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

***И.Т. Трубилина»,***

***Российская Федерация***

***Колесников Роман Юрьевич, ПИ2102***

***roman563412@gmail.com***

**Обработка текстовых данных. Интеллектуальная обработка текстов.**

**Заголовок**

Расширенные алгоритмы поддержки принятия решений и обработка текстов в системе Эйдос

**Резюме**

Лекция-демонстрация посвящена расширенному алгоритму поддержки принятия решений (ППР) в интеллектуальной системе Эйдос и её возможностям для анализа текстовых данных.

Введение и настройка системы:

Демонстрируется запуск системы Эйдос и установка лабораторной работы (3.03) для иллюстрации различных режимов.

Подчеркивается необходимость адекватной модели объекта управления для эффективного ППР.

Основные концепции ППР и моделирования:

Объект управления рассматривается как система со своими элементами, связями и эмерджентными (системными) свойствами, такими как эффективность.

Цель управления – повышение эффективности объекта, что эквивалентно повышению уровня его системности.

Цели (целевые состояния) ставятся руководством и могут выражаться в натуральном или стоимостном виде.

Обсуждается роль искусственного интеллекта (ИИ): ИИ 2-го поколения используется в управляющей системе для ППР, ИИ 3-го поколения будет способен к автоматизированной постановке целей.

Подчеркивается принцип необходимого разнообразия У.Р. Эшби: модель в управляющей системе должна быть сложнее (иметь более высокий уровень системности), чем объект управления, для адекватного отражения и управления.

Автоматизированный системно-когнитивный анализ (АСК-анализ) и расширенный алгоритм:

Расширенный алгоритм ППР (блок 2) включает все этапы АСК-анализа: постановку задачи (не автоматизирована), когнитивно-целевую структуризацию, формализацию предметной области, синтез и верификацию моделей, решение задач (идентификация, прогнозирование, ППР).

Демонстрируется режим 6.3, содержащий сам алгоритм.

Обработка целей и ограничений:

Проверяется совместимость поставленных целей с помощью кластерно-конструктивного анализа (режим 4.2.2.1), матрицы сходства, когнитивных диаграмм и дендрограмм.

Если цели несовместимы (альтернативны), требуется их корректировка руководством (уточнение, приоритизация) или развитие модели предметной области.

Если цели совместимы, выполняется SWOT-анализ (режим 4.4.8) для определения факторов, способствующих и препятствующих достижению каждого целевого состояния.

Анализируются имеющиеся ресурсы (технологические, финансовые) для реализации управляющих воздействий (факторов).

Если необходимые факторы (технологии) недоступны, система позволяет найти им замену среди доступных факторов, оказывающих сходное влияние (на основе анализа сходства факторов), и спрогнозировать результат с учетом этой замены.

Анализ текстов в Эйдос (Лематизация, Конструкты):

Система Эйдос позволяет анализировать текстовые данные. Демонстрируется установка ЛР 3.02 (атрибуция текстов).

Тексты могут обрабатываться целиком, по словам или по символам.

Вводится понятие лемматизации – приведение словоформ к их начальной форме (лемме) с использованием словаря (в Эйдос встроен словарь А.А. Зализняка на 2 млн словоформ). Это позволяет сократить размерность признакового пространства.

Объясняется понятие конструкта (по Дж. Келли) – биполярной шкалы для измерения понятий. Кластерный анализ в Эйдос выявляет такие конструкты.

АСК-анализ текстов позволяет решать задачи: формирование обобщенных лингвистических образов (семантических ядер) классов (авторов, жанров и т.д.), сравнение конкретных текстов с этими образами (атрибуция, идентификация), сравнение самих обобщенных образов, исследование предметной области через ее лингвистическую модель.

Практическая демонстрация и заключение:

Показан пример анализа текстов классиков русской литературы для определения авторства.

Демонстрируется высокая достоверность моделей, получаемых при обработке текстов, из-за переопределенности задачи (информации для идентификации обычно больше, чем минимально необходимо).

Подчеркивается, что Эйдос может работать с текстами на любых языках (человеческих, машинных, программирования) и решать задачи, связанные с их анализом и сравнением.

Детальная расшифровка текста:

I. Введение и настройка системы

A. Запуск системы и демонстрация интерфейса

Видно вам, да? Рабочий стол.

Видно, да?

Да.

Ну давайте тогда рассмотрим.

Значит, я запускаю систему и на маленьком примерчике мы рассмотрим, как это делается.

B. Установка лабораторной работы 3.03

Значит, запускаю систему, устанавливаю лабораторную работу 3.03. Эта работа нам нужна только для того, чтобы просто показывать различные режимы системы, которые необходимо использовать, чтобы принимать решения в развитом варианте, в развитой форме.

То есть мы сейчас устанавливаем эту лабораторную работу.

Ребята, вам желательно делать то же самое, всё-таки лабораторная работа. Вот, то есть постарайтесь сейчас что сделать? Запустить систему, именно используя файл start Eidos.

Вот, тогда у вас, когда вы это сделаете, произойдёт обновление и система должна работать. Если у кого-то не будет работать, сразу же мне об этом сообщаете.

Будем разбираться почему. Вот, иногда это удаётся сделать, как вы заметили.

C. Синтез модели и переход к алгоритму ППР

Значит, сейчас что я делаю? Я устанавливаю работу 3.03 и сразу же выполняю режим синтеза модели.

И открываю режим 6.3, где у нас будет сам алгоритм принятия решений содержаться.

Вот. То есть здесь его описание этого алгоритма.

II. Основные концепции ППР и моделирования

A. Роль модели объекта управления

И сейчас я буду по прямо по блок-схеме всё рассказывать и ссылаться на работы, какие-то идеи из этих работ вам сообщать и в самой системе показывать соответствующие режимы.

B. Целеполагание и Искусственный Интеллект

Значит, так вот, ребята, я вам начал на лекции говорить о том, что интеллект третьего поколения сейчас разрабатывается. И в этом интеллекте третьего поколения функция постановки цели тоже будет автоматизирована. То есть не человек будет ставить цель андроиду, к примеру, там или какому-то какой-то системе автоматизированной, а она сама будет выбирать цель и обеспечивать её достижение. Но, наверное, будут какие-то критерии выбора цели тогда.

Ну а сейчас, пока мы говорим о системах, в которых применяется интеллект искусственный второго поколения в составе управляющей системы для принятия решений, то мы говорим о том, что цели ставит руководство.

C. Целевые состояния объекта управления

То есть руководство ставит цели управления, определяет будущие целевые состояния объекта управления.

Обычно целевые состояния выражаются в двух вариантах: в натуральном выражении и в стоимостном выражении.

В натуральном выражении целевые состояния обычно изображают отображаются в виде количества и качества продукции.

А в стоимостном выражении обычно прибыль и рентабельность, какие-то единицы измерения денежных.

D. Объект управления как система

Объект управления представляет собой систему. То есть он состоит из многих элементов, взаимосвязанных друг с другом. И эти элементы, значит, они имеют определённые свойства. И свойства системы не сводятся к свойствам элементов, из которых она состоит. То есть у системы есть системные свойства, ради которых она и создаётся, если это искусственная система.

E. Эффективность как системное свойство

И эффективность, эффективность тоже является системным свойством. Если мы говорим об экономике, эффективность объекта управления, эффективность работы предприятия, то есть получение большего объёма более качественной продукции, повышение прибыли и рентабельности предприятия – это всё системные свойства, которые могут быть сформированы у предприятия за счёт изменения его структуры и повышения уровня системности.

