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

***Российская Федерация***

**39 Лабораторная работа № 3.02. Пример интеллектуальной обработки текстов. 2020-11-14**

## Заголовок

Система Eidos: Интеллектуальная обработка текстов, атрибуция и оценка моделей (Лаб. 3.02)

## Резюме текста

Лекция посвящена лабораторной работе 3.02 "Пример интеллектуальной обработки текстов" в рамках дисциплины "Интеллектуальные системы и технологии в науке и образовании". Проводится обзор системы Eidos, её обновления и установки более полной версии для выполнения работы.

**1. Введение и обзор системы Eidos:**
Лекция знакомит с системой Eidos, предназначенной для интеллектуального анализа данных, включая тексты. Обсуждается актуальная версия системы (от 13.11.2020) и процесс её обновления. Подчеркивается важность обратной связи и вопросов во время лабораторных работ и лекций.

**2. Теоретические основы обработки текстов:**
Рассматриваются концепции математической и когнитивной лингвистики как основы для интеллектуального анализа текстов. Система Eidos позволяет проводить такой анализ, включая атрибуцию текстов (определение авторства анонимных или псевдонимных текстов). Упоминается возможность работы с текстами на разных языках (латиница, кириллица), но с ограничениями по иероглифам и некоторым специфическим шрифтам.

**3. Функционал системы Eidos для анализа текстов:**

* **Атрибуция текстов:** Определение авторства, работа с анонимными и псевдонимными текстами.
* **Обработка текстовых полей:** Возможность использовать текст целиком, отдельные символы или слова как признаки для анализа. Для ЛР 3.02 используется режим, где слова рассматриваются как признаки.
* **Лингвистические образы:** Формирование обобщенных лингвистических образов (семантических ядер) классов текстов (например, авторов).
* **Сравнение и идентификация:** Количественное сравнение текстов с обобщенными образами для идентификации или классификации.
* **Датировка и жанр:** Возможность определять датировку, жанр и смысловую направленность текстов.
* **Лемматизация:** Приведение словоформ к их первообразной форме (лемме) с использованием встроенной базы (основанной на словаре Зализняка), что позволяет сократить количество признаков и повысить качество анализа (хотя в текущем примере не используется).

**4. Практическая работа с системой (ЛР 3.02):**

* **Ввод данных:** Демонстрируется простой способ ввода текстовых данных из Word в Excel, где каждый абзац помещается в отдельную ячейку. Подчеркивается необходимость отключения переносов в Word перед копированием.
* **Настройка анализа:** В диспетчере приложений Eidos выбираются параметры анализа: классификационная колонка (автор), описательная колонка (текст), включается режим специальной интерпретации текстовых полей (слова как признаки).
* **Пример анализа:** Анализируются фрагменты сочинений о русских писателях-классиках (Достоевский, Гоголь, Горький и др.). Система формирует модели и позволяет сравнивать тексты.

**5. Оценка достоверности моделей:**
Обсуждается важность оценки достоверности результатов. Вводится понятие F-меры Ван Рисбергена и её ограничения (зависимость от объема выборки, бинарность оценки). Предлагается авторское обобщение – нечеткая мультиклассовая инвариантная мера L2, которая лишена этих недостатков и показывает более стабильные и адекватные результаты, особенно на малых выборках. Графики демонстрируют нестабильность классической F-меры и устойчивость L2-меры при изменении объема данных.

**Заключение:**
Лекция закладывает основу для понимания интеллектуальной обработки текстов с помощью системы Eidos, демонстрирует практические шаги по вводу данных и настройке анализа, а также знакомит с методами оценки достоверности полученных моделей. Следующее занятие будет посвящено решению задач идентификации и другим аспектам работы.

## Детальная расшифровка текста

**Раздел 1: Введение и обзор системы Eidos**

**Подраздел 1.1: Приветствие и информация о занятии**

Здравствуйте, ребята.
Доброе утро.
Здравствуйте.
Доброе утро.

Так, сегодня у нас 14 ноября 2020 года. Суббота, кстати. Вторая пара, 9:40–11:10 по дисциплине "Интеллектуальные системы и технологии в науке и образовании".

**Подраздел 1.2: План работы и установка системы**

И мы начинаем изучать лабораторную работу 3.02. Это пример интеллектуальной обработки текстов. Очень упрощённый такой, учебный пример. Занятие ведёт профессор Луценко Евгений Вениаминович.

Вот. Если у вас какие-то возникают вопросы, вы можете их задавать, потому что лабораторная работа, да и на лекциях тоже можете. Никаких проблем, пожалуйста.
Вот.
Привет, Артём. Появился.
Здравствуйте.
Да. Так вот, ты прямо вовремя пришёл. Так вот, для того, чтобы эту работу пройти, мы сейчас установим более полную версию системы. Там минимальная версия стояла.

И эта работа, она встроенная, то есть она есть там в системе. В самой инсталляции.

Система новая, новая версия.

**Подраздел 1.3: Обновление системы и актуальность**

Значит, у меня довольно много работ есть по анализу текстов. Это к интеллектуальному анализу текстов. Это то, что относится, называется математическая лингвистика. Я бы даже сказал, не просто математическая лингвистика, а какая-то когнитивная лингвистика, вот так вот, или интеллектуальная лингвистика. То есть это применение интеллектуальных технологий для анализа текстов.
И вот я только-только запустил систему, и тут, видите, есть кнопочки, ссылки на некоторые подборки публикаций. И, в частности, подборка публикаций по анализу текстов. Вот я сейчас эту подборку по этой кнопочке кликнул. И появился список публикаций по этой тематике со ссылками на них. То есть их можно почитать. Я вам сейчас вот эту ссылочку кину в чат. То есть вы можете по ней кликнуть, и там довольно много работ есть по анализу текстов.

Обратите внимание, система 13.11.2020. То есть это достаточно новая версия, вчерашняя. Это внёс некоторые исправления, усовершенствования, как обычно. Вот всё время над ней работаю, замечаю что-нибудь не то, стараюсь исправить. Ну, не всегда есть возможность, конечно.
Вот. Теперь обращаю ваше внимание, что я обновил на своём сайте. Вы, наверное, ссылочку знаете на него, но я всё равно пошлю.
Вот. Обновил инструкцию по разработке. Само описание вот здесь обновил тоже системы и инструкцию обновил для учащихся, то есть для вас, по разработке собственных приложений.

**Подраздел 1.4: План изучения тем и лабораторных работ**

И вот здесь у нас, видите, план какой? Сначала мы подробнейшим образом изучаем лабораторную работу 3.03 модельную, очень упрощённую, которая легко просматривается с помощью естественного интеллекта. Потом изучаем работу по анализу текстов, интеллектуальному анализу текстов. Потом более-менее реалистичное приложение по разработке методики риэлторской оценки. А потом вот это вот всё – это разработка модели, обеспечивающей спектральный анализ изображений, обобщение спектров, то есть формирование спектров классов обобщённых, а потом решение всех задач: сравнение конкретных спектров с обобщёнными, обобщённых друг с другом и так далее, и так далее. То, что я рассказывал уже. После этого мы начинаем разрабатывать своё приложение по этой инструкции.
У вас по этой дисциплине зачёт или экзамен, ребята?
Зачёт.
Зачёт. Ну, всё равно для вас это было бы очень полезно, если бы вы провели маленькое исследование сами. Ну, это мы к этому подойдём попозже, наверное, если семестр не закончится. Вот. И могли что-то сделать сами, понимаете? Не просто вот слушали, смотрели, а сами пробовали. Кстати, сейчас вот у вас лабораторная работа, очень желательно было бы, если бы вы на своих компьютерах вот пытались повторить то, что я делаю.

Вот. Для этого надо скачать систему, установить её или обновить.
А как её обновить?
Обновить очень просто, ребята. Вообще элементарно. Вот заходите туда, где вы её скачиваете. А, давайте я покажу, как это в системе сделать. Я как раз это сделал вчера. Это режим 6.2. И вот здесь вот, видите, есть ссылочки на сам сайт, на монографии, учебные пособия, на научные статьи. Все ссылочки действующие. Вот. И скачать полный архив можно заново. А можно минимальную скачать инсталляцию, а можно скачать актуальное текущее обновление, патч, около 10 МБ. Это downloads.exe.

