (StartEidos.exe).

Проверка обновления системы до последней версии.

Удаление предыдущих приложений и добавление Лаб. 3.02 через режим 1.3.

Интерфейс и параметры:

Используется универсальный программный интерфейс (API 2.3.2.2) для ввода табличных данных (текстовых и числовых) и сценарного анализа.

Настройка параметров обработки текста:

Выбор полей для классов и признаков.

Опции интерпретации текстовых полей: целиком, по словам (с фильтром по длине), по символам.

Возможность включения/отключения лемматизации (приведения слов к начальной форме).

Для Лаб. 3.02 выбирается обработка по словам длиннее 3 символов, с лемматизацией для описательных признаков (текстов) и обработка поля целиком для классификационных признаков (авторов).

Создание модели:

Система автоматически считывает исходные данные из Excel-файла (фрагменты текстов о писателях, сгруппированные по авторам).

Происходит автоматическая формализация предметной области: создаются классификационные и описательные шкалы и градации.

Исходные данные кодируются с помощью созданных шкал, формируется обучающая выборка.

Анализ исходных данных:

Демонстрируется структура Excel-файла: фрагменты текстов (абзацы из школьных сочинений), нумерация фрагментов, указание автора (классификационная шкала).

Синтез и верификация моделей:

Запускается процесс создания статистических и системно-когнитивных моделей (режим 3.5).

Обсуждаются этапы автоматизированного системно-когнитивного анализа (АСК-анализа): когнитивное структурирование (постановка задачи), формализация (создание шкал и обучающей выборки), синтез и верификация моделей, решение задач (идентификация, прогнозирование и др.).

Подчеркивается важность оценки достоверности моделей с использованием различных критериев (F-мера Ван Ризбергена, ее нечеткие обобщения, инвариантные меры).

Просмотр результатов:

Демонстрируются матрицы моделей (абсолютных частот, условных/безусловных вероятностей, информативности, Х-квадрат).

Показываются нелокальные нейроны (семантические ядра классов/авторов) и дендрограммы для визуализации сходства/различия классов (авторов) и признаков (слов).

Объясняется понятие семантического потенциала слова (количество информации о принадлежности текста классу).

Решение задачи идентификации (атрибуции):

Демонстрируется, как система определяет автора текстового фрагмента на основе созданных моделей, сравнивая профиль фрагмента с обобщенными образами авторов.

Показывается высокая точность идентификации для данной задачи.

Обзор возможностей Eidos и дальнейшие планы

Система Eidos позволяет проводить глубокий интеллектуальный анализ текстов на любом языке (с алфавитом, не иероглифическим), включая создание обобщенных лингвистических образов, сравнение текстов и авторов, атрибуцию, анализ контента сайтов.

Упоминается использование словаря лемматизации академика Зализняка.

Кратко анонсируется следующая тема – обработка графики (по контурам, пикселям, спектрам).

Конечная цель курса – научиться разрабатывать собственные интеллектуальные облачные Eidos-приложения.

Объявляется перерыв.

Детальная расшифровка текста

Раздел 1: Введение и план занятия

Подраздел 1.1: Приветствие и проверка записи

[0m2s] Еще раз здороваемся.

[0m5s] Здравствуйте, ребята.

[0m7s] Что-то она не начинается запись.

[0m9s] Добрый вечер.

[0m10s] Добрый вечер. Но запись не идет.

[0m14s] Запись уже идет.

[0m15s] Идет 5 минут.

[0m17s] Да вы что?

[0m18s] Когда это она включилась?

[0m21s] Не видел.

[0m22s] Только что включил.

[0m23s] Так я включил, но она не шла.

[0m25s] Было это неактивно.

[0m29s] Ладно, вот теперь я вижу, что идет.

Подраздел 1.2: План занятия и обзор предыдущей работы

[0m31s] Ну, ребята, у нас, здравствуйте, во-первых.

[0m35s] Да, здравствуйте.

[0m36s] У нас сегодня по плану две лабораторные работы.

[0m40s] Мы подробнейшим образом изучили работу 3.03, на примере которой я объяснял все базовые понятия.

[0m47s] А сегодня у нас две работы. Одна для демонстрирует возможности обработки текстов.

[0m55s] ну я покажу вам, как это делается в текстах с текстами, которые в табличном

[1m1s] представлении, в таблицах.

[1m3s] И потом вторая работа - обработка графики. Ну я вам покажу только один из вариантов. В системе Eidos есть три варианта обработки графики: по контурам, по всем пикселям и по спектрам.

[1m15s] Мы посмотрим только по спектрам. Ну и сошлюсь на свои работы, покажу вам, где работы есть,

[1m21s] которые другие варианты обработки графики обеспечивают.

Подраздел 1.3: Дальнейшие планы и цели курса

[1m24s] После этого мы начнем разрабатывать собственное облачное Eidos приложение.

[1m29s] А сейчас вот у нас заканчивается, по сути дела, сегодня рассмотрение базовых возможностей системы Eidos по обработке

[1m39s] числовых и текстовых табличных данных

[1m43s] и по обработке графических данных. И после этого уже мы, по сути, будем готовы к тому, чтобы свое приложение разрабатывать

[1m52s] с моей помощью методической.

Раздел 2: Лабораторная работа 3.02 - Обработка текстов

Подраздел 2.1: Запуск и настройка системы Eidos

[1m55s] Значит, открываем сейчас, у кого есть компьютеры, желательно, чтобы повторяли за мной

[2m1s] то, что я буду делать.

[2m5s] Запускаем

[2m7s] папочки системы Eidos.

[2m9s] Запускаем Start Eidos

[2m13s] программу.

[2m17s] Система должна обновиться у вас

[2m19s] до версии двадцать первого десятого.

[2m23s] То есть вчерашняя версия.

[2m26s] Я буду отправлять в чат

[2m29s] скриншоты.

[2m32s] Стираем все приложения.

[2m35s] Удаляем.

[2m37s] И заходим в режим 1.3

[2m40s] и добавляем лабораторную работу 3.02.

[2m44s] Это работа, демонстрирующая возможности обработки текстов,

[2m49s] представленных в таблицах экселевских.

[2m55s] Устанавливаем эту лабораторную работу.

[2m59s] О'кей нажимаем.

Подраздел 2.2: Интерфейс API 2.3.2.2 и параметры обработки текста

[3m1s] Появляется вот такой вот такая экранная форма.

[3m5s] Эта экранная форма автоматизированного программного интерфейса 2.3.2.2,

[3m11s] который я называю универсальным программным интерфейсом, он очень

[3m15s] многофункциональный.

[3m17s] Обеспечивает ввод табличных данных и текстовых, и числовых, а также обеспечивает сценарный

[3m26s] автоматизированный системно-когнитивный анализ.

[3m30s] Мы эту форму уже видели

[3m33s] в работе 3.03.

[3m35s] Но там было параметры были были упрощёнными.

[3m40s] Обратите внимание, вот здесь мы сейчас, видите, птички стоят. Применить специальную интерпретацию текстовых полей

[3m46s] классов.

[3m48s] А здесь специальная интерпретация текстовых полей признаков.

[3m52s] И что мы здесь видим? Что можно значение текстовых полей целиком рассматривать,

[4m0s] а можно рассматривать признаки слова

[4m4s] и длиной больше тут заданного количества символов.

[4m8s] Ну я задал по больше трёх символов длину.

[4m12s] Вот. И можно как проводить лемматизацию, так и не проводить. Да, и как признаки можно рассматривать элементы символы, элементы значения полей символы.

[4m23s] Вот. Ну в данной работе, вот, которую сейчас мы будем рассматривать, мы будем слова рассматривать как признаки текстов.

[4m32s] Здесь у нас одна классификационная шкала, одна описательная шкала.

[4m37s] Но может быть и много и классификационных, и описательных шкал

[4m42s] при обработке текстов. Просто здесь очень упрощённый вариант я показываю.

[4m46s] Сейчас сошлюсь вам на список работ

[4m51s] по обработке текстов, и там есть развитые такие подробные примеры, которые показывают,

[4m56s] какие есть в этом плане возможности. И то, что я сейчас вам показываю, там есть.

[5m3s] И также там есть возможности

[5m6s] использовать текстовые файлы для обработки. Причём размер их может быть

[5m12s] каким угодно, в принципе, но в пределах физических ограничений компьютера, там, памяти,

[5m18s] языка программирования и тому подобное. Параметры, конечно, ограничивают, но книжки вполне можно обрабатывать.