F. Связь эффективности и уровня системности

Значит, я хочу вам сейчас выразить такую мысль, что эффективность объекта управления или повышение эффективности объекта управления, значит, ну, эффективность объекта управления – это его эмерджентное свойство, системное свойство. А, значит, цель системы управления – повышение его эффективности объекта управления, то есть повышение его уровня системности. Об этом говорится вот в этой моей работе, с которой вы можете ознакомиться. Там есть ссылочки.

G. Принцип необходимого разнообразия Эшби и адекватность модели

И для того, чтобы принимать решения об объекте управления, которые бы мы, которые в результате исполнения которых этот объект управления переходил бы в заданные целевые состояния, для этого в управляющей системе вот здесь, входящей в состав объекта автоматизированной системы управления, должна быть модель объекта управления адекватная, правильно его отражающая.

И вот что касается этой модели, что мы можем сказать? Что если у объекта управления, как системы, есть определённый уровень системности, то и модель его, которая его отражает в управляющей системе, тоже должна иметь определённый уровень системности. Иначе она, то есть именно выше должен уровень системности модели должен быть выше, чем уровень системности объекта управления.

Иначе эта модель не сможет его полностью адекватно отразить.

И если модель объекта управления, которая есть в управляющей системе, не полностью отражает объект управления или не вполне адекватна, то тогда и само управление будет тогда неполным и неадекватным. То есть вот эти решения, которые вырабатываются управляющей системой, они не смогут гарантированно переводить объект управления в целевые состояния. А будет просто повышаться вероятность перехода его в целевые состояния, но не гарантироваться переход в эти целевые состояния.

Почему я сейчас про это говорю? Потому что, да, значит, есть принцип методологический, теоретический принцип Уильяма Росса Эшби, который как раз об этом говорит. Ну это сейчас я чуть попозже об этом скажу. Вот. Но сейчас я об этом почему говорю? Потому что мы должны перед началом принятия решений иметь модель объекта управления, уже обеспечивающую высокую достоверность.

И это является основой для того, чтобы принимать адекватные решения об этом управлении этим объектом управления.

III. Автоматизированный системно-когнитивный анализ (АСК-анализ) и расширенный алгоритм

A. Этапы АСК-анализа

Шаг второй. Значит, мы смотрим режим 6.4, который я вам до этого показывал на предыдущей лекции. И видим в этом режиме ряд этапов. Я про них говорил. Это этапы автоматизированного системно-когнитивного анализа.

Первый этап он не автоматизирован в системе Eidos – постановки задачи. Этап когнитивно-целевой структуризации предметной области, когда мы решаем, что у нас является объектом управления, что факторами, которые на него влияют, и что его будущими состояниями.

А последующие этапы все автоматизированы. Это формализация предметной области в системе Eidos автоматизирована. Синтез и верификация моделей и решение различных задач.

Задачу идентификации я вам рассказал на прошлом лабораторном занятии, похоже так, да? А сейчас рассказываю про принятие решений. Да, и принятие решений в простейшем варианте тоже, наверное, рассказал. И сейчас рассказываю уже в развитом варианте принятия решений.

B. Блок 2 алгоритма как реализация АСК-анализа

То есть вот этот блок два в развитом алгоритме принятия решения, он представляет собой все этапы автоматизированного системно-когнитивного анализа вместе взятые. То есть все вот эти этапы все вместе, они и представляют собой шаг два развитого алгоритма принятия решений. То есть когнитивно-целевая структуризация, формализация предметной области, синтез и верификация моделей, определение наиболее достоверной из них. И повышение уровня системности и адекватности модели управления до уровня, который обеспечивает адекватное отражение объекта управления и, в соответствии с принципом Уильяма Росса Эшби, разумное, адекватное управление этим объектом управления.

C. Принцип У.Р. Эшби (детализация)

О чём говорил Уильям Росс Эшби, во-первых? И кто это такой? В двух словах я вам сейчас скажу, и что он вообще, в чём заключается суть его принципа? Это друг Норберта Винера, разработчика кибернетики. Но если Норберт Винер был математиком, то Уильям Росс Эшби был психиатром. То есть он занимался психопатологиями, лечением психопатологий. И они, видимо, обсуждали, видимо, вопросы кибернетики. И вот Эшби предложил принцип, который сейчас носит его имя, он вошёл в историю науки, в историю создания кибернетики.

К кибернетике по современной классификации научной, научных направлений относятся теория управления и интеллектуальные технологии в том числе. То есть вот что такое кибернетика с современной точки зрения? Это теория управления и интеллектуальные технологии.

Так вот, что говорил, собственно говоря, Уильям Росс Эшби? Он говорил о том, что если мы хотим адекватные принимать решения, то нам нужно хорошо знать объект управления. Если мы его знаем не очень хорошо, какие-то только аспекты его знаем, а другие не представляем себе, то и решения наши будут не вполне адекватными, и мы не сможем обеспечить полное управление объектом управления. То есть мы не сможем гарантированно переводить его в те или иные целевые состояния. Вот это вот очень важно понимать.

Ну здесь я бы ещё такой привёл бы пример вам, что если есть два человека, один развитый, а другой не очень, то тот человек, который не очень развит, он будет иметь представление о человеке развитом упрощённое, ущербное, неадекватное. Ну, скорее всего, он будет думать, что он такой же, как он по уровню. Вот. А человек более развитый, у него будет адекватное представление о человеке более простом. То есть более сложная система адекватно отражает в себе более простую, а более простая – упрощённо и неадекватно, ограничено отображает более сложную.

И вот сложность управляющей системы, а точнее сложность модели объекта управления, которая есть в управляющей системе, должна быть выше, чем сложность объекта управления, иначе управляющая система не может им управлять.

D. Связь этапов АСК-анализа с шагом 2 алгоритма

Вот эти все этапы, которые на шаге два выполняются, это и есть этапы, которые, ну, в общем, вот исходных данных в систему, разработка классификационных описательных шкал градаций обучающей выборки, синтез модели, выбор наиболее достоверной. Это всё мы с вами делали на лабораторной работе, когда изучали, так сказать, основы моделирования в системе Eidos.

IV. Обработка целей и ограничений

A. Проверка совместимости целевых состояний

Потом мы должны определиться. Вот эти целевые состояния, которые руководство нам поставило, они являются совместимыми друг с другом по системе детерминации или нет?

То есть можно ли их ставить цель одновременного достижения этих целевых состояний? И для того, чтобы на этот вопрос ответить, мы проводим кластерно-конструктивный анализ будущих состояний, ну вообще всех, и желательных, и нежелательных будущих состояний.

B. Кластерно-конструктивный анализ (Матрица сходства, Диаграммы, Дендрограммы)

Рассчитываем матрицу сходства таким образом, в режиме 4.2.2.1. Да, здесь, кстати, об этом говорится, что мы должны посчитать матрицу сходства в режиме 4.2.2.1. И потом на её основе вывести когнитивные диаграммы и дендрограммы и определить, насколько сходны будущие состояния объекта моделирования друг с другом по их системе детерминации, то есть по тем факторам, которые переводят объект управления в эти состояния.

C. Интерпретация результатов анализа совместимости

И здесь я хочу вам ещё вот что сказать, что, видите, мы смотрим здесь на шаге три, одно у нас целевое будущее состояние или нет. Если одно, тогда мы вот этого не делаем кластерного анализа, а сразу переходим на шаг шесть – SWOT-анализ, который я вам показывал на прошлом занятии и на лекции сейчас немножко показывал.