Downloads.exe, видите, вот здесь вот скачать lc.google.ru downloads.exe. Вот здесь вот кликнете и у вас скачается обновление. Потом нужно выйти из системы обязательно и развернуть это обновление в папке, где находится исполнимый модуль. И тогда этот исполнимый модуль заменится. И там ещё кое-какие файлы заменяются обычно, а что-то не только одно изменение произвожу в исполнимом модуле, а ещё могут какие-то файлы меняться.
Вот. Значит, теперь у вас, когда вы систему запустите, должна быть версия 13.11.2020.

**Раздел 2: Лабораторная работа 3.02: Интеллектуальная обработка текстов**

**Подраздел 2.1: Запуск и настройка ЛР 3.02**

Заходите теперь в диспетчер приложений и устанавливаем лабораторную работу 3.04. Это атрибуция анонимных и псевдонимных текстов.

При этом у нас используются новые возможности, ну, в смысле, новые не для системы, а для вас новые. Такие, которые раньше мы не изучали.

Значит, мы видим экранную форму программного интерфейса 2322. И очень странные параметры здесь приведены, которые раньше вы не видели. Потому что мы решали очень упрощённую задачу, а сейчас мы решаем более развитую задачу.

**Подраздел 2.2: Параметры анализа текста в Eidos**

Значит, смотрите, ребята. Значит, у нас классификационная колоночка одна задана. Ну, тип данных старенький Excel. Описательная колоночка тоже одна. Но я хочу сказать вам, что может быть и классификационных много, и описательных много. И вот в этой ссылочке на работы по обработке текстов, там такие примеры приводятся.
А потом самое вот важное. Смотрите, вот здесь вот стоит птичка: применить специальную интерпретацию текстовых полей признаков. И здесь, видите, когда эту ставишь этот пункт, то появляется, ну, если его убрать этот пункт, то это исчезает возможность поставить. Если применить, тогда здесь появляется возможность значение полей целиком использовать для как значение поля, как признак, как признак. А можно использовать символы. Символы, прямо отдельные символы как признаки рассматривать. А можно слова рассматривать как признаки. И вот у нас стоит этот вариант: рассматривать слова как признаки. Это и позволяет обрабатывать тексты. То есть признаки рассматриваются как признаки тех или иных объектов, тех или иных классов. Каждая строчка – это тоже, как обычно, у нас наблюдение, то есть это текст определённый. А дальше сейчас я буду рассказывать уже, рассматривая этот текст.

**Подраздел 2.3: Ввод и обработка данных**

Сейчас этот файл вводится в систему исходных данных. Этот файл текстовый, то есть, в смысле, экселевский, таблица, но значения ячеек все текстовые. И вот это очень упрощённая форма калькулятора внутреннего, где нет числовых полей, поэтому здесь не спрашивается, сколько диапазонов ни по классификационной шкале, ни по описательной.

Вот. И мы просто одна классификационная шкала текстовая, одна описательная. Вот. Поэтому мы просто двигаемся дальше на создание модели. Дальше у нас что? Создаются, как обычно, справочники классификационных и описательных шкал градаций, и с их помощью кодируются исходные данные. И получается обучающая выборка. Потом всё переиндексируется.

**Подраздел 2.4: Интерпретация результатов и дальнейшие шаги**

Потом, видите, нам для студентов я пишу эту экранную форму сделал, что нужно открыть файл исходных данных, посмотреть на него по пути. Вот этот путь будет всегда у вас соответствовать тому месту, где фактически на вашем компьютере находится система Eidos. Студенты обычно эту форму вообще не читают, а сразу о'кей нажимают. Я их заставляю опять повторить всё сначала, чтобы дойти до этой формы, чтобы они прочитали и выполнили то, что здесь написано. А здесь написано, смотрите, ребята, что надо открыть файл исходных данных. Давайте это и сделаем.

**Раздел 3: Практика ввода данных и пример анализа**

**Подраздел 3.1: Пример файла исходных данных**

Вот у нас этот файл исходных данных. Значит, казалось бы, как вот вводить так вот тексты в экселевский файл? Возможно, это как-то трудоёмко. Может быть, вам покажется, что это трудоёмко. Сейчас я вам покажу, как это делается. Значит, я закрываю этот файл. Сейчас просто открываю Excel. Открываю какой-нибудь вордовский файл. Ну, допустим, пособие какое-нибудь открываю.

Ну, просто некое пособие учебное. Ну, в общем, любой файл фактически в Ворде. В Ворде открываете этот файл. И я дохожу до такого места, где есть абзацы. Вот смотрите, я эти абзацы вот так выделяю. Ну, выделю вот на одной страничке. И в экселевский файл вставляю. Смотрите, ну, вставляю, например, с третьей колоночки. Всё. Вот. Теперь делаю шрифт какой-нибудь поменьше и растаскиваю. И потом минимизирую высоту колонки. У меня получается каждый абзац в отдельной строке.

То есть вот так можно вводить без всяких проблем, очень просто в Excel тексты. Вот я могу взять другую страницу. Беру вот эту страницу, к примеру. Вот так. Ну, больше я не буду показывать, просто убираю. Вот. А эту страницу покажу, как ввести. Вот ставлю сюда и нажимаю вставить. Вот у нас эти абзацы все, они здесь есть. То есть сама вставка текста очень упрощённая.

**Подраздел 3.2: Ограничения и возможности метода**

Ну, вопрос возникает такой: это же ведь не весь текст, это ж только фрагменты текста и абзацы в качестве фрагментов текста. Ну, это как раз нам и требуется для нашей задачи.

Значит, смотрим теперь на этот текст. Вы теперь понимаете, как он сделан. Значит, здесь взяты работы, произведения не самих этих писателей.

**Раздел 4: Теоретические аспекты и приложения анализа текстов**

**Подраздел 4.1: Задачи, решаемые интеллектуальной обработкой текстов**

Значит, что позволяет, во-первых, давайте сейчас рассмотрим. Что позволяет сделать, какие задачи решить позволяет интеллектуальная обработка текстов? Я вам вот эту ссылочку сейчас бросал там в этот чат. А сейчас я у себя открою её. И вы тоже откройте. Или вообще-то не надо, вам видно на экране всё. Вот смотрите, что позволяет это делать, этот подход? Какие задачи решать?
Во-первых, можно сформировать обобщённые лингвистические образы классов. Их ещё называют семантические ядра. И могу вам сказать, что, скажем, если вы пишете статью, то семантическим ядром являются ключевые слова или частью семантического ядра. Вот.
Эти лингвистические образы классов, обобщённые, их можно сформировать на основе фрагментов или примеров текстов, относящихся к каким-то категориям, обобщающим классам. И эти тексты могут быть на любом языке. В системе Eidos есть единственное ограничение, ребят, в этом плане, что этот язык должен быть без иероглифов и, ну, в общем, короче, должен использовать или латиницу, или кириллицу. Так по-простому, если сказать. Вот. То есть иероглифы мы не можем обрабатывать и всякие тексты типа грузинского там, арабского тоже. Потому что они не входят в стандартный набор шрифтов, так скажем. Но всё, что входит в стандартный набор шрифтов, можно обрабатывать.