[5m26s] Вот. Значит, здесь я сошлюсь на хелпы

[5m29s] к этой работе, чтобы полностью уже картина была,

[5m35s] где описаны

[5m36s] принципы, стандарты подготовки исходных данных в таблицах.

[5m44s] Ну когда сейчас мы эту работу установим, будем ещё этот вопрос рассматривать подробнее.

[5m52s] для классификационной шкалы мы рассматриваем специальную текстовую интерпретацию полей,

[6m0s] значение полей, но рассматриваем значение полей целиком.

[6m4s] Мы видим также, что можно сортировать

[6m7s] уникальные значения и не сортировать. Мы

[6m11s] выбрали

[6m13s] опцию выделять уникальные значения и сортировать. А можно не выделять уникальные значения и не сортировать.

[6m22s] Нажимаем о'кей.

Подраздел 2.3: Загрузка и преобразование данных

[6m25s] Скачиваются исходные данные

[6m28s] из папочки лабораторных работ.

[6m33s] И конвертируется во внутренний стандарт данных системы Eidos. Это

[6m38s] стандарт DBF

[6m42s] Dbase 4 NTX.

[6m45s] И видим, что изучен файл исходных данных системы и определено, что будет восемь колонок

[6m52s] в матрицах моделей и 43 строки.

[6m55s] Все шкалы текстовые. Ну, естественно, потому что мы обрабатываем как раз тексты сейчас.

Подраздел 2.4: Создание модели (формализация)

[7m1s] Выходим на создание модели.

[7m3s] И происходит, как я здесь иногда говорю, маленькое чудо.

[7m7s] Автоматически создаются классификационно-описательные шкалы и градации.

[7m13s] И потом исходные данные кодируются с их помощью, и получается обучающая выборка.

[7m18s] И таким образом исходные данные внешние

[7m23s] оказываются полностью подготовленными для

[7m26s] последующего

[7m29s] последующих этапов

[7m31s] автоматизированной системно-когнитивного анализа. То есть мы полностью вообще полностью прошли формализацию предметной области за минуты какие-то.

[7m41s] На этих на этом этапе формализации предметной области, я уже сказал, разрабатываются классификационно-описательные шкалы и градации

[7m49s] и с их помощью кодируются исходные данные, получается обучающая выборка.

[7m54s] Это небольшое окошко такое.

[7m57s] Когда студенты увидят, у них возникает какое-то неприятие к этому окошку, они сразу нажимают о'кей и не читают, что там написано.

[8m5s] А там между тем написано, что им делать дальше.

[8m8s] Но я знаю, что там написано, потому что сам это написал, поэтому сейчас мы это просто сделаем.

Подраздел 2.5: Просмотр исходных данных (Excel)

[8m14s] Посмотрим на сам файл исходных данных.

[8m17s] И потом немножко я дам информации по работам по обработке текстов, посмотрим шкалы, выборку и так далее. Создадим модели и будем дальше анализировать их.

[8m29s] Значит, сначала смотрим файл исходных данных.

[8m33s] Он находится по пути,

[8m36s] который был указан в этом окошке.

[8m39s] Значит, здесь мы видим, что в качестве файла исходных данных у нас используется экселевский файл,

[8m45s] в котором у нас есть фрагменты текста по различным писателям.

[8m52s] Эти фрагменты я пронумеровал по фамилии имени писателя, потом ещё 1 2 3 4 5.

[8m59s] И указал, что эти фрагменты относятся к одному и тому же писателю по классификационной шкале.

[9m5s] Откуда я взял эти фрагменты текста, ребят?

[9m9s] Я взял просто нашёл детские

[9m12s] школьные сочинения

[9m14s] про писателей известных.

[9m17s] И я сообщаю вам, ну это, наверное, вы знаете, но всё-таки я вам сообщаю, что если мы в Ворде возьмём, выделим блоком какой-то текст и потом вставим в Excel его через буфер обмена,

[9m31s] то есть скопируем и вставим, то у нас абзацы вордовского текста будут представлены в отдельных строках.

[9m38s] После этого остаётся просто написать здесь фамилию писателя, инициалы и протянуть её на все фрагменты текста, которые касаются этого писателя.

[9m51s] А в первой колонке, которая не является шкалой, они просто там содержится информация о том, откуда взята информация этой строки.

[10m0s] Мы тоже пишем так Достоевский-1, протягиваем вниз, получается автоматическая нумерация.

[10m8s] Вот. То есть я хочу сказать, что то, что я сейчас вам показываю, сделано очень технологично, просто и быстро.

[10m14s] Единственное, ребят, надо иметь в виду, что перед этой операцией нужно выключить переносы в Ворде, чтобы слова были без переносов.

[10m23s] Потому что эти переносы, они тоже попадут сюда.

[10m27s] Вот. Если же номера нет, просто написано Достоевский, то если мы просто протянем, так оно и остаётся, видите?

[10m34s] Поэтому это делается всё очень просто и быстро.

Подраздел 2.6: Этапы АСК-анализа (обзор)

[10m37s] Значит, теперь давайте расскажу в двух словах о первом этапе автоматизированного системно-когнитивного анализа

[10m43s] при обработке текстов. Это этап

[10m50s] когнитивно-целевой структуризации предметной области.

[10m54s] Это единственный не автоматизированный этап системно-когнитивного анализа. На этом этапе мы должны поставить

[11m6s] задачу,

[11m8s] которую мы будем решать.

[11m10s] И постановка задачи заключается в том, что мы должны выбрать, что мы будем рассматривать в качестве

[11m17s] факторов и их значений или в качестве значений свойств,

[11m21s] то есть признаков.

[11m23s] И что будем рассматривать в качестве объекта моделирования, на который действуют эти факторы или у которых есть эти свойства

[11m31s] с такими значениями. И что мы будем рассматривать в качестве обобщённых категорий, то есть классов, к которым относятся объекты, состояние объекта моделирования или в которые они перейдут.

[11m50s] Эта схема этапов автоматизированного системно-когнитивного анализа,

[11m57s] она есть в системе Eidos в режиме 6.4.

Подраздел 2.7: Установка Лаб. 3.02 и обзор этапов АСК-анализа (продолжение)

[12m1s] Как видим, лабораторная работа установлена, и можно её изучать. Будем изучать.

[12m6s] Но сначала я вам перечислю этапы

[12m10s] автоматизированного системно-когнитивного анализа, которые мы будем проходить сейчас.

[12m15s] Первый этап я вам рассказал, что мы в качестве, вернее, суть этого этапа рассказал. А что мы будем принимать в качестве свойств и значений и в качестве объекта моделирования, и в качестве

[12m30s] состояний объекта моделирования, это я ещё не говорил.

[12m34s] Значит, я вам сейчас скажу, что, значит, в качестве свойств и их значений будем рассматривать слова, в качестве объекта моделирования - тексты, принадлежащие, или вернее, не принадлежащие разным писателям, а о разных писателях, потому что это не их тексты, а тексты о них.

[12m53s] А в качестве состояний, к которым относятся объекты моделирования, будем рассматривать самих этих писателей или классы, соответствующие писателям точнее.

[13m6s] Теперь смотрим на следующий этап автоматизированного системно-когнитивного анализа. Следующий этап заключается в формализации предметной области.

[13m15s] Мы уже этот этап выполнили, это заняло несколько там минут, вот, даже секунд, может быть. Но суть этого этапа состоит в том, что мы разрабатываем автоматизированно, автоматически даже, или автоматизированно иногда, когда числовые шкалы есть, э классификационно-описательные шкалы и градации и кодируем с их помощью исходные данные.

[13m38s] Смотрим. У нас есть подсистема формализации предметной области, где мы смотрим, какие у нас классификационные шкалы и градации. Это уникальные текстовые значения из классификационной шкалы второй, которая вторая у нас является. То есть это наши все известные писатели и поэты.

[13m57s] А в качестве описательной шкалы мы рассматриваем слова, которые были во всех этих текстах.

[14m4s] Их 767 в данном случае.

[14m7s] И мы видим, что есть слова, которые имеют одинаковый корень: человек, человека, человечество. Вот, или, допустим, там читатель, читателю, читателям. Вот. Это слова, которые имеют одинаковое происхождение, у которых одно исходное слово, из которого они образованы путём словообразования.

[14m30s] Эти вот исходные слова называются леммами. Лемма - это слово, из которого образована некоторая словоформа.