А вот если много состояний, тогда мы сразу же определяемся, сходны ли эти состояния по тем факторам, которые обуславливают переход в эти состояния. И здесь есть два варианта: что какие-то состояния будущие целевые, целевые состояния, являются достижимыми одновременно, то есть совместимыми по обуславливающим их факторам, или они могут быть взаимоисключающими, альтернативными. Ну то есть у них разная система детерминации, разные факторы обуславливают переход в эти состояния. Вот. Если они совместимы, тогда это задача корректно ставится для перевода в эти состояния. Если нет, тогда, значит, некорректно.

D. Корректировка целей или модели при несовместимости

Значит, вот на шаге пять это и определяется. Поставленные задачи управления корректны, совместимы, достижимы одновременно? Если да, мы переходим на шаг шесть. Если нет, мы переходим опять на шаг один, где мы просим руководство э-э с более корректно поставить цели, так, чтобы они были одновременно достижимыми.

Либо какой-то какие-то приоритеты нам задать, сказать, что вот эта цель более важна, эта менее важна. Тогда мы будем, если эти цели альтернативные, то есть одновременно не достигаются, мы тогда будем достигать более важных целей в ущерб, так сказать, менее важным.

Либо мы должны понимать такую вещь, что то, что в нашей модели предметной области невозможно, может вполне быть возможным в другой, более развитой модели предметной области. Ну, например, в эпоху воздухоплавания, воздушные шары, господствует такое представление, что летательные аппараты тяжелее воздуха невозможны, потому что известен только один принцип полёта – закон Архимеда. И проходит там каких-то несчастных 200 лет, и выясняется, что ещё существуют десятки других видов полёта, принципов полёта: и аэродинамический принцип – самолёт, например, и вертолёт, и ракетный принцип. Вот, ракетные двигатели. Вот, не не просто, ну, реактивные и ракетные, и винтовые, может быть, даже на самолётах, вертолётах. Вот. И возможны другие способы полёта, такие, скажем, как за счёт взаимодействия с магнитным полем Земли там и тому подобное, более экзотические такие, электромагнитные виды движения.

Тоже возможно. И вот оказывается, что если такая цель будет поставлена – создать аппарат летательный тяжелее воздуха, то многие скажут: "Да это невозможно". И будут правы, но только в том случае, с учётом того, какие у них представления, скажем так. Для них это невозможно, потому что у них вот такие представления о принципах полёта, отражённые в их модели реальности. Вот. А у кого-то другие могут быть более развитые модели, и там для них это вполне возможно. То есть я хочу вам сказать, что если какие-то цели получаются в соответствии с результатами кластерного анализа невозможными одновременно, то это не означает, что они вообще невозможны одновременно, а лишь означает, что они невозможны одновременно в нашей модели. И можно либо сами цели корректировать, либо можно модели развивать. И тогда это, может быть, станет возможным.

E. SWOT-анализ (упрощенное ППР)

Ну, если целевые состояния оказались у нас совместимыми по системе детерминации, то тогда мы переходим на шаг шесть и решаем задачу поддержки принятия решений в упрощённом варианте. Я вам это показывал. 4.4.8 режим. SWOT-анализ. Выбираем целевое состояние, выбираем модель и видим, что у нас из модели очень быстро извлекаются значения факторов, которые способствуют переходу объекта моделирования в целевое состояние, выбранное, и которые препятствуют. И это можно в графической форме тоже посмотреть.

F. Объединение факторов для совместимых целей

Однако мы здесь видим только возможность одно состояние, да, задать в качестве целевого, а у нас может быть их несколько. Что в таком случае делать? Предлагается несколько раз применить SWOT-анализ для разных целевых состояний. Дело в том, что раз они не альтернативны, сходны по системе управляющих факторов, то тогда вполне это допустимо.

Дальше мы объединяем в SWOT-анализе, в режиме 4.4.8, те факторы, которые необходимы для достижения этих совместимых целевых состояний в одну систему факторов.

G. Анализ реализуемости факторов (ресурсы, технологии)

И потом смотрим, а мы все эти факторы можем применить на практике для управления объектом управления, для оказания воздействия на объект управления или нет? И мы оцениваем прежде всего технологические и финансовые возможности. Ну, желательно даже имеющиеся налицо, как говорится, уже технологические возможности. Почему? Потому что если у нас говорить не только о тех, которые уже имеются технологиях, но ещё и о финансовых возможностях, здесь имеется в виду, что мы можем купить нужные нам технологии и применить их. Но тогда достижение цели получения высокой прибыли и рентабельности может оказаться затруднённым, усложнённым. То есть потому что у нас появятся дополнительные затраты на достижение цели.

H. Замена недоступных факторов

Представим себе, что у нас каких-то технологий нет. Значит, что мы можем сделать? Либо купить эти технологии фактически и установить у себя и снизить прибыль и рентабельность. Либо мы можем попытаться заменить те технологии, которые у нас отсутствуют, но которые необходимы для достижения цели. Но перед тем, как их заменять, возникает вопрос: а надо ли вообще их заменять? Может быть, вообще без них обойтись можно как-то? Для того, чтобы этот вопрос решить, получить ответ на этот вопрос, для этого мы применяем на шаге восемь...

Значит, ну, видим, что у нас, э-э, значит, если у нас не все факторы мы можем применить на практике, которые у нас, которые рекомендуются, значит, тогда мы переходим на шаг девять. Если же все возможности есть применить, тогда мы сразу выходим на конец алгоритма, на проверку корректности той системы факторов, которую мы выработали путём вот суперпозиции из разных систем SWOT-анализа. А если у нас нет возможности применить факторы, тогда мы смотрим, а что будет, если мы вообще исключим этот фактор из факторов, которые управляют?

Вот нам рекомендованы были какие-то факторы, значения факторов. Мы возьмём вообще его и уберём. И спрогнозируем, что у нас получится. Вот берём какое-то фактор и убираем. То есть берём вот здесь вот вместо 45 код фактора, код значения фактора, берём, ставим там ноль. Всё. То есть мы это значение фактора убрали.

И потом проводим уже прогнозирование с такой системой факторов.

И получаем некий результат. И смотрим, этот результат нас устраивает или нет?

Значит, я вам могу сказать, что у нас мышка 2 имела признак, который делал её на первых позициях, было у нас класс мышка и был элемент компьютер на второй позиции. Мы этот признак убрали. Всё равно система идентифицирует его как элемент компьютера, но уже не на первой позиции, а на четвёртой. Но всё равно это ещё довольно высокий уровень сходства. О чём это говорит? О том, что мы достигаем целевого состояния в случае отсутствия этого фактора. То есть просто мы вообще выкидываем его и всё, и не пользуемся. И, ну, можем достичь целевого состояния, выбросив те значения факторов, которые мы не можем применить.

Если это удаётся достичь целевые состояния, тогда мы просто осуществляем выход из этого алгоритма принятия решений. То есть этим самым путём суперпозиции мы получили нужную систему факторов и убедились, что целевые состояния достижимы одновременно, не альтернативны, а совместимы.

Если же у нас оказалось, что целевое состояние не достигается, вот, допустим, сейчас нас это мы то, что мы сделали, посмотрели на результат, нас это не устраивает. Тогда надо этот фактор, по-видимому, заменить каким-то другим, который оказывает сходное влияние на объект моделирования.

I. Поиск замены с помощью анализа сходства факторов

Как узнать, каким фактором можно заменить? Для этого мы рассчитываем матрицу сходства. Здесь я во всех моделях рассчитываю, но можно в выбранной модели рассчитывать их. И тоже смотрим на дендрограмму нужной модели.