Второе. Можно количественно сравнить лингвистический образ конкретного человека или текста какого-то или описания объекта, процесса или явления с обобщёнными лингвистическими образами групп, то есть классов. То есть это задача идентификации, задача идентификации. То есть она в математической лингвистике называется задача атрибуции текстов. Ну, то есть определяется авторство текста. При этом сам этот текст может быть подписан автором, может быть не подписан, то есть быть анонимным, а может быть псевдонимным. То есть он может быть подписан, но другой фамилией, не того, кто писал этот текст. Причины могут самые различные быть. Ну, например, Владимир Ильич Ленин, Ульянов. Когда была стачка подавлена жестоким образом на реке Лена рабочих и крестьян, ну, она была расстреляна просто. Вот. Много жертв было. То он написал такую очень талантливую статью, публицистическую, резко критикующую правительство, и вот всё это вот в связи с тем, как они заботятся о нуждах народа. Почему-то люди вообразили себе, что они должны о них заботиться, правительство. Ну ладно, значит, это я так подшучиваю немножко горько по этому поводу. Так вот, и он подписал эту статью очень благоразумно, не написал там Владимир Ульянов, он написал там Владимир Ленин. И он этот псевдоним взял Ленин именно в связи с этой историей трагической, с этой стачкой. Вот. И до этого он был малоизвестен, но был широко известен в узких кругах своих единомышленников. А тут он стал известен на весь мир, наверное, я бы сказал. Ну, на Россию точно. Российская империя тогда была в два раза больше, чем сейчас по территории, по населению там, в общем. По населению примерно как сейчас была, но это давно ж было. Вот, а по территории в два раза больше, можно сказать, была. И вот, включая Финляндию, Польшу, там, короче говоря, пошире было. Ну и эти вот республики, которые сейчас Украина, страны там, Казахстан, всё это входило в состав России.
Вот. И, значит, он получилось такое пиар получилось для него, как сейчас сказали бы. То есть он стал широко известным публицистом, политиком вот после этой статьи именно. И все стали говорить, то есть на слуху было это слово. А кто написал? А кто вот это? Это Ленин, это Ленин. И вот все это вот это Ленин, оно к нему прилипло. И сейчас уже никто и не, может быть, и многие не знают из молодёжи, что это вообще-то Ленин - это Ульянов. Вот. Ну, то есть Ленин к нему это пристало основательно.

Вот. Дальше, то есть мы можем решать задачу атрибуции текстов, то есть определять авторство текста, истинное авторство текста. Вот. Или пытаться понять, на какого автора этот текст похож. На текст какого автора этот текст похож. Может быть, это не этот автор написал, но кто-то его пародировал там или пытался подражать как-то. Ну, эта система это определит.
Также можно сравнивать обобщённые лингвистические образы классов друг с другом и создавать их кластеры и конструкты. Исследовать моделируемую предметную область путём исследования её лингвистической системно-когнитивной модели. Проводить интеллектуальную атрибуцию текстов, то, что я уже сказал, это определение авторства. То есть определять вероятное авторство анонимных и псевдонимных текстов. А кроме того, слушайте внимательно, можно осуществлять датировку текстов, определять жанр этих текстов и смысловую направленность содержания. И всё это можно делать для любого естественного, искусственного языка или для любой системы кодирования. То есть мы можем взять фрагменты исходных текстов программ на различных языках программирования и точно так же создать их лингвистические образы этих программ на разных языках программирования. И выяснится, что определённые команды являются общими для разных языков программирования, ну такие как print, например, for next, там do while. А многие команды являются специфическими для конкретных языков. И в некоторых языках они есть, а в других нет. В других языках, скорее всего, есть команды, которые или функции, которые имеют сходное функциональное назначение, но они могут отличаться по написанию.

**Подраздел 4.2: Пример анализа текстов писателей**

И вот здесь вот у нас задача стоит какая в этой простенькой работе? У нас задача стоит создать обобщённые лингвистические образы различных писателей, классиков русской литературы XIX века и XX, начала XX. Значит, это Достоевский, Гоголь, Горький, Грибоедов, Лермонтов, Пушкин, безусловно, Тургенев, Толстой. Ну и всё. Значит, я как эти откуда взял эти тексты? Я просто нашёл школьный сайт школьных сочинений по литературе и оттуда вот так вот, как я сейчас вам показал, сюда эти тексты ввёл. Но при этом, значит, нужно иметь в виду один маленький момент. Маленький, но существенный. Перед тем, как вводить тексты в Excel, нужно обязательно выключить переносы в Ворде. Потому что иначе с этими переносами сюда слова и попадут. Един- чёрточка-ства, там, вот такие вот получатся у вас дела. Значит, это нежелательно. Поэтому нужно, конечно, убрать эти переносы, а потом, значит, уже вводить.

**Подраздел 4.3: Лемматизация (теория)**

Теперь смотрите, значит, у нас, соответственно, несколько абзацев текста о Достоевском. Это причём не сами их тексты, а это о них тексты, сочинения. Несколько абзацев о Гоголе и так далее. И как я вот их вводил, я вам показал. А вот здесь вот, смотрите, я пишу два раза Достоевский, вернее, пишу один раз Достоевский, потом копирую и вставляю сюда. А потом беру и протягиваю столько раз, сколько, столько абзацев его текста. А здесь то же самое делаю, пишу Достоевский 1, Достоевский 2, Достоевский 3. Значит, но я написал Достоевский 1, а вот это всё остальное я уже не писал. Я просто взял его вот так вот, протянул, вот, и он перенумеровался. Это Excel делает. Поэтому создать такую форму нетрудоёмко. Вот. Значит, в одной ячейке Excel можно разместить текст 32 КБ. Это примерно шесть страниц текста такого вот, который стандарт А4.

Значит, какая у нас, как у нас будет решаться задача наша? Сейчас смотрите. Значит, мы на основе вот этих фрагментов текстов, значит, здесь это у нас классификационная шкала вот эта. И мы, значит, вот эти фрагменты текста, где классом является Достоевский, это класс, классификационная шкала, значит, значение - это классы. Вот. Они будут использованы для формирования обобщённого образа лингвистического этого писателя. Также точно об остальных писателях. А потом мы проведём идентификацию текстов, атрибуцию. И у нас получится, что какие-то тексты идентифицируются с авторами, а другие, может быть, и не совсем. Ну, в общем, мы увидим.

Теперь я хочу вам вот что ещё сказать. Очень важный момент, ребята, слушайте сейчас внимательно. Значит, э человек пользуется различными моделями различных уровней формализации, различной степени формализации. И самый низкий уровень формализации - это интуитивные модели. Это настолько низкий уровень формализации, что вообще нет никакой формализации. То есть мы не знаем, в какой форме и где хранятся интуитивные модели. Ну, обычно думают, что где-то связано с головой это, но это есть большое сомнение по этому поводу. А у меня вообще есть уверенность, что это не так. В связи с тем, что есть э нетелесный опыт, out of body, есть опыт клинической смерти, когда очевидно совершенно, что человек мыслит не с помощью мозга, и вообще, когда у него тело не работает или вообще отсутствует, он может прекрасно мыслить, видеть, чувствовать всё. И, может быть, даже лучше, чем сейчас, потому что это тело, оно даже иногда усложняет этот процесс, является таким мутным стеклом, через которое мы смотрим на мир, грубо говоря. А без мутного стекла получается чётче всё видно.

Вот. Ну сейчас я не буду это детализировать, но, в общем, э скажу так, что никто не знает, где эти интуитивные модели, в какой форме представлены физически. То есть форма представления знаний в интуитивных моделях неизвестна. И локализация их неизвестна, не только в пространстве, но и во времени, кстати. Но люди об этом не подозревают, что они и во времени ошибаются. Интуитивные модели могут относиться, например, их структура, материальная, которая их поддерживает, она может вообще не относиться к нашему времени. Она может относиться к будущему, например. Моё мнение такое, что, наверное, она локализуется в будущем. Ну такие представления - это требует комментариев, чтобы это объяснять, что это такое. Ну сейчас я про это не буду дальше говорить. Скажу только, что у меня есть работа об этом, и можно их там почитать, и какое-то представление возникнет. Но не всё я там, конечно, в этих работах могу описать. По простой, по одной очень простой причине. Это трудно, ну просто трудоёмкость большая. И поэтому получается очень тезисно, но даже тезисно это получается очень большие объёмы.