[14m38s] Лемма всегда в единственном числе именительного падежа. От неё образуются разные словоформы и путём добавления различных элементов слова.

[14m51s] И в системе Eidos есть словарь лемматизации академика Залезняка,

[14m59s] который содержит 2 млн русских слов, словоформ.

[15m7s] И вот здесь вот мы о нём можем почитать.

[15m10s] Вот, профессор Залезняк, академик Академии наук Российской Академии наук. Несколько лет назад скончался.

[15m17s] У него огромный вклад в науку этого учёного. Он переводил новгородские рукописи.

[15m24s] Замечательный такой самоотверженный исследователь, учёный, посвятивший этому всю свою жизнь, с очень хорошими результатами научными.

[15m33s] И одной из его важных работ является создание словаря лемматизации русского языка,

[15m40s] в котором 2 млн слов.

[15m43s] И для каждого слова приведена словоформа. Я вот здесь вот указаны у нас

[15s] ссылочки на страницу на Хабре, где описан этот словарь, где я его нашёл, откуда его скачал. И, значит, он был в Экселе. Я его оттуда скачал и эту экселевскую страничку, экселевский файл. Ну, в Экселе это миллион слов, а у него 2 млн. То есть там было в двух таблицах этот словарь на одном листе. Я сделал небольшую программку, которая эту эти таблицы преобразовала в одну базу данных, база лемма. Здесь на неё ссылочка есть. Это довольно большая, около 200 МБ размер базы, которой система Eidos пользуется, и она может обеспечивать преобразование слов к словоформам, к леммам.

Подраздел 2.8: Просмотр обучающей выборки

[16m48s] Смотрим теперь на обучающую выборку. Обучающая выборка у нас рассматривается в режиме 2.3.1. Здесь мы видим фрагменты текстов, а здесь ещё две таблицы. Значит, да, мы видим наименование фрагментов текстов в верхнем окошке, а в левом нижнем окошке - коды классов, к которым относятся эти фрагменты, а в правом нижнем окошке - коды слов, которые есть в этом фрагменте.

[17m17s] Это всё таблицы, связанные отношением один ко многим, и таблица классов, кодов классов, и таблица кода признаков, связана с таблицей главной наименований источников данных, в данном случае это писатели. И размер этих таблиц, они принципиально особых ограничений не имеют. Связано только есть не очень жёсткие ограничения с самим языком программирования и со свободным местом на носителе, да и всё.

Подраздел 2.9: Синтез и верификация моделей (детально)

[17m47s] Следующий этап автоматизированного системно-когнитивного анализа после формализации предметной области - это синтез и верификация моделей. Этот этап мы проходим в режиме 3.5. На этом этапе создаются статистические модели - это матрица абсолютных частот, матрица условных и безусловных процентных распределений. И на их основе рассчитывается матрица системно-когнитивных моделей.

[18m12s] Потом мы эти модели посмотрим, когда они будут созданы.

[18m18s] Потом обязательно выбирается наиболее достоверная модель. Если она есть, эта модель наиболее, в смысле, достаточно достоверна, тогда мы в ней решаем задачи идентификации, прогнозирования, принятия решений, управления и задачи исследования предметной области путём исследования её модели, моделируемой предметной области.

[18m41s] Если же у нас достоверной модели нет, то мы просто выходим из этого процесса, переосмысливаем задачу и так далее.

[18m48s] Начинаем, по сути дела, заново её ставить и решать.

[18m53s] Теперь мы переходим к этому этапу синтез и верификация моделей.

[19m0s] Это окошко этого режима, в котором это делается.

[19m7s] Режим 3.5.

[19m10s] Здесь мы видим, что та модель, которая сейчас будет создаваться, составляет 80% от того, что может обрабатывать система Eidos. Это по поводу того, что ограничения у неё есть, но очень слабые.

[19s] Если же здесь у нас получится больше 100%, то есть она не может это обработать, то тогда вот здесь можно задать не 100%, а там 50, 20, там 10, 1%. Тогда будут после результатов после идентификации будут оставлены только те результаты идентификации, по которым по внутреннему критерию достоверности, который является адекватным этот критерий, хорошим, являются малодостоверными. Эти результаты будут удалены из результатов идентификации. Оставлены будут наиболее достоверные результаты. И тогда можно обрабатывать модели ещё большей размерности. Также, если очень большой размерности данные, то вот здесь вот последняя строчка, здесь меняется эта строчка, будет выдана рекомендация использовать бутстрепный подход. То есть часть обучающей выборки использовать для создания модели, а часть для оценки их достоверности. Если такая рекомендация будет дана, тогда можно осуществлять только синтез модели или синтез и верификацию, но верификацию на каком-то подмножестве исходных данных, а не на полном объёме исходных данных.

[20m40s] Но сейчас мы будем действовать по умолчанию все параметры возьмём. Я вам рекомендую, ребята, использовать центральный процессор по возможности. Если уже не получается, долго считается, тогда графический процессор. Это связано с тем, что не всегда он работает из-за того, что не у всех компьютерах видеокарты на чипсете Nvidia. Если у вас там GeForce, ZTX, там Nvidia 770 или 1.500, там или что там, какие там сейчас получше есть карты, то можно пользоваться. Даже если там 1040, всё равно там 96 шейдерных процессоров, а на ZTX 770 уже 1.500 шейдерных процессоров. То есть будет примерно, ну, 4, там 5, 7.000 раз быстрее, чем на центральном процессоре считаться. Но само качество расчётов лучше на центральном процессоре.

[21m38s] Сейчас я запускаю на центральном процессоре синтез моделей и их верификацию. Сначала они создаются, потом проверяются на достоверность.

[21m47s] Пока там считается, это не очень долго будет считаться.

Подраздел 2.10: Обзор возможностей Eidos для анализа текстов

[21s] Я вам даю ссылочку на свои работы в области обработки текстов. Эти ссылочки есть вот здесь на этой странице, где мы скачиваем систему. А также эта ссылка есть в самой системе Eidos. Когда мы её запускаем эту систему или когда она уже запущена, мы можем использовать режим 1.1 запустить. Это тот же то же самое окно у нас откроется, которое открывается, когда мы запускаем систему.

[22m30s] И вот я вам даю ссылочку на ссылочки на работы, связанные с анализом текстов.

[22m38s] Вот.

[22m45s] И вот у нас закончился процесс синтеза и верификации моделей.

[22m50s] Это заняло 55 секунд.

[22m53s] Ну, конечно, лабораторная работа, надо признать, небольшой объём содержит текстов.

[23m0s] Вот. Мы это видели, что он небольшой.

[23m5s] Здесь вот у нас 44 строки, около 800 слов там разных.

[23m13s] Можно попробовать и на гораздо больших объёмах. Вот эти статьи, они некоторые из них посвящены решению реальных задач. И вы, значит, можете там посмотреть в этих статьях, как это выглядит.

[23m28s] Но сейчас я вам, ребята, расскажу немножко о том, что вообще можно делать в системе Eidos в плане интеллектуальной обработки текстов.

[23m40s] Значит, мы можем формировать обобщённые лингвистические образы классов, которые ещё называют семантические ядра. Сейчас в интернете, если задать запрос семантическое ядро там создать, то там масса есть информации об этом. Но это ключевые слова к семантических ядрах. Ну, в общем, я могу сказать так, что это слова наиболее характерные для определённого автора, определённого направления науки, определённого жанра, определённых статей или книг, определённого периода и так далее, и так далее.

[24m13s] Значит, система позволяет создавать обобщённые лингвистические образы. То есть это что означает? Что если мы возьмём какие-то литературные произведения определённого периода, определённых жанров, определённой смысловой направленности, определённых авторов, то мы сможем создать обобщённые образы по всем этим категориям.

[24s] И это можно делать на основе фрагментов или примеров текстов, относящихся к этим категориям.

[24m39s] Потом она позволяет количественно сравнивать лингвистический образ конкретного автора, человека или или словесное вербальное описание объекта, процесса какого-либо с обобщёнными лингвистическими образами групп, классов, обобщённой группы или классов. То есть, допустим, мы можем описать словами какие-то места и создать соответствующую модель, которая будет определять, какое место вы описываете. Потом вы описываете какое-то место, это ваше описание сравнивается с описаниями теми, которые есть в модели.

[25m16s] Или можно взять работы разных авторов и создать образы этих авторов, как вот сейчас мы делаем с писателями. И потом берём любой текст и определяем автора. Это называется атрибуцией текстов. Или можем брать сайты и смотреть, анализировать их контент текстовый и определять смысловую направленность этих сайтов. Или если, значит, они какие-то криминальные, можем прямо сразу статью административного или уголовного кодекса определять, которые они нарушают.