Такая картинка получается нечитабельная, то есть здесь слишком много информации, и она, так сказать, не воспринимается на глаз. Можно, конечно, увеличить размер поля отображаемого до 4К. А можно взять и уменьшить число отображаемых признаков.

Вот, и мы увидим тогда уменьшенное подмножество этой системы всей признаков. Но тот признак, на котором стоит курсор, будет у нас обязательно показан. И другие признаки, которые оказывают на него сходное влияние, то есть на объект моделирования сходное влияние. И здесь уже мы можем выбрать либо этот признак, который у нас нет возможности использовать, либо другие признаки, оказывающие сходное влияние на объект моделирования. Также можем перейти в режим кластерного анализа значений факторов и посмотреть...

Так, 4 3... на дендрограмму признаков. Тоже она отражает сходство, различие признаков по их влиянию на объект моделирования.

Сейчас немножко притормаживает, потому что у меня расчёт идёт трудоёмкий. Вот результат кластерно-конструктивного анализа значений факторов. Есть факторы, которые очень сходны друг с другом по влиянию на объект моделирования. И если один из них нет возможности применить, то можно применить другой. Будет будет сходный результат получен. Но будет ли он получен или нет, это можно проверить.

J. Прогнозирование с учётом замены факторов

Вот мы взяли, заменили, значит, какие-то признаки, значения факторов, которые мы не можем применить, на другие, которые оказывают сходное влияние на объект моделирования. И провели прогнозирование, ребята. То есть решили задачу прогнозирования с такой вот системой факторов. И у нас получилось два, один из двух вариантов: либо это вот изменённая система факторов обеспечивает переход объекта моделирования в целевое состояние или состояние, либо не обеспечивает. Если обеспечивает, тогда выход из алгоритма принятия решений. Если нет, то у нас есть опять же те же самые два варианта: либо просить руководство, чтобы оно скорректировало цели или вообще убрало менее значимые или сделало, указало приоритеты в достижении целей. Либо нужно ставить вопрос о том, чтобы совершенствовать модель объекта управления, переходить так, чтобы эта модель была более развитая, более совершенная, более адекватная, может быть. И обеспечивала решение тех задач, которые ставятся руководством, в отличие от той модели, которая сейчас у нас есть. Ну, идёт речь о развитии технологий, по сути дела. То есть более развитая модель объекта управления предполагает наличие у нас других технологий воздействия на этот объект.

K. Итог алгоритма

Вот. И потом повторяются все эти вот шаги, которые мы рассмотрели. И пересматриваются целевые состояния, может быть, те же самые, а вот факторы, влияющие на объект моделирования, уже будут другие, потому что уже модель другая. Вернее, наоборот, модель другая, потому что другие факторы, то есть дополнительные факторы у нас появились. И эта другая модель, она, может быть, и позволяет решить эту задачу. Ну, примерно, как вот у инженеров, разрабатывающих летательные аппараты, модель полёта только на основе закона Архимеда, или у них модель, которая основана на законах аэродинамики, например, или ракетных технологиях. Вот если у них модель более развита, то они могут создать летательные аппараты тяжелее воздуха, которые сказала начальство разработать вот такие аппараты. А если модель была бы на законах Архимеда и всё, и больше других нет, им такое задание дают, они говорят: "Это невозможно, потому что либо будет летать, либо будет тяжелее воздуха". Вот. Ну тогда летать не будет, по в соответствии с их моделью.

Ну вот, ребята, я вам сейчас рассказал развитый алгоритм принятия решений.

V. Анализ текстов в Эйдос (Лематизация, Конструкты)

A. Переход к анализу текстов (ЛР 3.02)

И показал вам режимы системы Eidos, в которых мы решаем все те задачи, которые необходимо уметь решать в этом алгоритме.

Вот. Теперь давайте вопросы, какие у вас есть, ребята? Задавайте.

Сейчас мы рассмотрели с вами...

Спасибо.

Вопросов нет, да, Диана? Или есть?

Нет.

Нету, да? У вас у всех система работает, да, на компьютерах?

Ну сейчас, пожалуйста, поднимите руку те, кто пытается за мной повторять на своём компьютере то, что я показываю.

Ну, не густо. Да. Ну так, уже получше. Получше уже. Ну, фактически... Да, вообще даже неплохо уже. Опускайте руки.

Ну теперь, наверное, будем рассматривать следующий вопрос. То есть сейчас мы с вами проходили очень много вопросов различных, но на примере одной лабораторной работы. То есть я вам давал базовые понятия, иллюстрировал их в системе, в этой системе на примере этой простой лабораторной работы обработки табличных данных. То есть внешние данные табличного вида.

А теперь посмотрим, ребята, на другие задачи, которые можно решать в системе. Вот, вернее так, задачи те же самые можно решать, но на основе других данных исходных: данных текстовых и данных графических.

Вот. И давайте сейчас перейдём к рассмотрению работы обработки текстовых данных, интеллектуальная обработка текстов.

У нас ещё почти час времени, да? Правильно?

Да. До 17:00.

Тогда будет ближе к концу занятия, вы мне тогда напомните, пожалуйста.

Хорошо.

B. Настройка системы для анализа текстов

Значит, переходим мы, то есть запускаем систему. Давайте, запускайте систему. Если запущена, выходим в режим 1.11.

То есть этим самым мы удаляем все приложения, которые там у нас были, вот эти, на которых мы сейчас, на примере которых мы сейчас вот вопросы изучали, развитый алгоритм принятия решений. Заходим в режим 1.3. Добавляем лабораторную работу 3.02. Работа 3.02. Атрибуция анонимных и псевдонимных текстов. Всё, нажимаем Окей, Окей, по умолчанию. То есть ничего там выдумывать не нужно.

C. Интерфейс ввода текстовых данных и опции обработки

И вот у нас появляется такая экранная форма, ребята, управления автоматизированным программным интерфейсом ввода данных в систему Eidos из внешних данных, из внешних источников данных текстового типа, ребят. Ну, правда, тексты тоже в таблице Excel находятся. А есть ли возможность в системе Eidos вводить текстовые данные, которые не в таблице Excel, а в виде файлов? Есть, ребят, но это другой интерфейс. Но сейчас я вам покажу на примере вот этого интерфейса, который вроде как табличные данные, но они не числовые, а текстовые.

И я вам скажу сразу, они могут быть и числовые, и текстовые одновременно в этом программном интерфейсе.

И вот здесь смотрите, ребята, у нас здесь есть птичка, которая здесь сейчас установлена. И если мы эту птичку вот здесь установим, то у нас появляется ещё окошко с дополнительными параметрами. И мы видим, что здесь написано против этой птички: применить специальную интерпретацию текстовых полей признаков. А здесь вот специальная интерпретация текстовых полей классов. И тоже можно задавать разные опции здесь.

Какие же это опции? Смотрите. Значение текстовых полей рассматривать целиком. Так обычно и делается. То есть когда мы обрабатываем числа, числовую информацию из таблиц, то обычно текстовые значения рассматриваются целиком. Но если мы применяем специальную интерпретацию текстовых полей, то они могут рассматриваться тексты, которые в полях таблицы, как слова, то есть как совокупность слов. Вот. И мы здесь можем задавать минимальный размер этого слова. Ну, опыт показывает, что очень короткие слова нет смысла рассматривать. Вот, значит, здесь больше трёх символов уже рассматриваются. Мы можем рассматривать даже символы как признаки этого текста.