Вот. Следующий уровень формализации - это вербализация, то есть выражение в словах. То есть когда мы что-то описываем словами, то это уже некая начальная, самая начальная уровень формализации. Для чего мы это делаем? Ну, чаще всего для того, чтобы это сохранить и передать кому-то в пространстве и времени. То есть когда люди освоили вербализацию, то есть вообще речь появилась у них, то это дало очень многое. То есть мы смогли идентифицировать объекты после этого не только как относящиеся к каким-то конкретным классам. Ну то есть вот, допустим, я кого-то вижу. Вот сейчас я, ну не знаю, например, говорить или нет. Ну вот, допустим, кого я вижу? Я вижу сейчас только изображение Александры. Вот. И вот я её вижу и идентифицирую, что это Александра. То есть я идентифицирую это изображение с конкретным человеком. Вот. А когда у нас появляется язык, то мы можем тогда использовать обобщающие категории, то есть не только конкретные категории называть объекты и явления и процессы, но и обобщённые. Что принципиально важно? Я говорю так: это студентка. А кто тут у нас ещё студентки? А больше нету. Вот. А ещё есть у нас студенты, но они у нас тут инкогнито, фотографии свои не разместили. Ну, я имею в виду Артёма, Вячеслава, Павла, Сергея. Вот. То есть, видите, категория студент является обобщённой. Это не конкретный студент, это вообще студент. И когда я вижу, допустим, э какого-то конкретного студента, там, Артёма, например, то я понимаю, что это конкретный студент Артём. Это можно было и без использования языка идентифицировать. То есть для этого язык не требуется, слова. Вот. А вот когда я говорю, что это студент, то вот этот уровень понимания, когда я его отношу к какой-то обобщающей категории, он возможен только тогда, когда мы освоили вербализацию, и у нас в языке есть соответствующие слова, термины для обозначения обобщающих категорий. То есть роль языка огромна. Человек начал обобщать после того, как у него появился язык, он стал обобщать. То есть у него на основе конкретных примеров стали появляться обобщённые образы.

Вот. А дальше слушайте очень интересный момент. Значит, мы можем решать задачи. Значит, для чего мы вербализуем? Для того, чтобы сохранить по времени. Когда появился язык, то мы научились пересказывать. Есть народы, у которых письменности нет, примитивные, ну там где-нибудь в Амазонке там или в Папуа Новой Гвинее, когда вот Миклуха-Маклай там исследовал. У них не было письменности. И они просто пересказывали свою историю новым поколениям. При этом у них был достигнут очень высокий уровень качества этого пересказа. То есть они, их считалось идеальным, когда ни одного слова не добавлено, не убавлено. То есть вот прямо вот тождественно пересказ. Это было идеал. Вот. Им это удавалось во многом достичь этого. То есть это даже очень странно, насколько хорошо им это удавалось сделать. И Миклуха-Маклай как раз вот и записывал эти вот их легенды, которые были, в общем, дословными пересказами того, что было сказано там 100 лет назад, 200 лет назад кем-то, каким-то их историком, там, жрецом каким-то этого племени.
Вот. Когда же появилась письменность, то появилась возможность записи, да, вербализованных моделей. И тогда появились вот книги, ну сначала какие-то там на камнях вырезали, скрижали там, делали таблички глиняные, потом папирусы, потом на бумагу придумали китайцы. Вот. Ну, в общем, короче говоря, стали передавать эти слова из поколения в поколение. И сейчас мы можем прочитать то, что было написано 400 лет назад, там, может быть, и тысячи лет назад. Ну, скажем, Библия, там, священные эти какие-то писания, тексты, они имеют древнее происхождение. Вот. И больше даже, потому что на камнях надписи есть у тех гробниц есть надписи египетских. То есть там 3.000 лет, 4.000 лет где-то есть тексты, которым такое такая датировка этих текстов. И вот это стало возможным именно благодаря вербализации.

Теперь вербализация представляет собой один из самых универсальных видов моделей. То есть мы можем описать, ну, практически всё, что угодно из нашей области представлений, из той области, где у нас есть система конструктов. Вы помните, да, про конструкты я вам рассказывал? Это очень важное понятие. Это связано с размерностью когнитивного пространства и его объёмом. Вот если у нас есть соответствующие понятия, то мы можем описать явление в этих понятиях. Если же их нет, то мы тогда, значит, говорим, ни словами сказать, ни пером описать. То есть вообще невозможно это как-то описать. Вот. Так вот, чаще всего удаётся описать э более-менее процессы и явления. Так вот, когда мы их описываем, то мы уже даём им такой уровень формализации. Дальше слушайте внимательно. То уже такой уровень формализации у нас появляется. Э-э, уже, во-первых, это модели. То есть лингвистические модели, вербализация - это модели. Это модель мира, который мы воспринимаем, осознаём, внутреннего и внешнего. И там есть, соответственно, в языке соответствующие конструкции: существительные, глаголы, прилагательные. То есть есть описание, есть сам объект, его идентификация, есть его динамика этого объекта, есть свойства этого объекта. В языке есть конструкции, позволяющие всё это описать: и сам объект идентифицировать термином каким-то, и э-э сказать, как он движется, что с ним происходит. Ну, допустим, автомобиль, вот тот красный автомобиль сейчас вот движется по улице Ставропольской. То есть автомобиль - это существительное, движется - глагол, то есть что с ним происходит, да? Вот. И какой он? Красный автомобиль. По-моему, это Mazda. Вот. Марка Mazda, а модель я не пойму какая. По-моему, тройка. Вот он раз, проехал, я так. По-моему, тройка Mazda. Вот, короче говоря, мы, используя различные конструкции языка, лингвистические, грамматические, вот эти конструкции, слова сами и их способы взаимосвязи, которые называются грамматика в предложениях, мы можем описать процессы и явления.

Так вот, система Eidos, она позволяет такой уровень формализации использовать для дальнейшим для решения всех вот этих задач, которые я сейчас перечислял. То есть можно и обобщённые образы создавать, и сравнивать конкретные образы с обобщёнными. Ну сначала, конечно, конкретные образы создавать, потом их обобщать, потом сравнивать конкретные с обобщёнными, обобщённые друг с другом, формировать их кластеры и конструкты. И конструкты образуют парадигму реальности. То есть можно добавлять конструкты в нашу парадигму реальности, увеличивать размерность и объём нашего когнитивного пространства, то есть повышать уровень системности его и делать его, скажем, наши модели более адекватными для отражения более сложных явлений и процессов внешнего и внутреннего мира.

**Раздел 5: Анализ моделей и оценка достоверности**

**Подраздел 5.1: Сформированные модели (классы и признаки)**

Ну теперь давайте перейдём, посмотрим, что ж у нас там за классы сформировались. То есть будем выполнять вот эту инструкцию, которая здесь написана у нас для студентов, которую они никогда обычно не читают, а сразу кликают и идут дальше. Ну я особо их не ругаю, говорю: "Ну ладно, ребята, тогда будем сейчас э-э да, работа установлена. Тогда будем э просто делать то, что я покажу, сейчас вам скажу".
Вот режим 2.1. Мы здесь видим просто вот этих авторов в порядке алфавита, причём по первой букве, то есть не по фамилии, а по той букве, которая там была в ячейке первой. Ну, это здесь никакой роли не играет, в каком они порядке. Значит, но я могу вам сказать, что это те уникальные наименования, которые были в колоночке классификационной шкалы, классификационной шкале. Если бы нам нужно ещё было датировку текста, мы тогда здесь вот ещё бы добавили колоночку и написали бы там, э-э, ну, столетие, например, там, или какая первая половина, второго там XIX века, вторая половина XIX века, там, первая половина XX века, там, ну что-нибудь там, серебряный век, там, ну, в общем, написали бы там что-то. Вот. И третья, ну, можно вообще год написать, когда работа там, если здесь речь о конкретное произведение, то можно было бы год выпуска этого произведения написать. А потом можно было бы написать жанр. И там написать роман, там, повесть, э-э, стихи, там, такой исторический роман, там, или там ещё какой-то там. Ну, в общем, как-то жанр описать. И можно было бы направленность смысловую описать, э-э, основную идею этого романа, допустим, или стихов, фабулу, как бы сказать, или там аннотацию какую-то написать там, о чём речь там вообще. Комедия, там, ну да, вот жанр, может быть, комедия, кстати, вот Горе от ума - это вроде как комедия. Ну такая комедия грустная какая-то. Смех сквозь слёзы, как говорят. Горе от ума - это как раз и смех сквозь слёзы. Было бы смешно, если б не было грустно.