[25m52s] Мы можем также сравнивать обобщённые лингвистические образы этих вот категорий различных друг с другом и применять для этого кластерный и конструктивный анализ, то есть дендрограммы можем построить сравнение различных текстов, лингвистических образов различных категорий текстов. Мы можем исследовать моделируемую предметную область путём исследования её лингвистической системно-когнитивной модели. Можем проводить интеллектуальную атрибуцию текстов, то есть определять вероятное авторство анонимных и псевдонимных текстов. Анонимных - это вообще без автора, без указания автора, псевдонимных - подписанных другими авторами.

[26m31s] Также можно определять датировку, жанр, смысловую направленность, содержание текстов.

[26m36s] И всё это можно делать, ребята, слушайте внимательно, для любого естественного или искусственного языка или системы кодирования.

[26m42s] То есть мы можем этот анализ проводить на любом языке, лишь бы у него только было, ну, в общем, не иероглифы, короче говоря, чтобы была фонетическая азбука. Почему? Потому что анализируются слова, написанные буквами.

[27m0s] Вот. Но зато на любом языке это можно делать. В системе Eidos есть поддержка, мультиязычная поддержка интерфейса. И я когда её делал, я проверил все языки, которые были в онлайн-переводчиках, там около 180 языков было, 183, по-моему. Вот. Из них я выделил 50 языков, которые нормально переводятся, и получается, что эти языки основаны на латинском алфавите, кириллическом или кириллическом алфавите. Вот, и всё это прекрасно работает с этими языками. Их 50 оказалось таких языков. Это все основные языки мира, на которых говорит 90% населения Земли.

[27m46s] Вот. Значит, и также можно очень интересные задачи решать сравнение описаний каких-то каких-то текстов на искусственных языках, допустим, на языках программирования. Мы можем составить обобщённые лингвистические образы различных языков программирования и определить их сходство, различия, а также по тем командам и функциям, которые там применяются в этих языках. А также можем просто брать любой фрагмент текста на любом языке программирования, и система будет сравнивать его со всеми обобщёнными лингвистическими образами и определять, на каком языке программирования это написано. То есть довольно такие интересные и разнообразные задачи, довольно актуальные стоят, в общем-то, сейчас и решаются путём интеллектуальной обработки текстов.

Подраздел 2.11: Просмотр результатов анализа (матрицы, нейроны, дендрограммы)

[28m35s] И вот здесь приведены статьи, в которых показано и на численных примерах, как были разработаны семантические ядра научных специальностей ВАК Российской Федерации, автоматическая классификация статей и диссертаций по научным специальностям с применением АСК-анализа интеллектуальной системы Eidos. Там около 680 паспортов научных специальностей проанализировано. Сейчас эта работа актуальная, потому что, когда идёт, готовится к защите диссертация, обязательно создаётся комиссия в составе трёх профессоров, докторов наук, членов диссертационного совета. И эта группа определяет, является ли эта диссертация работой по данной научной специальности, которая есть в совете. Ну и так далее. Ну вот этот вопрос, является ли она диссертацией по этой научной специальности, он может решаться с применением консультирующей системы, которая определит степень сходства текста этой диссертации с паспортами научных специальностей. И, может быть, она э может быть, она не по одной специальности, по двум, по трём, что, между прочим, проблему составляет, потому что нужно, чтобы либо в совете эти специальности были, либо кооптировать туда, включать временно на защиту, на одну защиту учёных по этой специалистов по этой специальности.

[29s] Вот здесь вот у нас вторая статья - формирование семантического ядра ветеринарии путём анализа паспортов научных специальностей тоже.

[30m1s] Вот, интеллектуальная привязка некорректных ссылок к литературным источникам. Это на примере Российского индекса научного цитирования. Это очень интересная задача. Мы знаем, что программное обеспечение РИНЦ не очень хорошо с этой задачей справляется, она просто ведёт разбор, грамматический разбор ссылки. А я, значит, здесь написал в этой статье, что эта задача распознавания образов, и может быть решена методами искусственного интеллекта. Продемонстрировал это, и это получается намного более качественно и намного быстрее, чем с помощью грамматического разбора.

[30m39s] Показано, как можно применить систему Eidos или подходы, которые в ней реализованы, описано в АСК-анализе для решения задач, в общем виде, задачи идентификации литературных источников и авторов по стандартным, нестандартным, некорректным библиографическим описаниям. Это довольно актуальная задача.

[30s] И также можно анализировать анонимные, псевдонимные тексты, рассматривать тексты как модели реальности. И поэтому мы можем говорить о том, что система Eidos сможет решать задачи с очень низким уровнем формализации моделей, исходных данных, то есть просто словесное описание обрабатывать каких-то реалий, которые мы моделируем, а не количественные какие-то использовать их описания. И можем использовать это для того, чтобы осуществить интеллектуальную датировку текста, определение вероятного авторства, жанра. И вот здесь вот есть статьи, где это продемонстрировано на примере русской литературы XIX-XX веков. И вот здесь вот у нас есть статья на английском языке тоже посвящённая этой тематике. Ну это краткий обзор работ моих по этой тематике.

[31s] Вот. То есть мы создали модели. Теперь посмотрим на эти модели, что они собой представляют.

[32m0s] У нас классы представлены соответствующими авторам, представляют собой колонки матрицы модели,

[32m10s] а слова являются строками в этой матрице.

[32m14s] Ну здесь вот у нас матрица такие пустоватенькие. Это что значит? Как это можно интерпретировать? Это можно интерпретировать таким образом, что данных можно было бы представить значительно больше текстовых, чтобы эти всё-таки пустот меньше было. Ну у нас сейчас маленькая учебная задачка, поэтому мы считаем, что это приемлемо. Вот. Но при желании вы можете взять значительно больше объёмы текста, обработать и посмотреть, что получится.

[32m48s] Это модель Abs. Смотрим теперь, какова, значит, её эту модель Abs нельзя использовать для принятия решений. Почему? Потому что у нас разное количество фрагментов по разным авторам. По Грибоедову пять, по Пушкину семь, по Тургеневу шесть, по Горькому три, по Лермонтову семь, по Гоголю четыре. Вот. Поэтому одна и та же единица, если какое-то слово один раз встретилось у разных писателей, вот, допустим, этого у Пушкина встретилось один раз и у Гоголя один раз. Ну так это не одно и то же, потому что у Пушкина семь фрагментов, а у Гоголя четыре фрагмента. То есть для Гоголя это слово более характерно. Мы можем перейти к относительным частотам, условным и безусловным

[33s] и посмотреть,

[33s] какова вероятность встречи этого слова у данного автора.

[33s] Ну, правда, здесь надо говорить не о вероятностях, а об относительных частотах.

[34s] Понимаете? Относительная частота - это не вероятность. Вероятность определяется как предел, к которому стремится, стремится относительная частота при неограниченном увеличении объёма выборки. Неограниченного, конечно, не может быть увеличения объёма выборки. Вот слово этот, ребята, смотрите, значит, 14% вероятность того, что оно у Пушкина встречается, и 25 у Гоголя. То есть у него выше вероятность встречи этого слова. То есть для него оно более характерно. А в среднем по всей выборке шесть, 7%. Значит, мы имеем дело с тем, что у нас условные и безусловные относительные частоты, и мы можем сравнивать. И это сравнение, оно в системе Eidos автоматизировано.

[35s] Каким образом оно автоматизировано?

[35s] Значит, сначала рассчитывается матрица абсолютных частот, которую я вам показывал.

[35s] Значит, здесь на пересечении, значит, каждая строка соответствует слову, колонка - автору или другому классу. Классами могут быть не только авторы, но и периоды времени, и жанры, и направление смысловой направленности текстов, различные направления.

[35s] И здесь вот у нас, видите, есть сумма по строке, сумма по колонке, сумма числа слов и суммарное количество фрагментов текста, на основе которых формировался соответствующий обобщённый образ. Мы можем анализ проводить связи слова с принадлежностью текста с этим словом к тому или иному автору, используя непосредственно матрицу абсолютных частот, используя, скажем, критерий связи хи-квадрат Пирсона. То есть сравнивая фактическую частоту встречи слова с теоретической. А можем перейти к условным и безусловным относительным частотам, просто разделив фактическую частоту на суммарное количество слов, соответствующих этому классу, или число фрагментов текста, соответствующему этому автору, этому классу. А также можем вычислить безусловно относительную частоту встречи этого слова по всей выборке среди всех авторов. И мы можем сравнивать также фактические и теоретические частоты либо путём вычитания, как это принято в хи-квадрат, либо можем путём деления их сравнивать. И тогда получается другие критерии связи. И сами условные и безусловные относительные частоты тоже мы можем сравнивать либо путём вычитания, либо путём деления. И тоже получаются разные критерии.