D. Лематизация: объяснение и реализация в Эйдос

И здесь вот внизу, видите, появляется у нас окошко, где мы можем задать: проводить лемматизацию или не проводить?

Вот. То есть разные опции. Если значение поля целиком, то сортировать, не сортировать, выделять уникальные, не выделять. А если слова, тогда проводить лемматизацию или не проводить?

И, наверное, из вас не все знают, что такое лемматизация, ребят, да? Правильно?

Поднимите руку, кто знает, что такое лемматизация?

Лучше скажите.

А?

Лучше скажите.

Значит, ну да, слово такое, которое вроде бы слышали, лемма, да? Но то, что я сейчас вам рассказываю, относится к математической лингвистике. А вы её, наверное, не изучали, поэтому, наверное, и не знаете, что такое лемма в лингвистике. Значит, я вам сейчас показываю. Я вот выхожу на свой сайт, на ту страничку, где скачать систему Eidos. Вот. И вот здесь мы видим про базу лемматизации есть информация. Пункт пятый. Тут всякие пункты есть там, вот. В общем, пункт пятый, где говорится о том, что в состав системы входит база лемматизации. И даётся ссылочка на эту базу и ссылочка на источник, откуда я её взял.

Эту базу разработана академиком РАН Зализняком, который года два назад умер. Это замечательный учёный, который действительно энтузиаст, внёс большой вклад в науку. Он был расшифровщиком новгородских рукописей, их исследователем. И он же и разработал словарь лемматизации русского языка, в который входят, ребята, сейчас вдумайтесь чуть-чуть так, 2 млн русских слов.

А сколько всего русских слов есть, знаете, нет? В языке?

Какое у вас предположение? Сколько слов входит в русский язык?

Рассказать?

Ну я так думаю, что десятки миллионов, десятки миллионов слов входят, ребята. При этом для того, чтобы общаться в быту, достаточно 200-300 слов. На таком более-менее приемлемом уровне, ну 2.000 слов хватает. Минимальный уровень 200-300 слов, уже можно общаться в быту. 3.000 слов – вполне нормально, без особых ограничений можно общаться. А вообще-то, вот Пушкин использовал, скажем, 25.000 слов. Это больше всего, чем, больше, чем какие-либо другие писатели. Вот. Ну и можно найти данные об этом, какие слов, сколько, вернее, сколько кто использовал слов различных. И вот, насколько я знаю, Пушкин больше всех использовал различных слов, больше, чем другие писатели и поэты. А там идёт речь о том, что там 18, там 20.000, там 15.000. Повторяю, для того, чтобы нормально, на нормальном уровне общаться в быту, 2-3.000 слов вполне достаточно.

Вот. Что же такое сделал Зализняк, когда разработал словарь в 2 млн слов? Значит, я могу вам сказать так: он использовал, в этом словаре включил все слова, не только исходные, но и производные от них. И вот я сейчас как раз и объясняю, что такое лемма, что такое исходное слово, производные слова. Значит, есть способы различные словообразования в русском языке: падежи, числительные, рода, то есть можно множественное число, единственное число, можно склонять это слово, да, по-разному. Вот, можно добавлять какие-то суффиксы к ним или предлоги. Вот. В общем, можно с этими словами очень много разных вещей производить, операций, при которых возникают новые варианты этих слов. Их называют словоформы. А вот те слова, то слово, которое было исходным, изначальным, базовым для всех словоформ, называется леммой.

И вот я вам приведу пример очень простой. Стол, например. Слово стол. А какие слова образуются на основе этого слова стол путём там его видоизменения, словообразования? Столы, столики, уменьшительное, да, выражение. То есть видите, числительные, уменьшительные, да? Вот. Ну и также есть совсем другие слова, которые уже не стол обозначают, а совсем другие вещи, но другие объекты, но тоже связаны со словом стол. Столица, например. Это город, это не стол, это это город. Какой город? А там, где стол стоит, да? Ну, стол, видимо, царя, за которым царь сидит, правитель. Вот. Или или много правителей за этим столом сидят, министры там, например. Вот. Или столяр. Кто такой столяр? Это тоже не стол, но это тот человек, который делает столы, правильно? Вот. А ещё можно сказать, что есть столовая. Что такое столовая? Это либо комната, либо помещение, в котором, значит, стоят столы и там кушают просто, да, на этих столах. Вот. А ещё есть столешница там, ну и так далее, и так далее. То есть можно понять, что из этого слова стол огромное количество разных слов производных можно создать разными способами.

И вот что сделал Зализняк? Он взял все эти слова написал. Э-э, которые созданы на основе слова стол, а слева написал слово стол у всех у них. То есть мы берём, находим слово какое-то, которое у нас есть в русском языке, переходим в первую колонку и видим, что это слово производным является от слова стол. То же самое: конь, конюшня, конюх там. Вот. Понятно, да? Князь. И пошло-пошло. Понятно, да?

Так вот таких производных слов у него в словаре 2 млн. Ну, гораздо больше, чем вот этих вот лемм. Вот лемм как раз, может быть, там не так уж и много, там тысяч 20, действительно, там или 30. Вот. Теперь слушайте дальше внимательно. Значит, я сделал в системе Eidos возможность лемматизации при обработке текстов. То есть она обращается к словарю Зализняка, я его включил в состав системы. Для этого я сделал небольшую программку, которая его преобразовала в базу данных этот словарь. Он представлен был в Экселе, в Экселе на в виде двух таблиц, потому что там не больше миллиона строк в таблице. Вот я сделал небольшую программку, которая вытащила оттуда эти слова, леммы и словоформы и сделала одну базу данных. Эта база данных, она размещена у меня на сайте. Называется Лема ДБФ. Вот ссылочка на неё прямая на эту базу. Не знаю, будет ли она скачиваться? Не будет, я её удалил, потому что она 200 МБ занимает, но она входит в состав полной инсталляции. Там у меня на сайте уже нет места, не хватает места на хостинге.

Вот. А это вот статья Хабра про это вот, про эту базу лемматизации. И здесь возможность её и скачать, что я и сделал.

И система, работая с текстом, если задана опция лемматизировать, проводить лемматизацию, то она использует эту базу лемматизации Зализняка, ищет там леммы и заменяет слова на их леммы. Получается заметное сокращение числа признаков. То есть если у нас слова как признаки текстов рассматриваются, то получается там раза в полтора меньше слов.

При этом некоторые слова она не находит в словаре Зализняка. Тогда что она делает, ребят? Как вы думаете, что делает система, если она не находит слово в словаре лемматизации?

Возможно, оставляет неизменным?

Ну она, во-первых, вы не меняет ни на что, потому что она не может этого сделать. Но она его просто добавляет в словарь лемматизации и об этом сообщает, что столько-то слов, там 300, например, или 400, добавлено в словарь лемматизации. И в системе я сделал простенький интерфейс, позволяющий написать леммы для этих слов и размножить. Понятно, да?

То есть это большая работа, конечно. Для этого нужен коллектив, но это вполне такая реальная работа. То есть можно этот словарь э-э развивать э-э Зализняка. Если это сделать, то там получится тогда гораздо больше 2 млн слов. Я уверен абсолютно, что там даже и тремя миллионами не ограничится. Будет там и пять, и семь миллионов слов. Вот. Ну, я ещё скажу, что когда вот я смотрел, что там написано в словаре Зализняка, то я понял, что там в основном слова такие несколько архаичные, я бы сказал. Такие классические, классические. А всяких вот этих новомодных слов там нету. Там нет слова сайт, например, или слово файл. Ну это, я не очень уверен, что это русские слова, по-моему, это, конечно, если строго говорить, то это не русские. Но многие из них вошли органично в состав русского языка, правильно? Мы ж не применяем какое-либо русское слово вместо слова сайт. Ну, соответственно, оно вошло в состав русского языка.