Это вот у нас, ребята, описательная шкала. В данном случае она одна. Вот, их тоже может быть несколько, сколько, ну, вообще-то много может быть. И здесь просто слова, которые в этих текстах были. Причём эти слова, ребята, здесь они в том виде, в каком они там были в тексте. Вот было с большой буквы, вот оно с большой буквы. Значит, рассортировано по алфавиту. Вот. Значит, здесь мы видим, смотрите, отношение, отношению. Значит, э-э поэма, поэмы, поэта, да? Вот. То есть есть слова, которые имеют общее происхождение. Простая, простой. Произведение, произведением, произведений, произведения. То есть склонение разное, да? Слово одно и то же, но разное склонение. Противоречивые, противоречивый. То есть иногда там род разный или число. Почему, почему я сейчас это говорю об этом? Значит, ранее, ранним, там, раннего, раскольников, раскольникова. Значит, потому что есть такое понятие лемматизация. Что такое лемматизация? Это приведение слова к первообразному слову, от которого оно возникло.

Система Eidos поддерживает лемматизацию. Сейчас здесь 767 у нас слов. Это довольно мало, если честно. Значит, там могут быть десятки тысяч слов. Система это поддерживает, то есть она способна обрабатывать книги прямо, не только вот такие фрагменты небольшие, а прямо файлы книг. И я это делал, и в этих статьях это написано, описано.

Ребята, сейчас я вам покажу про базу лемматизации, в двух словах расскажу. Значит, система Eidos использует базу лемматизации, содержащую около 2 млн слов, которая разработана академиком Залезняком. Это очень интересный, замечательный учёный, который недавно, ну как недавно, года два назад, по-моему, примерно или полтора скончался. Вот. Знаменитый очень учёный тем, что он расшифровывал новгородские рукописи. Замечательный был человек. И он проводил многоплановые работы в разных направлениях лингвистики. И, в частности, он разработал словарь лемматизации русского языка, состоящий из 2 млн слов, ребята. Можно эту базу скачать, она входит в состав полной инсталляции системы Eidos. Я нашёл её на Хабре, Хабре. Вот. И она представляла собой экселевский файл, в котором, значит, вы знаете, что в Экселе только миллион строк возможно, к сожалению. Это вообще-то маловато для современного уже нашего времени. Ну, может, они его расширят всё-таки адресацию. Но миллион - это мало. Я многократно сталкивался с тем, что этого не хватает. И вот эта база лемматизации, в частности, значит, она была в виде двух листов представлена. Значит, я сделал программу, которая из Экселя это всё вытащила и в нормальную базу запихнула. Вот эта нормальная база, она есть в системе Eidos, база лемматизации. Она довольно-таки здоровенная. Вот смотрите, сейчас я по размеру рассортирую. Она самый большой файл здесь. Она 182 МБ занимает. Это самый большой файл в системе Eidos. И здесь вот мы видим что? Мы видим справа лему, а слева словоформу. Лема - это первообразное слово. Вот, допустим, самовращение слово. А вот здесь мы видим самовращение и, самовращением, самовращению. Ты идёшь по коврёшь, он идёт по коврёт, мы идём по коврём, они идут по коврут. Ну, в общем, я так подшучиваю немножко. Ну, то есть склонение и словообразование. Есть игры соответствующие. Вот, допустим, давайте сейчас я вам э вот 2 млн, даже больше чуть-чуть. Вот. 97.000 слов. Что такое лемматизация? Мы находим в тексте вот такое слово, допустим, заменяем на такое. Вот. Находим такое, заменяем на такое. Это вот и есть лемматизация. То есть это замена слов на первообразные.

Сейчас мы, когда вводили эти данные, то я лемматизацию не включал, но она там есть. Можно было её включить. Вот здесь вот поставить птичку, и была бы тогда включена лемматизация.

Ну, слов тогда будет поменьше, но и информация будет поменьше, и будет более такая будет обобщённая информация в исходных данных. Вот. Так вот, вчера я видел передачу по телевизору, где на телеканале Спас, религиозный канал, где священники участвуют и, скажем так, люди, которые как бы являются учёными, но религиозными, ну, богословы, теологи по западной манере, если называть. То есть это это интерпретаторы, комментаторы священных текстов. Они иногда не имеют никакого сана. И сейчас у нас в перечень научных специальностей внесён внесена специальность теология. Почему-то она называется теология, а не богословие. Я не знаю, почему. Ну нам более научно звучит, может быть. Но теология, она переводится как богословие на русский язык. Это прямо сейчас ваковская специальность. То есть присваивается учёная степень кандидата теологии, доктора теологии. Наш вот митрополит Павел, который сейчас у нас после Исидора, он кандидат, по-моему, теологии, да, богословия.
Так вот, что я хочу сказать? Что там была передача посвящена тому, кому приписываются те или иные священные тексты и в самой, в псалмах, в Библии самой, и в различных апокрифах. И вот известно, допустим, вот этот текст приписывается, допустим, апостолу Павлу. А вот современные лингвисты его анализируют, говорят: "Да нет, что-то как-то не похоже, понимаете?" То есть или есть ряд особенностей, которые не характерны для апостола Павла. Вот если сравнивать с обобщённым лингвистическим образом текстов апостола Павла, ну вот к евреям, например, есть послание. Это о нём именно шла речь конкретно. И возникает, спорят об этом, вообще-то, авторстве. То есть как-то не похоже на апостола Павла, понимаете? Автор, по сути дела, неизвестен. А мог быть определён этот автор, если сравнить с другими текстами, возможно, мы бы нашли ещё тексты в том же в том же стиле и определили бы, что это текст такого-то автора. То есть это задача атрибуции текстов. Это задача интеллектуального анализа текстов.
И там выступал компетентный вполне человек, хотя он сказал, что он некомпетентный в самом начале. Я в этом вообще-то не разбираюсь. А потом стал рассказывать, и было видно, что, конечно, он хорошо разбирается в этом, в том, о чём говорит. Ну, видимо, он имел в виду, что он всё-таки не учёный-лингвист и не специалист в области интеллектуальных технологий и атрибуции текстов, понимаете? Вот. А так, в общем, он, конечно, безусловно, очень компетентно рассказывал. Мне понравилось вообще, как он рассказывал. Вот оттуда я даже сейчас, видите, упомянул его.
Вот. Так вот, что самое интересное, что система Eidos позволяет такие задачи решать. И все вот эти, ну, допустим, мы даже можем взять и описать словами какие-то места, ну, допустим, там было дерево. Помните этот фильм, когда они искали этот бриллиант, этот, нет, это Джентльмены удачи. И там Крамаров говорил: "Там было дерево, вот такое вот". И они искали, ездили на машине, на служебной, искали это дерево, потом в конце концов нашли это место, которое они искали. Там оказалось в окне цветочек такой типа пальмы. Вот это вот дерево, вот оно. Ну там Пушкин был, памятник Пушкину. Потом что, он стоит или сидит Пушкин? Он говорит: "Он так ты вообще, понимаешь, что ты спрашиваешь? Ну то, что посадят, он же давным-давно уже умер. Ну что его посадят?" В общем, там вот этот вот такие моменты были. И вот тоже это связано с лингвистикой, юмор этот. Так вот, я хотел вам сейчас что сказать интересное, что в системе Eidos можно задать описание различных мест словами, я подчёркиваю, не фотографиями. Можно и фотографиями, кстати, вводить и анализировать. А можно просто словами описать, что там было. И она сравнит различные описания, понимаете? Она сравнит эти различные описания и найдёт место, которое по описанию больше всего подходит. То есть она будет идентифицировать объекты по описаниям. Вот. Можно, допустим, статьи административного уголовного кодекса ввести, и она будет определять, какая, по какой статье проходит то или иное преступление, если дать его описание этого преступления, фабулу. Как говорится, коротко описать, в чём его суть. Она находит статьи, например, соответствующие, понимаете? То есть именно потому, что там сходные есть фразы, слова и сочетания слов. Кстати, нечто подобное делает система антиплагиат, проверка на оригинальность. Но она меньше функций выполняет. Система Eidos, она выполняет интеллектуальные функции и очень много различных выходных форм. Она такая более специализированная. Но некоторые моменты там, в общем-то, присутствуют: выделение сходства, там, фрагменты ищет и так далее.