[37s] Значит, как мы это можем делать? Смотрите. Значит, фактическая частота n и т житое, ну n и т - это сумма по строке, n житое - сумма по колонке, и n - это сумма по всей матрице. Теоретическую частоту Пирсон предложил вычислять по такому по такой формуле: сумма по строке, умноженная на сумму по колонке, разделённая на сумму по всей матрице. И условная относительная частота встречи итого слова в житом классе или итого признака в житом классе равна фактической частоте, разделённая на сумме по колонке житой. А частота безусловная встречи этого признака итого - это частное отделение суммы по строке на число признаков или объектов по всей матрице модели. И дальше мы можем сравнивать фактическую частоту, отнимать теоретическую. Это известная, знаменитая мера Пирсона хи-квадрат. А можем сравнивать эти две величины, фактическую и теоретическую частоты, путём деления. Тогда у нас получается вот это выражение: n и т житое разделить, умножить на n и разделить на сумму по строке и сумму по колонке. А если мы обратимся, посмотрим на определение на формулы для определения условной и безусловной относительной частоты, то мы видим, что вот это вот выражение, когда мы делим фактическую на теоретическую, оно у нас соответствует, ну здесь вот это и написано, фактическую на теоретическую, оно у нас соответствует делению условной вероятности, относительной частоты встречи итого признака в житом классе на безусловную частоту встречи этого признака. Можно, в принципе, применять термин вероятность, но нужно понимать, что всегда будет некая погрешность, когда мы его применяем. И эта погрешность будет уменьшаться по увеличению объёма выборки. Значит, мы также здесь используем логарифм, чтобы когда слово никак не связано или признак никак не связан с принадлежностью объекта с этим признаком к житому классу, чтобы критерий давал ноль. Для критерия хи-квадрат, когда мы сравниваем фактическую и теоретическую частоты путём вычитания, это получается автоматически. А когда мы сравниваем условные и безусловные вероятности, будем говорить, путём деления, тогда надо нормировать к нулю в случае, если нет взаимосвязи. Это можно сделать двумя способами. Первый способ - это взять логарифм этого отношения, которое будет равен нулю, если отношение равно единице. Либо просто вычесть единицу из этого отношения. В первом случае у нас будут получаться количество информации по Харкевичу, его замечательная семантическая мера целесообразности информации. А во втором случае будет получаться коэффициент возврата инвестиций, известный в экономике, как ROI, Return on Investment. Значит, чем это интересно? На мой взгляд, здесь замечательные обнаружены закономерности, которые я не видел в литературе, чтобы они где-то были описаны. Но в моих работах это вот есть. Значит, мы видим, что критерий хи-квадрат взаимосвязан с количеством информации по Харкевичу и с коэффициентом возврата инвестиций. Математическая связь существует между ними. Кроме того, мы можем сравнивать условную и безусловную относительные частоты путём вычитания их друг из друга. И тогда у нас будет получаться известный в электротехнике коэффициент взаимосвязи.

[40s] Вот. Таким образом мы получаем семь вариантов расчёта сравнения фактических и теоретических частот, условных и безусловных процентных распределений. Что интересно? Если мы возьмём разницу между фактической и теоретической частотами, то у нас получается хи-квадрат. Если же возьмём отношение, то есть будем их сравнивать путём деления, то получается у нас формула Харкевича для количества информации.

[41s] Вот. Таким вот образом, ребята. И, значит, используя эти формулы, я их называю частные критерии знаний, которые применяются в настоящее время в АСК-анализе и в системе Eidos для расчёта моделей, для сравнения фактической и теоретических частот, для сравнения условных и безусловных процентных распределений, получаем, соответственно, семь матриц системно-когнитивных моделей.

[41s] Вот. Это я коротко описал вам, как это всё выглядит, как эти расчёты проводятся.

Подраздел 2.12: Просмотр результатов идентификации и достоверности

[42m0s] И теперь мы можем посмотреть на матрицу F1, где мы видим, что в каждом слове есть какое-то количество информации о принадлежности текста с этим словом к тому или иному автору. И здесь уже получается разное количество информации, и положительное, и отрицательное. Если слово у какого-то автора встречается чаще, чем в среднем по всей выборке, тогда количество информации положительное. Если реже, чем в среднем, тогда отрицательное.

[42s] посмотрим на слово этот. Вот у нас обнаружение этого слова даёт нам 0,290 тысячных бита о том, что это произведение Пушкина, 0,369 тысячных бита о том, что это произведение Тургенева, и 0,528 тысячных бита о том, что это произведение Гоголя. Здесь всё нормировано к битам уже, ребята, вот там в модели F1 мера информации по Харкевичу нормирована к битам с помощью весового коэффициента, который я вывел и назвал коэффициент конвергентности Хартли.

[43s] Вот. И теперь и многие почему-то теперь думают, что его предложил Хартли, а не я.

[43s] Значит, ну, теперь, значит, смотрим на меру хи-квадрат. То же самое слово этот. Видим, что это слово для всех писателей не характерно, кроме Пушкина, Тургенева и Гоголя. Причём для Гоголя оно наиболее характерно, на втором месте идёт Тургенев, на третьем Пушкин. Мы видим, что хи-квадрат замечательно сбалансированная мера взаимосвязи, потому что в каждой строке сумма равна нулю, и в каждой колонке равна нулю. И средняя также в каждой строке и колонке тоже равно нулю. Это очень хорошая, хороший результат. Очень интересные получаются модели. Значит, мы другие не будем смотреть.

[44s] Теперь посмотрим, что касается достоверности, что мы можем о достоверности сказать? Что у нас есть разные критерии достоверности моделей. И мы их все применяем эти критерии различные. Ну, какие именно? F-меру в адрес Бергена применяем и её нечёткое мультиклассовое обобщение применяем, которое мною предложено. А также применяем инвариантное относительно объёма распознаваемой выборки, нечёткое мультиклассовое обобщение меры в адрес Бергена. И эти вот критерии, они выделены здесь у нас цветами колоночки. И соответствующие строки, где это значение максимальное, выделены тоже цветом.

[45s] И сейчас я вам пришлю Help по этому, по этому режиму.

[45s] И пришлю второй Help по этому режиму в чат.

[45s] При увеличении объёма выборки сумматоры True Positive, True Negative, False Positive, False Negative растут с разной скоростью. Быстрее всего растёт True Negative. Это означает, что истинно отрицательные решения, они, э встречаются, ну, встречаются, скажем так, в соответствии с моделью, они получаются легче всего. То есть модель не ошибается, меньше всего ошибается при отрицательных решениях. При положительных она ошибается чаще. Вот. Поэтому, значит, у нас False Positive растёт быстрее, чем False Negative. И, соответственно, True Positive и растёт медленнее, чем True Negative. И вот так вот растут, изменяются характеристики достоверности модели при разных объёмах выборки. Мы видим, что F-мера при малых объёмах выборки очень неустойчива. F-мера метрика Ван Ризбергена. То есть очень большие колебания у неё. Где-то до 500 наблюдений. Потом колебания прекращаются, прекращают, уже переходят они в затухающие просто величину. Но не колебательный процесс, а просто уже затухание идёт. И нечёткое мультиклассовое обобщение меры Ван Ризбергена, оно идёт выше. Почему? Потому что к сумматорам True Positive, True Negative, False Positive, False Negative суммируются не единички, а суммируется степень уверенности системы в этом решении. А эта степень уверенности, она достоверна, адекватна. Поэтому получается, что если система не уверена в решении, то мы её наказываем не единичкой False там Positive, ложно положительное решение, а наказываем её в соответствии степень её уверенности в этом решении. А уверенность может быть там 1%. Тогда мы суммируем не единичку, а 1 сотую. И эти меры значительно более устойчивы. Вот нечёткое мультиклассовое обобщение при очень хорошо работает при малых объёмах выборки. И мы видим, что уже при 500 наблюдениях инвариантно относительно объёма выборки мера стабилизируется, тогда как э нечёткое мультиклассовое обобщение F-меры Ван Ризбергена и сама эта мера, они стабилизируются где-то около 3.000, в районе 3.000 наблюдений.