E. Демонстрация анализа текстов классиков

Дальше мы что делаем? Дальше мы нажимаем Окей. И происходит всё то же самое, что и при обработке таблиц. Показывается нам внутренний калькулятор, в котором нам показано, сколько у нас шкал различных типов, числовых и текстовых, классификационных и описательных. Их может быть много и описательных шкал, и текстовых, и классификационных шкал при обработке текстов может быть много.

Вот, ну сейчас посмотрим, что у нас получится. Сейчас она числовых шкал здесь не обнаружила. Выйти на создание модели, создала классификационные описательные шкалы градации. В качестве градаций выступают слова в описательных шкалах, а в классификационных – поле целиком, значение поля целиком.

Потом была создана обучающая выборка, всё переиндексировано. Теперь мы смотрим, что у нас получилось в результате. Классы. Классы – это у нас авторы, классики русской литературы: Грибоедов, Пушкин, Тургенев, Толстой, Горький, Лермонтов, Гоголь, Достоевский. А в качестве признаков у нас слова, которые они применили в своих сочинениях. Но я откуда взял эти тексты, ребята? Я эти тексты в качестве примера, который я привёл для формирования образов лингвистических этих писателей, я взял просто школьные сочинения. То есть это не их тексты, а это тексты о них, о них.

F. Идентификация авторства и анализ результатов

И, ребята, хочу вам показать, где ссылочка на мои работы есть, где в системе можно найти ссылку на мои работы, связанные с обработкой текстов. Для этого мы выходим на авторизацию. Это можно также сделать, когда мы просто входим в систему первый раз. И вот здесь вот у нас, смотрите, есть кнопочка такая: подборка публикаций по АСК-анализу. Видите, да? И кратко об АСК-анализе. Кратко об АСК-анализе – это работа моя, АСК-анализ как метод решения проблем. Здесь коротко описан АСК-анализ. Какие задачи он позволяет ставить и решать.

Вот. А вот здесь мы выходим на подборки публикаций. Видите, прямо сразу на сайт попадаем. И здесь у нас монографии, свидетельства, патенты, учебники, учебные пособия, тематические подборки публикаций. Видите, да? И вот тут есть у нас работы по АСК-анализу текстов. Видите? 6.3. Нажимаем и видим работа по АСК-анализу текстов. И появляется у нас ссылочка и ряд работ. Ну, как, не то чтобы там прямо так уж очень много, но достаточно много. Для чего достаточно?

Вот. Сейчас я вам ссылочку эту дам на эти работы по анализу текстов. И что мы здесь видим, ребята? Мы видим, что интеллектуальный системно-когнитивный анализ текстов позволяет решить ряд задач.

Да, все эти статьи, на которые я здесь ссылаюсь, и работы, они все в открытом доступе, то есть их все можно почитать.

Вот. И вот, э-э, какие задачи можно решать? Смотрим. Первая задача: формирование обобщённых лингвистических образов классов. Класс – это обобщённая группа, то есть обобщённый образ, созданный на основе примеров. Ещё их называют семантические ядра. Обобщённый лингвистический образ класса – это семантическое ядро, которое можно создать на основе фрагментов или примеров текстов, относящихся к этим классам. Тексты могут быть на любом языке.

И, знаете, что интересно, ребят? На любом человеческом языке или на любом машинном языке. Или на любом языке программирования.

Вот. Теперь давайте я вам покажу ссылочку на место, где приведена... Ну, ссылочку на сами работы я показал. А теперь покажу вам ссылочку на место, где...

Где лекции просто содержатся, а видеозанятия, э-э, сдвоенное, на котором рассматривается обработка текстов и обработка графической информации. То есть работа 3.02, то есть мы вот этот пункт второй выполняем всё время сейчас на лабораторных. Первый на лекциях, а пункт второй мы выполняем, вот мы его выполняли уже работу 3.03 освоили, я там довольно подробно всё рассказывал. Сейчас работу 3.02 мы изучаем. А потом вот ещё я вам даю ссылочку на видеозанятие, где ещё э-э проводится АСК-анализ изображений. И сейчас я вам эти ссылочки скину на само этот, на само это задание. Вот ссылочка на видеозанятие. И на само задание вот это тоже задание по разработке облачных Eidos-приложений.

Так, что-то такое я не то сделал.

Вот можете нажать сейчас на Яндекс Диск. Это мой Яндекс Диск, связанный с моей почтой. И там находятся вот эти видеозанятия. Вот то, что я сейчас вам рассказываю, там всё это есть. Ну, правда, я когда рассказываю, я немножко рассказываю лучше, чем написано. И на одних видеозанятиях одно рассказываю, на других немножко другое. То есть иногда у меня состояние получше, иногда похуже, там замученный, не замученный, вечером, не вечером, вот такое вот. Ну сейчас я немножко уже замученный, конечно, потому что я с самого утра, с 8:00 часов принимал экзамен, потом вёл, читал лекцию, ехал туда-сюда, ещё занятия провёл. То есть вот такое вот уже нагрузка.

Вот. Так вот, показываю вам э-э дальше. Э-э файлы исходных данных.

Да, извините. Сейчас мы э-э опять перейдём на тексты, на работы, связанные с обработкой текстов. Я вам немножко подробнее расскажу, что там за задачи могут решаться.

Значит, первое – могут создаваться лингвистические образы классов, то есть семантические ядра. А потом можно сравнивать, ребята, количественно, количественно лингвистический образ конкретного человека или автора, или описание объекта на естественном языке с обобщёнными лингвистическими образами классов. То есть можно что сделать? Можно создать модель, и эта модель будет количественно сравнивать э-э авторов по их э-э фрагментам их текстов с какими-то обобщёнными образами авторов. То есть он может она устанавливать авторство таким образом.

И я вам сказал, что это может быть естественный язык и искусственный. То есть это может быть э-э любой язык без разницы какой, там французский, немецкий, русский или английский, или китайский. Вот. И также э-э могут быть э-э созданы обобщённые образы различных языков программирования, к примеру. Вот. Или каких-нибудь систем кодирования, различных систем кодировки, например. И мы тогда сможем идентифицировать, допустим, кодировку. То есть система Eidos может определять, допустим, какая кодировка: ANSI 1252 или OEM 866 или ещё какая-нибудь там. Понимаете, да? Она это может вполне определять.

Вот. И может сравнивать э-э конкретные исходный текст программы на каком-то языке с разными обобщёнными образами разных языков программирования. И может определить, на каком языке программирования написана та или иная программа. То есть она прямо вам даст там Python там или C++ или C#. То есть прямо определит она, на каком языке программирования написана программа. Или JavaScript, или просто Java. Прямо будет определено, на каком языке написана программа. Вот. Но это для этого привести примеры написания программ на разных языках. Вот. И она это обобщит и выделит, э-э, чем они отличаются и будет определять язык программирования.