**Подраздел 5.2: Оценка достоверности моделей (F-мера и L2-мера)**

Ну, дальше, слушайте интересный момент. Значит, э-э обучающая выборка. Вот смотрите, значит, у нас вот здесь вот уже побольше этих вот кодов. Это каждый код соответствует слову какому-то в справочнике слов. Вот. И все фрагменты текста у нас закодированы. Дальше мы создаём модель. Причём мы что делаем? Вот здесь вот указываем 30. Ну, она показывает нам система, что эта модель составляет 80% по размерности от того, что максимально могло бы обрабатываться в системе Eidos при тех ограничениях на размеры баз данных, которые обусловлены языком программирования. И в системе Eidos есть критерий достоверности модели. И собственный есть, э-э, то есть есть критерии собственной уверенности системы в том или ином решении. Это уровень сходства, интегральный критерий, аддитивный уровень сходства объекта с классом. И вот этот уровень сходства может быть выше или ниже. И максимальный, конечно, он 100%, минимальный - это -100%, максимально отличается от образа класса. И вот модуль этого уровня сходства является адекватным критерием степени уверенности системы в этом решении, потому что мы потом, я вам покажу, есть режимы, позволяющие исследовать зависимость доли истинных решений среди всех решений в зависимости от уровня сходства. Оказывается, чем выше уровень сходства, тем больше доля истинных решений, как положительных, так и отрицательных. То есть как о принадлежности объекта к классу, так и о непринадлежности. Это означает, что уровень сходства является адекватным критерием достоверности решения, степени истинности решения.
Вот. И вот система имеет возможность вот здесь вот внизу, это как раз показано, удалять из результатов идентификации те, которые, по-видимому, недостоверные, в соответствии с этим критерием достоверности, который является адекватным. Это доказано в очень большом числе различных применений, подтверждено это, что это очень хороший и адекватный критерий достоверности.

Значит, мы осуществили синтез и верификацию моделей. То есть синтез - это созданы эти модели, а верификация определила их достоверность. Смотрим на эти модели. То есть я сейчас полностью иду, следую в соответствии с хелпом, который у нас есть в режиме 1.3. Вот эти пункты, просто я их по очереди выполняю, ребята. Это красная линия такая исследования. Можно отступать от этой линии вправо и влево, вверх-вниз, вперёд-назад, сколько позволяет вам размерность вашего когнитивного пространства. Но некоторые операции можно выполнять только после того, как будут выполнены некоторые предыдущие. Ну, скажем, анализ можно или решение задач можно осуществлять только тогда, когда модели созданы. Модели можно создать только тогда, когда данные введены. Определить достоверность модели тоже можно только тогда, когда они созданы. Ну и так далее.
Так вот, мы смотрим теперь на достоверность, на сами модели смотрим, что ж там получилось у нас. Значит, у нас три модели статистических. Это матрица корреляционная, её же называют часто матрица сопряжённости. Я называю матрица абсолютных частот. Почему? Потому что, чтобы отделить её от матрицы относительных частот, матрицы условных и безусловных процентных распределений. Это матрица абсолютных частот АС, матрица условных безусловных процентных распределений ПНЦ1 и ПНЦ2. И системно-когнитивные модели. Про то, как рассчитываются модели, я вам рассказывал, да, ребята? Про частные критерии. Вот сейчас мне дайте подтверждение. Помните вы что-нибудь такое? Про хи-квадрат, про меру Харкевича, про информ. Да, да, рассказывали. Вот. Ну довольно подробно, да, я наверное, объяснял, как они взаимосвязаны друг с другом там, что можно сравнивать путём вычитания, путём деления, да, вот всё это я рассказывал, да? Да. Да. Вот. Ну сейчас мы просто на сами модели посмотрим. Значит, смотрите, это модели довольно большой размерности уже. Ну, классов не очень много, восемь. Но зато вот слов-то довольно много уже, 767. И мы видим, какие слова встречались у каких авторов и сколько раз. Вот, допустим, слово стихотворение встречалось три раза в сочинении школьном о Лермонтове, Лермонтове. И стихотворение и тоже три раза встречалось в этом же классе. И вот так вот мы видим. Ну здесь статистика, скажем так, ну, в общем, много пустых мест, короче говоря, в матрице, потому что мало, маленькие примеры очень, мало примеров.
Вот. Ну и хи-квадрат, замечательный критерий, частный, очень хорошие модели получаются, они очень хорошо сбалансированы. Сумма равна нулю и по колонке, и по строке. Это прямо как магический квадрат какой-то. Но здесь вот, смотрите, у нас только там могут быть разные значения, а здесь именно ноль. Значит, теперь смотрите, ребята, значит, у нас здесь вот нули написаны, видите? А здесь ничего не написано. Почему? Потому что в системе расчёты внутренние ведутся с семью знаками после запятой. Ну, вернее так, они ведутся там гораздо большим числом знаков, но в базу данных заносятся именно с семью знаками. Вот. А сами расчёты, конечно, там идут 23 знака, по-моему, после запятой. И вот дальше смотрите, значит, кое-где мы видим нули. Эти нули, значит, система сделана как, что когда вообще нет ничего там или ноль, то она просто выводит пустое место, ну, пробел, грубо говоря. Ну там даже и пробела нет. А когда где-то там в седьмом знаке там единичка какая-нибудь выскочит, то она тогда вот пишет уже число, но здесь при этом форматирование, здесь три знака после запятой показано на экране, уже это всё-таки показано некое число. Ну это, ну, в данном случае ноль.