[47s] И мы можем построить э частотное распределение,

[47s] которое покажет нам, сколько истинно положительных, сколько истинно отрицательных, ложно положительных и ложно отрицательных решений получается при различных уровнях сходства. Мы видим, что начиная с уровня сходства 50% примерно, наблюдается большое количество истинно положительных решений. Ложно положительные решения наблюдаются, но только при малых уровнях сходства. А ложно отрицательных решений вообще не наблюдается. Это означает, что модели очень хорошие. Ну это видно и по значениям самих критериев достоверности 0,95, они нормированы к единице эти критерии достоверности. Вот. И даже вот мы видим модель F4 с интегральным критерием семантический резонанс знаний, вообще единица по критерию L1 - это нечёткое мультиклассовое обобщение. То есть вообще нет э ложно отрицательных решений, а ложно положительных очень мало.

[48s] Вот. Если мы посмотрим на разницу между числом истинных и ложных решений, то мы увидим, что соответствующие кривые у нас есть частотных распределений, которые линейно растут при увеличении объёма выборки. Что означает, что уровень достоверности, который даёт система Eidos, является адекватной мерой степени истинности решения. Ну здесь не очень наглядно это видно, но, в общем, видно, что э разница между числом положительных и что между числом истинных и ложных решений э растёт с увеличением уровня сходства, и начиная с уровня э около 20% уровня сходства модели F1, э эта мера информации Харкевича вообще нет ложных решений. Ну мы это видели по кривым этим, и так это глазами видно. Но когда взято отношение, то это ещё удобнее наблюдать.

Подраздел 2.13: Решение задачи идентификации (атрибуции) текста

[49s] Таким образом мы, значит, сейчас с вами рассмотрели вопрос о достоверности моделей. Этот вопрос необходимо рассматривать, потому что если мы не знаем, какая достоверность модели, или знаем, что она низкая, то использовать эту модель для решения реальных задач нельзя, некорректно. Значит, э, ну иногда говорят так: вот смотрите, вот мы её применили, и получилось хорошо. Значит, у нас модель была достоверная. Это верно, но дело в том, что это было очень рискованно, большой риск применять модели, достоверность которых неизвестна. Ну, по крайней мере, для решения выработки каких-то рекомендаций в реальной предметной области этого делать нельзя. Можно поэкспериментировать с такой моделью, но это как раз означает измерить её достоверность таким способом. Если она окажется достоверной, тогда уже можно её применять. Наши модели оказались очень высокой достоверности, и их можно вполне применять для решения задач.

[50s] Задачи мы будем решать идентификации текстов и типологический анализ будем рассматривать, какие слова наиболее характерны, не характерны для тех или иных текстов. И будем проводить их кластерный анализ самих текстов и авторов. То есть посмотрим на исследование объекта моделирования путём исследования модели.

[51s] Как это будет выглядеть, ребята? Ну, прежде всего, значит, давайте посмотрим на модель режим 448. Это режим, который выводит семантическое ядро.

[51s] Вот, допустим, мы берём, ставим здесь вверху курсор на классе на каком-то, соответствующем писателю. Нажимаем модель F3, в которой мы будем решать задачи. И видим, что семантическое ядро Пушкина - это Руслан, Евгений, Людмила, Онегин. Понятно, да? Если мы сейчас зададим построение свод диаграммы графической, то будет представлено в графическом виде это. Вот в таком.

[52s] То есть здесь мы слева видим слова характерные для Пушкина, справа не характерные. Цвет - это характерность, не характерность. Красная - характерность, синяя - не характерность. Толщина линии - степень характерности, не характерности. То есть мы просто видим наглядно, что вот слово Руслан, оно имеет характерность там в два раза выше, чем слово Онегин, например. По этим фрагментам Пушкина. Вернее, этим фрагментам о Пушкине, которые мы здесь использовали в качестве исходных данных.

[52s] Значит, фактически мы можем построить семантическое ядро каждого писателя, которого у нас здесь был в качестве исходных данных представлен. Вот. И мы видим прекрасно, значит, как это выглядит.

[53s] Вот, вот, допустим, Тургенев, семантическое ядро: Базарова, Базаров, взаимоотношения, время, общество, отцы, родители, дети, поколения, аристократический, безделие и так далее. А вот справа находится антиядро, семантическое антиядро Тургенева. То есть это слова, которые для него наиболее не характерны: свои, жизни, самое, Руслан, героем, Грибоедов, стихотворение, своё, Раскольников, Пушкин, поэмы. Это слова наиболее не характерные для Тургенева. И вот так вот по каждому писателю мы можем посмотреть. Или если мы будем модель создавать более развитую, в которой не только будут писатели, но и периодизация будет, то есть временные периоды, и жанры, и направление смысловой направленности текстов, различные направления. Вот, и так далее, и так далее.

[54s] Вот. Значит, теперь посмотрим, как решается задача идентификации. Для того, чтобы решить задачу идентификации, простейшую задачу, для этого нам на основе того, что у нас уже созданы семантические потенциалы, положительные и негативные авторов, э среди потенциала ядра, то есть классы. Сейчас мы видели, по сути дела, классы, какие у нас созданы, и какие слова характерны, не характерны, в какой степени для того или иного автора. На основе этого можно решать задачу атрибуции текстов, то есть поиска класса, соответствующего автору, к которому принадлежит тот или другой фрагмент текста. Для этого мы используем наиболее достоверную модель, которую мы возьмём в качестве такой модели, модель F3. Здесь на самом деле особой разницы нет, какие модели брать, потому что я вам раскрою небольшой секрет, ребята, такой, тайна Полишинеля, которая заключается в том, что у нас каждый класс описан настолько большим количеством признаков, и среди них довольно много уникальных, детерминистских признаков, что, в принципе, никакой проблемы не представляет собой решение задачи идентификации. Сейчас мы выбрали модель F3 в качестве текущей, и в этой модели будем решать задачу идентификации. Для чего так сделано? Дело в том, что задача идентификации является наиболее трудоёмкой в вычислительном отношении. То есть если у нас, допустим, 1.000 есть объектов распознаваемой выборки и 100 классов, и, допустим, 1.000 признаков, что вполне реально, как вы видите, то, значит, у нас цикл по классам, то есть цикл по объектам распознаваемой выборки сначала идёт. Вот. Потом каждый объект сравнивается со всеми классами, поэтому дальше идёт цикл по классам. А потом дальше идёт цикл по признакам объекта, чтобы сравнить объект с классами. Ну, конечно, это крайне трудоёмкая вычислительном отношении операция. Поэтому я в своё время, когда проектировал эти режимы системы Eidos, а я их проектировал очень давно, где-то в восьмидесятых годах. Первый акт внедрения - восемьдесят седьмой год. Я вам даже покажу. А расчёты в восемьдесят первом году проведённые на основе этой модели.

[56s] Первый акт внедрения.

[57s] Значит, это тогда работал главным конструктором Кубанского аэрокосмического центра. Вот внизу там моя подпись есть. И по заданию, по договору с институтом социологических исследований Академии наук СССР, Краснодарским сектором, которого руководил кандидат философских наук Айтыч Ибович Кагуров. Сейчас его сын - первый проректор Кубанского госуниверситета Тимур Айтычевич. Ну потом, конечно, Айтыч Ибович защитил докторскую по социологии, стал профессором. И сын у него тоже доктор социологических наук, профессор, первый проректор. Вот смотрите, ребята, значит, ещё в восемьдесят седьмом году формировались, видите, информативности признаков вычислялись, формировались позитивные, негативные информационные портреты социальных категорий. А эти вот позитивные, негативные информационные портреты, они в свод-анализе слева и справа расположены. То есть это как раз и есть списки признаков, расположенных в порядке убывания количества информации о принадлежности объекта к тому или иному классу. Ну там в качестве классов были социальные категории. И также определялась ценность признаков для решения задач идентификации, вопросы оптимизации решались описания. То есть можно было выделить наиболее существенные признаки, остальные выкинуть. То есть при пилотном исследовании это осуществлялось, а при промышленной эксплуатации уже в анкете были только наиболее значимые вопросы, те, которые наиболее ценными были для того, чтобы отличить одну категорию респондентов от другой.

[58s] Вот. То есть это, в общем-то, давняя история. И вот, значит, мы здесь видим слева фрагменты текстов разных писателей, а справа результаты их идентификации. Сейчас я с помощью курсора поперемещаюсь тут немножко по этой картинке.