Вот. Кроме того, она может сравнивать сами обобщённые лингвистические образы классов друг с другом, создавать их кластеры, конструкты. Ну то есть, допустим, если мы берём разных учёных и смотрим на их работы, то система может определить э-э группы учёных, работы которых сходны, ну, скажем, по какой-то терминологии базовой. Я здесь такую работу проводил. Вот синтез семантических ядер научных специальностей ВАК. Видите? Что это значит? Это значит, что комиссия диссертационного совета, которая даёт задание советам определить, принадлежит ли диссертация, соответствует ли диссертация, представленная в совет, тем специальностям, по которым проходят защиты в этом совете. И, значит, делается соответствующее заключение. А вот я эту задачу решил автоматизированно, автоматизировал её. То есть берём текст диссертации или автореферата, загоняем в систему, она говорит нам: вот такая-то специальность больше всего похожа, вот такая-то похожа поменьше, а такие-то, такие-то вообще не похожи. Понятно, да? Вот. И то же самое, значит, можно сказать, что эта работа такого-то автора, такого-то писателя, такого-то учёного. И может получиться, что таким путём будет найдено, будут найдены работы, которые собой представляют плагиат из этой работы или с большим количеством заимствований из неё, из работ какого-то автора. Вот, допустим, мои вот эти работы, они все связаны с автоматизированным системно-когнитивным анализом текстов. Если кто-то возьмёт, тоже напишет работу по автоматизированному системно-когнитивному анализу текстов, то это будет, э-э, то есть это можно будет обнаружить, сравнивая исходный текст этой работы, статьи, с моими статьями. И такая работа, она проводится. Вот система антиплагиат выполняет сходные функции, поисковые системы выполняют сходные функции.

Также можно исследовать саму моделируемую предметную область путём исследования её лингвистической модели этой области, системно-когнитивной. То есть можно описать эту область какую-то, ну, допустим, место, где там что-то произошло, можно описать словами и потом сравнить это описание с описаниями других мест и идентифицировать, где же это было то, что там описано, определить место, где это описано. А можно исследовать, то есть сказать, что эти слова наиболее характерны для описания этого места, это менее характерны. И всё это можно делать для любого естественного или искусственного языка или системы кодирования.

Вот я вам сейчас показал источники информации, видеозанятия, то есть всё-всё показал вам.

Вот. И теперь продолжим дальше. Что-то вы, по-моему, не очень-то всё это выдерживаете, разбегаетесь потихоньку, да? Или не очень? Не разбегаетесь? Кто разбежался, поднимите руку, да? Кого нет, поднять руку. Вот.

Ну и продолжим, продолжим. Смотрим на сами исходные данные этой лабораторной работы 3.02. Что мы здесь видим? Мы здесь видим фрагменты текста о разных писателях. Эти фрагменты помещены в ячейки экселевской таблицы. Как это сделать, ребята? Значит, я вам описываю очень удобный такой комфортный, сервисный способ это сделать. Открываете вордовский файл. Вот, допустим, файл такой у нас есть система искусственного интеллекта. Это наш учебник, по которому мы занимаемся. Здесь очень много всего написано. Конечно, то, что там написано, это едва ли удастся мне всё вам рассказать. Я в таком сокращённом варианте несколько обычно всё это рассказываю.

Вот. И вот я, смотрите, значит, я беру, открываю Excel. И вы видели, я взял просто, то есть некоторый фрагмент текста вордовского, выделил блоком и помещаю в экселевскую таблицу. Что получилось у нас? У нас получился замечательный эффект. Смотрите, что у нас получилось, ребят. Вам видно, нет? Видно, что получилось?

Да.

Ну как это словами назвать, что получилось?

Получилось так, ребята, что каждый абзац занесён в отдельную ячейку новой строки. Вот что у нас получилось. То есть Excel так сделан, что он абзацы записывает в виде отдельных строк, помещая тексты из абзаца в отдельную ячейку таблицы Excel. Понятно, да? То есть эта операция автоматизирована, поэтому очень легко формировать исходные данные для работ научных и лабораторных.

Вот здесь что я сделал? Я взял сочинение школьное, посвящённое Достоевскому. Кусочек этого сочинения, не всё, а кусочек, страничку, включил в Excel. Написал здесь Достоевский, сдублировал сюда и протянул. А здесь написал Достоевский 1 и протянул. У меня получилось автоматически 1 2 3 4 5. Вот. Понятно, да? Также я сделал с другими писателями. То есть такую вот таблицу можно сделать за несколько минут. Можно использовать, конечно, не сочинения об этих писателях, а тексты самих этих писателей.

И мы здесь видим, что у нас одна колоночка классификационная – это сам писатель, фамилия, имя, отчество писателя. Одна описательная шкала, в ней фрагменты текстов школьных сочинений об этом писателе. И одна у нас не шкала, а просто это первая колонка является колонкой, где содержится информация о том, откуда взята информация в соответствующей строке. То есть это, допустим, третий абзац текста о Достоевском, к примеру, это второй абзац текста о Достоевском.

Значит, таких вот классификационных шкал и текстовых, и описательных шкал может быть много, ребят. Вот, скажем, может быть, у меня есть работы, вот там, где я вам показывал ссылочки на них, список этих работ, где в качестве классификационных шкал выступает не только автор, но и, скажем, жанр, э-э, или, значит, э-э, направленность произведения. Ну, допустим, там роман, вот, или поэма, там, или какой-то там фельетон там, или что там, понимаете? Вот, комедия, трагедия, то есть жанр, да? Вот. И, э-э, значит, э-э, в качестве описательных шкал тоже может быть не одна, а несколько. Таким образом, мы можем решать задачи, которые я вам перечислял, э-э, задачи не только формирования обобщённых образов писателей и идентификации текстов с этими писателями, но и определение года написания или жанра, или смысловой направленности работы.

И мы видим теперь э-э обучающую выборку. Значит, смотрите, у нас здесь вот, видите, фрагменты текста, а в правом нижнем окошке у нас коды слов, которые есть в этих фрагментах текстов. И вот здесь вот у нас таблица один ко многим, три таблицы. Их размер э-э жёстко не ограничен в системе. Вот, а то есть можно книжки анализировать, то есть просто книжки сравнивать друг с другом. То есть очень, то есть тексты очень большого объёма. Правда, в Экселе эти тексты в одной ячейке не может быть больше 32 К текста. Но это примерно шесть страниц текста. Ну тоже довольно-таки немало. Таких абзацев не встречается. Вот. Но я сделал интерфейс в системе 2 3 2 1, видите? Это интерфейс ввода данных из текстовых файлов. И в этом интерфейсе я обрабатывал большие файлы текстовые, в том числе книги прямо. То есть он позволяет это делать. И здесь есть другие интерфейсы, которые вот мы используем, будем использовать сейчас, и уже кое-что мы изучали, а кое-что ещё нет.

И вот мы увидели, какая у нас обучающая выборка. Теперь мы создаём модели. Модели. И смотрим, что у нас получится. Значит, модели получаются уже большей размерности, потому что у нас колоночки – это писатели, а строки – это слова, а их там около 800.

И мы можем посмотреть, какова достоверность моделей. Достоверность моделей многих моделей единица равна, видите? То есть это чрезвычайно высокая достоверность моделей. Почему так получается, ребята? Потому что все задачи обработки текстов можно отнести к переопределённым задачам. То есть для того, чтобы идентифицировать писателя, достаточно там трёх слов, например: Раскольников там или Лукоморье, или Базаров, и уже можно его идентифицировать, понимаете? Вот, или Мцыри. А, значит, э-э, у нас там сотни слов и тысячи слов. То есть у нас более чем достаточно для решения надёжного решения задачи идентификации. То есть задача переопределена, то есть объекты описаны сверхподробно, сверхдетально. Поэтому задача решается безошибочно. Модели получаются с очень высокой достоверностью. Ну здесь видно количество истинных положительных решений значительно больше, чем ложных. И истинных отрицательных решений вообще ложных отсутствует, только истинные решения отрицательные. Вот. А есть только небольшой фрагмент, когда очень низкий уровень сходства, где ложных положительных решений больше, чем чем истинных. Но это только при очень низких уровнях сходства. Что и говорит о том, что критерий сходства объекта с классом является адекватной мерой степени истинности решения.