Смотрим на достоверность моделей. Значит, вот у нас есть несколько разных интегральных критериев оценки достоверности. Про них я рассказывал, ребят, скажите, пожалуйста, мне. Про них нет. Нет? Да вы что? Не может быть. Если я кластерный анализ вам показывал, значит, должен был про критерии достоверности тоже рассказывать. F-мера достоверности Ван Рисбергена, её обобщение. Что, неужели я не рассказывал про это? Что Ван Рисберген предложил рассчитывать достоверность модели, оценивать, вернее, её по тому, как она решает задачу идентификации. Значит, ну есть и другие задачи, которые решаются с помощью модели, но с их помощью труднее оценивать достоверность модели. Проще всего это делать путём решения задачи идентификации. Какие ещё задачи есть? Прогнозирование, принятие решений. Вот. Значит, если мы, ну есть возможность исследовать достоверность модели на ретроспективных данных, тогда можно использовать и задачу идентификации, и задачу прогнозирования, потому что известно задним числом, что там произошло, да, на фондовом рынке за последний год и более длительные сроки. А вот принятие решений ретроспективные данные не позволяют исследовать. Почему? Потому что если мы сейчас, допустим, примем то или иное решение, то мы можем только через некоторое время узнать, к какому результату оно привело. Но если у нас где-то будут базы данных, где будет написано, какие решения были приняты, и что потом получилось, то можно и задачу принятия решений использовать для оценки достоверности модели. Но это всё, как говорится, более сложно, а самый простой вариант - это такой естественный - это вариант использования задачи идентификации для оценки достоверности. То есть что должна делать модель? Она должна правильно относить объект к тем классам, к которым они относятся, и правильно не относить к тем, к которым они не относятся. И она может и ошибаться и в первом, и во втором случае. Рисберген предложил считать, что если уровень сходства объекта с классом, интегральный критерий, больше нуля, тогда это положительное решение, если меньше нуля, тогда отрицательное. Если решение истинное, то это тру, ложное - false. Соответственно, он предложил четыре сумматора ввести: тру позитив, тру негатив, false позитив, false негатив. И если система относит объект к классу, к которому он на самом деле относится, то это тру позитив. Если не относит объект к классу, к которому он на самом деле не относится, то это тру негатив. Если она относит объект к классу, а на самом деле он к этому классу не относится, то это false позитив, это ложное срабатывание. Очень опасная ошибка, ребята. Вот, скажем, летит самолёт, наш самолёт. И наши средства ПВО пытаются идентифицировать наш это самолёт или нет. А там транспондер не работает по какой-то причине, ну, повреждён, например. Вот. И система идентифицирует, что это не наш самолёт, и сбивает его, добивает, грубо говоря. Из-за того, что ошибка идентификации. Понимаете? Ложноположительное решение. То есть она считает, что это противник и сбивает. Вот. А это не противник. Поняли, да, ребята? О чём речь идёт? Вот. И ложноотрицательное решение. Система не относит объект к классу, потому что уровень сходства ниже, чем, меньше нуля. А фактически на самом деле он к нему относится. Ясно дело, что когда какой-то объект идентифицируется, то какой-то из этих сумматоров увеличится на единицу. В результате все, если их все сложить эти сумматоры, то получится число объектов распознаваемой выборки или тестовой ещё называют в нейронных сетях.
Так, вспоминаете, ребята, кто-нибудь вспоминает, рассказывал про это или нет? Ну, наверное, не вам рассказывал. Дежавю. Значит, ну вам, наверное, тоже, потому что кластерный анализ я ж вам показывал, а его я показываю после того, как расскажу про модели, достоверность модели, решение задачи идентификации, прогнозирования, принятия решений, потом только про кластерный анализ. Понимаете? То есть похоже, что что-то вы там пропустили всё-таки. Ну ладно, коротко расскажу.
Так вот, на основе этих сумматоров Рисберген предложил формулы для полноты модели, точности модели и достоверности модели. Называется это мера F-мера Ван Рисбергена. Полным-полно литературы про неё в интернете. Но я сейчас вам сошлюсь на свою статью. Потому что сейчас дальше пойдёт речь именно о том, что там описано. Значит, эта статья называется "Инвариантное относительно объёмов данных нечёткое мультиклассовое обобщение F-меры Ван Рисбергена достоверности модели Ван Рисбергена в АСК-анализе в системе Eidos". Почему я эту статью написал? Про идентификацию рассказывали. А про достоверность нет, да?
Значит, слушайте внимательно. Значит, дело в том, что если мы посмотрим на результаты идентификации, то мы увидим, что да, система ошибается, но она ошибается слабо, мало. Вот смотрите, допустим, фрагмент текста Достоевский 1 очень похож на обобщённый образ Достоевского и чуть-чуть похож на обобщённый образ Горького. Видите, да, на 3%? Что делает Рисберген? Он берёт, суммирует единичку к истинно положительному решению, что, так сказать, вполне адекватно, потому что уровень сходства очень высокий. Но потом он делает странную вещь такую, которую я не одобряю. Он берёт и суммирует единичку к ложноположительным решениям, потому что система отнесла фрагмент текста Достоевский 1 к классу Горький, но она сама сильно сомневается в том, что он относится к этому классу, и уровень уверенности всего 3,5%. То есть она сама очень сильно сомневается в том, что этот фрагмент текста относится к классу Достоевский, то есть к классу Горький. Вот. Но он всё-таки уровень сходства больше нуля. То есть это ложноположительное решение. Ван Рисберген берёт и суммирует единичку к ложноположительным решениям. Я говорю: "Ну я говорю, вообще-то не согласен. Я говорю, против. Я против этого. Я считаю, что более справедливо было бы просуммировать в данном случае 0,82 там и всё остальное, все знаки остальные, которые там есть, их там я сказал, семь после запятой, просуммировать к истинно положительным решениям, не единичку, а 0,82. И просуммировать 0,30, 0,036 и так далее, то есть 3 к сумматору false positive. Понимаете? А вот здесь вот я бы просуммировал что? По критерию сумма знаний. Я бы просуммировал Толстого, ещё похож Достоевский 1 на Толстого. Ну он-то похож на 1,8%. Я просуммировал бы 18 тысячных к сумматору false positive по этому интегральному критерию.
Понятно? Это называется нечёткое обобщение F-меры Ван Рисбергена. А что такое мультиклассовое? Дело в том, что Рисберген сам предполагал, что один объект может относиться только к одному классу. По-моему, это очень такое сильное обобщение, упрощение. Значит, я считаю, что любой объект может относиться к очень большому числу классов. Вот. Ну, допустим, если предприятие работает, то у него есть определённая деятельность, и есть результаты этой деятельности. Результаты есть в натуральном выражении и в стоимостном выражении. В натуральном - это количество, качество продукции, а в стоимостном - это прибыль и рентабельность. Причём это очень упрощённо я рассказываю. Может быть очень много показателей качества продукции, ну десятки буквально, ну там штук пять точно. То же самое может быть, кроме прибыли и рентабельности, там капитализация, там прибыль и рентабельность за определённый период, с определённого подразделения там или поля там и так далее, и так далее. То есть может быть очень много классов, к которым относится объект одновременно. Поэтому необходимое мультиклассовое обобщение F-меры Ван Рисбергена. Значит, я предложил нечёткое мультиклассовое обобщение меры Ван Рисбергена. Вот. Но мало того, дело в том, что здесь же абсолютные величины, эти сумматоры - это просто суммы абсолютных величин различных вариантов исходов идентификации. А при увеличении объёма выборки они растут линейно эти все сумматоры. Ну одни быстрее, другие медленнее. И, конечно, иногда отклоняются от линейного, чисто линейного вида графики эти. Но, в принципе, они растут линейно. Вот. Вот если, допустим, это у нас будет объём выборки, ну, скажем так, монотонно они растут. То есть с увеличением объёма выборки увеличивается и значение этих сумматоров.
Вот. И вот берём, значит, я вам могу сказать из опыта, что быстрее всего растёт сумматор истинно отрицательных решений. То есть это проще всего системе получить истинно отрицательное решение. Она очень хорошо определяет, что объекты не относятся к какому-то классу. Это очень высокая достоверность таких решений. Это это во всех моделях, это я сейчас обобщаю на большое число моделей, которые я видел. То есть истинно отрицательные решения растут, их число растёт быстрее всего. А вот истинно положительных помедленнее растёт. Это когда хорошая модель. Когда модель плохая, тогда там совсем наоборот всё. Вот. Потом растёт false positive, ложноотрицательные решения. И медленнее всего, как вы, наверное, догадываетесь, ложно, ложноотрицательные решения растут. Ложноположительные растут медленнее, то есть быстрее, чем ложноотрицательные. Ложноотрицательных решений вообще практически нет. Можно так сказать, их очень мало. Ну, обычно, обычно их очень мало.
Вот. То есть эта картина приводит к вопросу, создаётся, возникает вопрос о том, а как ведёт себя точность, полнота и достоверность модели в этих условиях, когда вот эти слагаемые, сумматоры, которые используются в формулах при расчётах, вот эти вот формулы, э-э, recall, precision, там, F-мера, то есть полнота, точность и достоверность. Как будут вести себя вот эти вот выражения математические при увеличении объёма выборки? И я исследовал этот вопрос, скорости сходимости. Оказалось, что она довольно плохая эта сходимость. То есть F-мера Ван Рисбергена классическая, она плохо сходится к некоторому пределу при увеличении объёма выборки. Ну, примерно около 3.000 наблюдений нужно. Это нехорошо, понимаете? Обычно столько нет. Обычно там, ну, 200, 300, 400, там, вот такое какое-то количество, но не 3.000. Это редко бывает. Вот если проводить исследования психологические, там, лингвистические, там обычно вот эти вот связанные с трудоустройством и тому подобное. Я сталкивался с тем, что обычно там в пределах тысячи наблюдений. Вот. Но 3.000 - это уже редкость.
Вот. Поэтому необходимо обобщение. И я эти обобщения разработал, описал вот здесь. Да, значит, поставил эти проблемы: проблема мультиклассовости, проблема нечёткости и проблема зависимости F-меры от объёма данных. Она, оказывается, зависима от объёма данных. Что вообще очень нехорошо. Ну это примерно, как вот мы говорим, допустим, о относительных частотах и говорим, что относительная частота стремится к вероятности при неограниченном увеличении объёма выборки. А потом спрашиваем: а стремится-то она при неограниченном, а вот всегда ж она ограничена реально. А насколько быстро она к нему стремится к этому вот пределу, асимптотически приближается? И оказывается, что она, допустим, там нужно 100.000, например, наблюдений, чтобы она приблизилась на 5% погрешность, чтобы была. Да ничего себе! А может быть, нужно 20 наблюдений, она приблизится. То есть насколько быстро сходимость вот, понимаете? Вот. И я исследовал этот вопрос, скорости сходимости. Оказалось, что она довольно плохая эта сходимость. То есть F-мера Ван Рисбергена классическая, она плохо сходится к некоторому пределу при увеличении объёма выборки. Ну, примерно около 3.000 наблюдений нужно. Это нехорошо, понимаете? Обычно столько нет. Обычно там, ну, 200, 300, 400, там, вот такое какое-то количество, но не 3.000. Это редко бывает. Вот если проводить исследования психологические, там, лингвистические, там обычно вот эти вот связанные с трудоустройством и тому подобное. Я сталкивался с тем, что обычно там в пределах тысячи наблюдений. Вот. Но 3.000 - это уже редкость.
Вот. Поэтому необходимо обобщение. И я эти обобщения разработал, описал вот здесь. Вот. А потом их решил: решение проблемы мультиклассовости, решение проблемы учёта нечёткости, асимптотическое выполнение для меры, которую я предложил, принципа соответствия с F-мерой Ван Рисбергена. Вот. То есть принцип соответствия выполняется. То есть мера, которую я предложил, переходит в меру Ван Рисбергена, когда у нас все решения либо 100, либо -100, когда у нас один объект относится к одному классу и так далее. Вот. А потом взял и разработал инвариантное относительно объёмов данных обобщение чёткой мультиклассовой меры, которую только что предложил. Назвал её L2 мера. И потом привёл численный пример, который мы сейчас и посмотрим. Вот, я не знаю, получится? А, ну сейчас, наверное, вот так сделаем. О! Мы видим, ребята, что при увеличении объёма выборки классическая мера F-мера Ван Рисбергена, которая жёлтеньким здесь обозначена, она очень сильно колеблется, и получается типа затухающего колебания. То есть она очень неустойчиво себя ведёт при малых объёмах выборки. То есть вообще ей нельзя верить фактически, если меньше там 500 объектов наблюдения, то вообще верить ей нельзя совершенно, потому что отклоняется на при максимуме единица и минимуме ноль, но отклоняется на 2/10. То есть получается то 0, то 0,75, то 0,57, понимаете, достоверность? Ну то есть это вообще, ну То есть она вообще не работает, я бы сказал так. Это термометр, который говорит так: либо там у вас 25°, либо у вас там 45°. Вы понимаете, ну есть же разница, 25 и 45, правильно? Либо, значит, там, либо там у него 20, 30, там 16° температура, либо у него 46. Понимаете? На 20 разница. Вот. Либо, значит, там, либо у него 25°, 26°, либо 46°. То есть либо он уже охлаждается, либо он вообще перегрелся, понимаете? То есть это ж ну такой термометр, он нам не очень-то нужен вообще-то, честно говоря. Зачем он? Он не позволяет температуру определить. Вот. Так вот, нечёткое мультиклассовое обобщение меры Ван Рисбергена, оно, во-первых, выше идёт на 1/10 примерно. Почему? Потому что более справедливо оценивает ошибки. То есть вот когда у нас, мы же знаем, это вот как раз подтверждает то, что я говорил. Мы знаем, что уровень сходства объекта с классом является адекватной мерой достоверности. Поэтому из этого следует что? Что истинные решения обычно имеют высокий уровень сходства, а ложные - низкий. Поэтому получается, что если мы будем суммировать э к истинным э к сумматорам тру позитив, тру негатив, то там обычно суммируется большая величина. А когда мы суммируем э false positive и false negative, там обычно суммируется маленькая величина. Вот это вот и означает, что уровень сходства является адекватной мерой достоверности решения того или иного. Поэтому получается, что мера F-мера Ван Рисбергена классическая является заниженной. То есть она даёт заниженную оценку достоверности модели. А вот эта, которую я предложил, мультиклассовое нечёткое обобщение, именно за счёт нечёткости в данном случае, получается гораздо более адекватная оценка достоверности. И она более высокая. И эта мера, мультиклассовая нечёткое обобщение, она довольно-таки имеет высокую устойчивость при низких э объёмах выборки. При малых объёмах выборки. И я вам скажу, что недавно систему Eidos исследовали в Пермском национальном университете, где я сейчас недавно провёл несколько занятий. Вот, вот здесь вот. Шесть пар провёл с ними на курсах повышения квалификации. Вот. И они там её исследовали под увеличительным стеклом с пристрастием. И у них получился очень интересный результат, который они мне сказали. Для меня это было некая новостью. Она оказывается быстрее в пять раз обучается, чем сети э с свёрточной сети с глубоким глубокого обучения. В пять раз быстрее она обучается. То есть у неё достоверность модели определённая достигается в пять раз быстрее на объёмах выборки в пять раз меньше, чем в этих сетях, которые сейчас очень популярны.
Вот. А вот синенькая - это э инвариантная относительно объёмов данных, нечёткая мультиклассовое обобщение классической меры Ван Рисбергена. Смотрите, она даёт нам критерий для оценки достоверности, который выше на 3/10, чем оценка классическая. Вот у нас тут 0,39, грубо говоря, а здесь 0,79. Понимаете? Значит, на здесь 0,39, ну 4, а здесь 0,7. На 3/10 выше. Но это вообще-то не шутки, понимаете? Либо там у нас 0,4 достоверность, либо 0,7. 0,4 вообще задачу нельзя решать. Если 0,7, то она уже так более-менее, ну на три балла, грубо говоря, достоверность. А вы знаете, какую я исследовал выборку здесь? Случайную. Я для формирования выборки использовал операторы языка программирования, которые генерируют псевдослучайные последовательности. А шум в среднем должен быть равен нулю, кстати. И матрица абсолютных частот должна равна нулю быть, и все вот эти коэффициенты во всех этих моделях, э системно-когнитивных, они должны стремиться к нулю при увеличении объёма выборки.