[59s] И вы, а вы смотрите справа, верхнее и нижнее окошко. Мы видим, что всегда во всех случаях для всех фрагментов текста наиболее похожим классом является класс, соответствующий данному автору. И никогда нет никакого другого варианта. То есть эта задача решается чрезвычайно точно, с очень высокой точностью, безошибочно. Хотя вот, допустим, фрагмент Достоевский 1, он похож на Достоевского на 82%, а на Гоголя на 1%. И по другому интегральному критерию тоже он похож на 82% на Достоевского, а на Гоголя на 0,9% похож. То есть, видите, это это 0 это 0,9 - это ложно положительное решение или вот 1,1%, да? Но по Ван Ризбергену мы суммируем единичку к False Positive, а я предлагаю, что мы в этом случае суммируем 1 сотую. Почему? Потому что сама система крайне не уверена в том, что это фрагмент относится к горькому. Всего 1% сходства. Этот критерий сходства, он как раз является критерием степени достоверности предлагаемого решения. Это мы видели уже неоднократно на всех, на самых различных задачах.

Подраздел 2.14: Сравнение объекта с классом

[1:00s] Теперь рассмотрим, как система сравнивает конкретный фрагмент текста с классом. Ну сначала ещё посмотрим на другую разрез этой же самой формы. Слева мы видим классы, а справа видим фрагменты текста, относящиеся к этому классу. Мы видим, что некоторые фрагменты текста являются более типичными для обобщённого образа класса, некоторые являются как бы исключением. Вот наиболее типичный - 85% сходства, а наиболее нетипичный - 66 по верхнему интегральному критерию семантический резонанс знаний. Ну а по критерию сумма знаний мы видим, что 100% сходства наиболее похоже, а наиболее не похоже - 65. И вот так вот мы видим по всем писателям. Вот у Горького, видите, есть сходство и с Пушкиным, и с Грибоедовым, и с Тургеневым, и с Достоевским, но очень маленькое: 4%, 2%, 2. Вот. То есть это ложно положительное решение с очень низким уровнем сходства.

Подраздел 2.15: Теоретические основы сравнения (Лемма Неймана-Пирсона)

[1:01s] То есть система работает очень хорошо. То есть она, скажем так, очень уверенно решает задачу атрибуции текстов, отнесения текстов к тем или иным авторам. Теперь давайте, как она это делает, рассмотрим. Значит, мы уже видели в режиме 5.5, что система определила, какое количество информации содержится в том или ином слове о принадлежности текста с этим словом к тому или иному автору. Вот мы видим это прямо в битах. Вот слово автор Пушкину 0,468 бита, Толстой 0,509 бита. И вот так вот по всем словам мы имеем эту информацию. А если у нас известно, что не одно слово, а много слов в фрагменте, тогда как определить, к какому автору принадлежит этот фрагмент текста? Тогда мы, основываясь на лемме Неймана-Пирсона, э, которую я сейчас вам приведу, эту лемму, воспроизведу, решаем, каким образом это сделать. Значит, вот я нажимаю на интегральный критерий сумма знаний, смотрим, как он устроен. Значит, лемма Неймана-Пирсона. Если у нас есть много разных гипотез о принадлежности объекта к разным классам, эта лемма связывается с критерием Пирсона хи-квадрат. Но замечательный математик с блестящим, так сказать, оригинальным мышлением, неординарным и сверхспособностями вроде бы в области вычислений умственных, Фон Нейман, на компьютерах, которые он придумал, мы все сейчас работаем. По сути дела, это он описал, как эти компьютеры могут выглядеть, какой у них там процессор, оперативная память, внешняя память, что такое команды, программы, всё это Фон Нейман, ребята. Так вот, Фон Нейман с Пирсоном доказали следствие из теоремы, которое называется леммой о том, что наиболее вероятной является та гипотеза о принадлежности объекта к тому классу, той категории, о принадлежности к которой в системе признаков этого объекта наибольшее количество информации. То есть они предложили аддитивный интегральный критерий, основанный на теории информации. Правда, они применяли, использовали не теорию информации, а меру хи-квадрат. Но мы знаем, что она связана с теорией информации, с информацией. Остаётся только удивляться их интуиции, их пониманию глубокому этих всех процессов и этого предмета вообще, этой предметной области. Хотя они не рели об информации, но, по сути дела, они говорят, так скажем так, об информации не рели не совсем обоснованно, потому что в то время ещё этого не было показано этой связи. Вот, когда они об этом говорили, но они её как бы интуитивно чувствовали, что это вообще слово вполне уместно в данном случае. И были правы. Так вот, ребята, мы можем просто взять и посчитать суммарное количество информации в признаках какого-то объекта о принадлежности его к какому-то классу. И сделать это для всех классов. А потом ранжировать или рассортировать классы в порядке убывания количества информации о принадлежности к ним. И считать, что это и есть результат распознавания. И считать, что объект относится к тому классу, о принадлежности к которому больше всего информации содержится в его системе признаков. Но в какой-то степени он, может быть, и к другому классу принадлежит, и к третьему, но в меньшей степени, вот так скажем. Как можно просуммировать количество информации признаков объекта о принадлежности к классу? Это можно сделать двумя способами. Значит, нам нужно знать, во-первых, есть этот признак у объекта или нет. Для этого удобно ввести некий массив. Я назвал этот массив массив локатор, L обозначил, который описывает сам объект, распознаваемый. И в этом массиве можно элементы либо True False взять, если признак есть True, нет признака False. Либо, значит, ну, в данном случае у нас в качестве признаков выступают слова. Либо можно поставить там в качестве значений элементов координаты 0,1. Тогда как вектор можно рассматривать этот массив. И если признак есть один, если нет - ноль. Так вот, если там True False, то можно поставить в цикле if, если признак есть, если там один там или True, например, то суммируем. Если False, не суммируем. Вот, количество информации. Либо можно просто умножить элементы матрицы системно-когнитивной модели, допустим, по житой колонке, итая строка, просто просуммировать, то есть просто умножить на элемент массива L итой, который либо нулю, либо единице равен. И у нас либо просуммируется это количество информации, либо не просуммируется, если ноль, то умножено будет на ноль, ничего не просуммируется. Вот это вот выражение, сумма произведений двух координат двух векторов - это вообще-то скалярное произведение, которое имеет геометрический смысл косинуса угла между двумя векторами: вектором класса из матрицы системно-когнитивной модели и вектором, описывающим объект. Эта мера имеет интересные свойства, ценные, дающие определённые преимущества этому интегральному критерию. Во-первых, этот интегральный критерий имеет неметрическую природу, то есть он не является мерой расстояния, а просто, но хотя и является мерой сходства векторов, но не является расстоянием. Расстояние измеряется, в частности, вот эвклидово расстояние в пространстве, плоском пространстве эвклидовом, ортонормированном, измеряется по теореме Пифагора, как корень квадратный из суммы квадратов координат или суммы квадратов разности координат. Вот. И, в общем, все применяют эту меру широко, не задумываясь особенно, что она некорректна для неортонормированных пространств. А на практике, я вам скажу, всегда получаются неортонормированные пространства. Что это значит? Это значит, что эти вектора классов, которые вот у нас получаются в матрицах системно-когнитивных моделей, они всегда между собой коррелируют, не взаимно перпендикулярны. То есть угол между ними не равен 90°. Ну это из определения скалярного произведения я сейчас основываюсь. Поэтому эта мера, межвекторное расстояние я называю, или или информационное расстояние, она гораздо более обоснована, чем эвклидово расстояние и даёт корректные результаты в неортонормированных пространствах, которые всегда и встречаются на практике. Второе интереснейшее свойство этого этого интегрального критерия заключается в том, что он по своей математической форме является определением белого шума. Но когда мы определяем белый шум, то ещё здесь вот этот результат суммирования, он равен нулю пишется, то есть сумма равна нулю. Что это значит? Это значит, что если мы возьмём любой фрагмент случайного сигнала, шума, и возьмём два фрагмента с одинаковым числом отсчётов, значений, построим на них два вектора, посчитаем скалярное произведение, то это скалярное произведение будет стремиться к нулю при увеличении длины этих векторов, то есть числа координат у них, количество координат у них. Что это по-простому так означает? Это означает, что любой фрагмент шумящего шумового сигнала никак не связан с другим любым фрагментом этого сигнала. И когда мы берём небольшое число координат, то связь формально получается, но она имеет случайный характер. И когда мы асимптотически устремляем длины этих фрагментов к бесконечности, то тогда корреляция между любыми двумя фрагментами шумящего сигнала стремится к нулю. Это означает, что этот интегральный критерий подавляет шум, который всегда присутствует в исходных данных. Представить себе, что исходные данные без шума - это вообще нереально абсолютно. Всегда присутствует в исходных данных шум. И поэтому всегда он и проникает в модели сами. Ну хотя вот в системе Eidos модели Abs считаются таким образом, что уже выделяется сигнал из шума. Причём применяются методы при этом, которые применяются в геофизике для подавления шума. То есть мы там не средние вычисляем значения, а суммируем, суммируем, суммируем. При этом у нас сигнал при этом суммировании как бы умножается на число наблюдений, а шум стремится к нулю, который в этом сигнале есть. Даже уже в матрице частот. А потом, когда мы уже вычисляем остальные модели и применяем вот такой интегральный критерий, то ещё больше у нас шум подавляется. Поэтому система Eidos очень хорошо работает на зашумлённых данных и показывает возможность выявления зависимостей в сильно зашумлённых данных. Значит, свойство подавлять белый шум проявляется у данного критерия интегрального тем ярче, чем больше в модели градаций описательных шкал, то есть признаков. То есть больше длина вот этих векторов классов и описания объекта. Этот интегральный критерий, третье его интересное свойство, представляет собой количественную меру сходства, различия конкретного объекта и обобщённого образа класса. И имеет тот же самый смысл, что и функция принадлежности элемента множеству в нечёткой логике Лотфи Заде, например. Не сама функция, а значение этой функции. Однако в нечёткой логике эта функция задаётся исследователям априорно, ну то есть просто так от фонаря или на основе каких-то его соображений не не формализованных, на основе опыта, интуиции. Вот. Здесь же мы имеем интегральный критерий, который хорошо обоснован теоретически и рассчитывается непосредственно на основе эмпирических данных. применяется, точнее, на основе моделей, рассчитанных непосредственно на основе эмпирических данных. Это очень большая ценность, в этом заключается большая ценность такого подхода. То есть мы не просто придумываем какие-то там на основе опыта, интуиции, профессиональной компетенции, какие-то функции принадлежности, а непосредственно на основе эмпирических данных получаем эту степень принадлежности объекта к классу с использованием этого интегрального критерия. В-четвёртых, этот интегральный критерий представляет собой адекватную самооценку системы, степени уверенности системы в предлагаемом решении, положительном или отрицательном, о принадлежности или непринадлежности объекта к классу. И соответствует адекватно риску ошибки при этом решении. Кроме того, интереснейший момент в этом есть, ребята. Значит, сейчас я вам дам ссылочку на монографию новую свою, в которой описано, можно считать, что система Eidos разлагает образ объекта, идентифицируемого или ситуации в ряд по функциям классов. То есть функцию объекта или состояния идентифицируемого разлагается в ряд по функциям классов. Что здесь имеется в виду? Если мы возьмём теорию рядов, таких наиболее известных, как Фурье, Макларена, Лагранжа, Бином Ньютона, вот, и менее известных, таких как ряды по спецфункциям, Бесселя, например, экспонентам, вот, Чебышева, то мы увидим, что коэффициенты при функциях, которые базисными считаются, по которым происходит разложение в ряд, они представляют собой что-то вроде коэффициента корреляции между функцией, разлагаемой в ряд и данной базисной функцией. Мы даже просто чисто математически формально это видим по самой форме этих вот коэффициентов разложения в ряды. Из этого вытекает очень интересный вывод, что мы в системе Eidos, когда получаем вот такую картинку, то можем считать так, что вот этот фрагмент текста - это тоже некая функция, в которой есть или нет какие-то слова, которые у нас есть в справочнике, то есть вектор. И этот вектор разлагается в ряд по векторам классов. И мы видим, что в этом разложении коэффициент корреляции с вектором класса Достоевский очень большой имеет значение, а горький очень маленькое, а остальные вообще отрицательные. То есть если мы проведём действительно сложение этих вот функций из системно-когнитивных моделей с этими весовыми коэффициентами, мы получаем с очень высоким приближением, качественным профиль идентифицируемого объекта. Вот, ребята. Значит, теперь вот в этой монографии, которую я сейчас вам послал ссылочку на неё в ResearchGate и, значит, сейчас я пошлю вам со своей, со своего сайта. Вот, значит, сейчас я вам пошлю ссылочку на эту монографию сначала целиком, как она есть. А потом пошлю обложку этой монографии. Обложку этой монографии. Что-то я не особо. Ладно, я просто ссылку на неё пошлю. Потом можно класнуть и посмотреть обложку. Хорошо. Так, ребята, у нас сейчас перерыв. Но я так думаю, что не надо выключать запись, а продолжим на следующем занятии, и потом я попробую вам рассказать про, то есть начнём следующее. Нет, ну, видимо, продолжим эту работу, описание, рассмотрение этой работы продолжим. И дальше, значит, во втором занятии. Вот. Так что, ребята, перерыв небольшой, 10 минут.