Вот. Ну возьмём модель L3 и будем сейчас в ней исследовать. Ну, во-первых, посмотрим, зададим режим в режиме 5-6, зададим модель L3 наиболее, то есть той моделью, в которой будем проводить исследования. Проведём в ней решение задачи идентификации этой модели.

И посмотрим на результаты решения этой задачи идентификации.

Вот. Мы видим фрагменты, слева фрагменты текста, а справа обобщённые образы лингвистические разных писателей. И видим, что слева сейчас так полистаем и посмотрим. Видно, что всегда правильно идентифицируется фрагмент текста с писателем. И обратная, обратный разрез тоже это показывает.

Вот. То есть задача идентификации решена идеально. То есть это что значит? Что мы можем по фрагментам текстов, по фрагментам текстов точно определять авторство. И таким образом мы можем решать задачу атрибуции текстов, которая состоит в том, что мы определяем автора этого текста, который себя указал, не указал. Если он не указал, то это текст анонимный. Э-э, то есть мы определяем атрибуцию текстов – это определение истинного авторства. Вот. И в том числе, когда текст вообще не подписан или подписан псевдонимом, то есть подписан псевдонимом. То есть псевдоним. То есть это не его имя, каким-то чужим именем подписан. И вот такой подход позволяет элементарно совершенно определить истинного автора. Хоть он там подпишется 1 2 3, хоть он там подпишет там Ленин, он моментально определяется, что это, кто это такой, по самому тексту.

Вот. Теперь мы можем, то есть задачу идентификации конкретных текстов с обобщёнными образами мы решили. Теперь посмотрим, а что ж там за обобщённые образы у нас созданы? Это задача обратная задача идентификации, прогнозирования. Пушкин F3. Как вы думаете, какое слово, какие слова наиболее характерны для текстов о Пушкине? Руслан, Пушкин, Евгений, Людмила, Онегин. Вот. Кто-то что-то мне написал.

Вы просили напомнить о конце пары.

И сколько ж там до конца?

Минут восемь.

Понятно. Спасибо.

Вот. И также мы можем и других посмотреть писателей. Вот, что для них, какие слова для них наиболее характерны? Ну, например, для Тургенева, какое слово наиболее характерно? Базаров. Видите? Отцы там, дети, поколение. Понятно, да? То есть всё как бы ожидаемо. Но здесь что интересно? Что это не человек этот анализ проводит, а система искусственного интеллекта. И она может сравнивать также обобщённые образы друг с другом. Мы сейчас вам показал, что она сравнивала конкретные тексты с обобщёнными образами, показала нам, что это за обобщённые образы, то есть семантические ядра писателей, ядро, то есть то, что характерно, не характерно. А теперь мы также можем сравнить друг с другом целиком сами эти обобщённые образы. И таким образом мы можем сравнить писателей друг с другом.

Вот. И посмотрим на результат этого сравнения. Если мы возьмём когнитивную диаграмму, то они все окажутся непохожи друг на друга писатели. То есть они обладают ярко выраженной индивидуальностью, стилем своим. А если мы посмотрим на дендрограмму, то они хотя и все и непохожи друг на друга, но непохожи в разной степени. Одни непохожи в большей степени, другие в меньшей степени. Ну те, кто непохож в меньшей степени, можно сказать, что они в какой-то степени похожи, да? Вот. И вот сейчас мы посмотрим на них, на дендрограмму, которая показывает нам сходство, различие этих писателей друг с другом.

Вот. Мы видим, значит, что у нас мы видим? Пушкин на Лермонтова похож, Тургенев на Достоевского, Гоголь на Грибоедова, Толстой на Горького. И у нас две компашки. Одна компашка – это красные кластер: Тургенев, Достоевский, Пушкин, Лермонтов, а синий – Толстой, Горький, Грибоедов, Гоголь. Но при том, что Пушкин и Лермонтов похожи друг на друга, всё-таки они сильно друг от друга отличаются. То есть этот кластер образован на большом уровне различия вот этих классов, которые входят в этот кластер. Если бы они были очень похожи, тогда бы здесь вот было соединение вот здесь на низком уровне различия.

Понятно, да?

Смотрим теперь, э-э, какой смысл слов? Это называется семантический потенциал слова. То есть какой смысл слова?

4 4 9.

Вот, допустим, берём слово Онегин. Как вы думаете, какой смысл этого слова? А скорее всего, это Пушкин. И очень маловероятно, чтобы это был Толстой, Тургенев, Лермонтов.

Ну это вполне очевидные вещи, но смысл в том, что э-э система на основе созданных моделей определила смысл слов, который заключается в том, что в них содержится определённое количество информации об авторе текста, где встретилось это слово. Вот Толстой, он применял слово "война" в своих произведениях чаще, чем другие авторы. Другие авторы, может быть, и не применяли. Ну у нас здесь маленький примерчик такой. Вот. Если взять реальные их э-э произведения, большие, ну прямо книжки, то там получится, конечно, поинтереснее это всё.

Вот. Ну так вот мы видим тоже, что она вполне смысл слов определяет.

Вот. То есть мы сейчас с вами рассмотрели э-э решение задачи интеллектуальной обработки текстов. И решение на основе этого нескольких задач: задачи создания семантического ядра или обобщённого лингвистического образа писателя, жанра, времени написания. И потом решается задача идентификации текстов по жанрам, писателям и времени и другим параметрам, по смысловой направленности. Вот. И решается задача сравнения самих обобщённых образов друг с другом и определение смысла слов. И, в общем, и другие задачи, которые решаются э-э задачей классификации, кластерного анализа текстов или обобщённых образов решается. Но самих слов э-э кластерный анализ провести затруднительно, потому что их очень большое количество, ребят. При таких размерностях матрицы сходства считать, когда даже 700 на 700, это уже много уже. Уже проблематично это. Вот я сейчас такую задачу примерно как раз и решаю вот здесь вот. Здесь 470 э-э классов, а здесь да, 470 классов, она решается уже туговато. Правда, здесь около 30.000 примеров в обучающей выборке. Поэтому, ну, в общем, требуется время для решения таких задач большой размерности.

Какие есть вопросы у вас, ребята?

Никаких вопросов нет.

Вот, говорю, плохо владеть телепатией, потому что когда я вас спрашиваю, какие есть вопросы, я сразу моментально узнаю, какие у вас есть вопросы. У вас вопрос, когда бы это занятие закончилось, в конце концов, уже сколько можно. Да?

Нет.

Нет? Нет? Ну ладно, бывает всякое, бывает, я ошибаюсь немножко. Ну ладно, ребят, на этом занятие наше наше заканчивается. Вот. И через перерывчик будет занятие с другой группой у нас. С третьей, по-моему.

Хорошо, спасибо, до свидания.

Всего вам хорошего, до свидания.

До свидания.

До свидания.

До свидания.

До свидания.

Ах, ребята, девчонки, с наступающим праздником! Передайте всем, кто сейчас слышит, что остальным передайте, что поздравил вас с наступающим праздником. Хорошо?

Хорошо. Спасибо большое.

Что-то как-то я вообще это самое, запреподавался. Желаю вам здоровья и счастья. Всем-всем, пожалуйста, передайте.