**Раздел 6: Заключение и дальнейшие шаги**

Вот, ребята, у нас заканчивается занятие. А у нас ещё куча вопросов осталась нерассмотренных. Мы только с вами начали рассматривать обработку текстов. Но это из-за того, что я отвлёкся на оценку достоверности модели. Теперь смотрим, у нас здесь вот эти критерии классические - это голубые линии и колоночки. А зелёные - это L1, это нечёткое мультиклассовое обобщение. И вы видите, у неё достоверность L1 единица. Это знаете, что означает? Что по этому критерию вообще не было ошибок. Вообще не было ошибок идентификации, не идентификации. То есть ни ложноположительных, ни ложноотрицательных. Вообще критерий Ван Рисбергена имеет высочайшее значение 0,989. Это означает, что задача решена качественно. И сейчас мы посмотрим, как распределены у нас результаты решения на в этих моделях. Вот в модели, это не в той модели. А, или в той? Да, в той. Нет, в четыре. Вот смотрите, значит, у нас синие - это ложные решения, красные и коричневые - истинные решения. Мы видим, что ложные решения есть, но при очень низких уровнях сходства, до 10%. Дальше идут только истинные решения. Это положительные вправо, вот где у нас уровень сходства больше нуля. А влево у нас вообще нет ложных решений. Вот. Значит, это очень хороший результат идентификации. И смотрим на ещё одну кривулечку вам покажу. Ну здесь она не очень наглядная, но смысл такой, что чем больше уровень сходства, тем больше доля истинных решений. Начиная с 20% уровня сходства, нет ложных решений.
У нас сегодня с вами будут занятия с вашей группой? Нет. Нет.
Ну тогда, значит, я сейчас запишу, и мы продолжим с этого места рассмотрение лабораторной работы 3.02 на следующем занятии. То есть начнём уже дальше рассматривать решение задач. Потому что раз у нас модели достоверные, то можно решать задачи. То есть мы с вами выполнили преобразование исходных данных в информацию, нашли зависимости, и применим эту информацию для решения задач. То есть мы по словам определяем авторов, провели формализацию предметной области, синтез и верификацию моделей, исследовали их на достоверность. И дальше нам осталось только решить задачи: задачу идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области. Что мы и сделаем на следующем занятии, я надеюсь. Так, это нам не то, не то. Так, ну сейчас я ещё хочу одну форму поместить туда в экран. Не знаю, получится или нет, но попробую. В чат, я имею в виду. О, получилось. Эта форма, ребята, она описывает порядок действий в системе Eidos не только в этой задаче, но и в других задачах. Всё, ребята, на этом у нас занятие закончено. До свидания, всего самого хорошего.
До свидания. До свидания. До свидания. Спасибо. До свидания.
До свидания.