Раздел 3: Продолжение занятия и ответы на вопросы

Подраздел 3.1: Возобновление занятия

[1:15s] Ну что, продолжим?

[1:18s] Как вы на то смотрите, чтобы продолжить?

[1:23s] М?

[1:25s] Ребят, кто за то, чтобы продолжить?

[1:27s] Охотно.

[1:29s] Охотно.

[1:30s] Ну ты идёшь на сам экзамен прямо. А кто ещё охотно?

[1:36s] Да, можно продолжить.

[1:39s] Ну давайте.

[1:42s] Значит, э

[1:44s] Так, и видео запись идёт у нас, да? И экран вы видите.

Подраздел 3.2: Ответ на вопрос студента (про обработку картинок)

[1:49s] Значит, ребята, второй интегральный критерий. Почему здесь два окошка? Выше и ниже? Потому что ниже сумма знаний интегральный критерий. Это скалярное произведение между вектором класса и вектором объекта.

[1:57s] Вверху другой интегральный критерий, который я назвал семантический резонанс знаний. Он похож очень на первый по смыслу. И, в общем, он получается из первого путём замены координат векторов на их стандартизированные значения. Это делается путём вычитания из значения координаты среднего по координатам этого вектора и делением на среднеквадратичное отклонение координат этого вектора, значение координат этого вектора. Это делается для вектора класса и для вектора объекта. Если мы подставим сюда это под сумму, вынесем за сумму стандартное среднеквадратичное отклонение, которое по индексу G, а не по индексу I. Вот, а сумма по I вынесем их. Вот, то, значит, мы фактически получаем формулу из скалярного произведения вот этого, которое я только что рассказывал, получаем формулу коэффициента корреляции Пирсона. Опять мы видим, слышим фамилию Пирсона, да? Который тоже эту меру ввёл, как не меру расстояний, а меру сходства двух массивов или двух векторов. По сути дела, это скалярное произведение двух стандартизированных векторов, то есть таких, которые путём замены координат на стандартизированные значения, приведены к единичным векторам. Это так называемые единичные векторы, которые очень часто математики исследуют и используют.

[1:54s] А так, в общем, смысл тот же самый. Это скалярное произведение стандартизированных векторов класса и объекта.

[1:56s] Ну вот, собственно, и всё.

[1:58s] Ну, да, я это записал полностью всё, Анатолий.

[2:00s] Ну только не все способы анализа, а только спектральный анализ.

[2:04s] Да, конечно, я полностью описал. И вот в книжке, которая там приведена, сейчас там в чате очень много информации. В книжке, которая приведена, ссылочка на которую есть в чате, там есть пятая глава, в которой это подробно описано. И в этой пятой главе есть список литературы, в которой тоже это подробно описано.

Подраздел 3.3: Заключительные слова

[2:06s] Сегодня вместе с нами, ребята, работали в Калифорнии, в США, а также э что это за городок тут у них? Омаха, США. Вот. И также мы работали вот, как мы видим здесь вот в Краснодаре. И ещё работали в Москве семь человек. Так что сегодня прямо такой день, немного, много в мире запускали систему Eidos.

[2:31s] Пожалуйста, какие вопросы у вас есть, ребята?

[2:34s] Давайте вопросы.

[2:39s] По вот этим вот моментам, которые мы прошли. Это довольно полно всё изложил. Вы вообще там ещё, нет? Кто там? Да. Ясно. Да. Ну это