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

***Российская Федерация***

**40 Как интеллектуальные технологии применяются для интеллектуального анализа текстов и для чего это нужно. 2020-11-14**

Заголовок: Система Eidos: Интеллектуальный анализ текстов и повышение формализации моделей для принятия решений

Резюме

Лекция профессора Луценко посвящена применению интеллектуальных технологий, в частности системы Eidos, для анализа текстов в контексте системного анализа и принятия решений. Подчеркивается, что любое решение основывается на модели, даже если она не осознается или не формализована. Вербализация (описание словами) рассматривается как первый этап формализации, который, однако, сопряжен с неизбежной потерей информации. Язык является универсальной системой моделирования, и система Eidos способна обрабатывать текстовые (вербальные) модели, повышая степень их формализации до статистического, информационного и аналитического уровней. Это позволяет решать различные задачи, включая идентификацию и прогнозирование.

Система Eidos увеличивает системность моделей, что важно для адекватного отражения сложных объектов (согласно принципу Эшби). Профессор объясняет концепцию лемматизации (приведение слов к базовой форме) на примере словаря Зализняка и демонстрирует возможности Eidos: формирование обобщенных лингвистических образов классов объектов на основе текстовых примеров, сравнение конкретных текстов с этими образами, кластеризацию текстов, исследование предметной области через ее лингвистические модели и интеллектуальную атрибуцию текстов (определение авторства, датировки, жанра).

На практическом примере анализа фрагментов текстов русских писателей показано, как Eidos с высокой точностью определяет авторство и выявляет семантическое ядро (наиболее характерные слова) для каждого автора. Упоминается возможность использования системы для анализа данных с платформы Kaggle, например, для определения возрастной группы читателей детских книг по тексту. Несмотря на эффективность формализации, подчеркивается, что процесс коммуникации через язык опосредован и зависит от индивидуального восприятия и состояния говорящего/пишущего. В ходе лекции также кратко обсуждается возможность преподавания курса по интеллектуальным информационным технологиям и упоминается недавняя кончина коллеги. Следующая лекция будет посвящена методике риэлторской оценки.

Детальная расшифровка текста

I. Введение и Контекст

Так, здравствуйте, ребята.

Здравствуйте.

Здравствуйте.

Здравствуйте.

Здравствуйте.

Здравствуйте.

Так. Сегодня у нас 14 ноября 2020 года. И что интересно, это суббота. И уже четвертая пара, которая идет 13:10-14:40 по дисциплине системный анализ и принятие решений.

Занятие ведет профессор Луценко Евгений Вениаминович.

II. Обсуждение предыдущих тем и проверка знаний

Давайте всё-таки я задам вам ещё один вопрос. А принятие решений как, вернее, цель, цель управления повышение уровня системности как цель управления, об этом я рассказывал вам или нет? Повышение уровня системности как цель управления?

М-м-м.

Там про таксистов, про лёт там, про...

Про таксистов, да, вы рассказывали.

А, ну тогда понятно, да.

III. Введение в интеллектуальный анализ текстов с использованием системы Eidos

Тогда давайте сейчас расскажу вам о том, как можно интеллектуальные технологии применять для интеллектуального анализа текстов. И для чего это нужно?

A. Предоставление ресурсов (ссылки)

Значит, прежде всего, давайте я вам расскажу о том... Значит, ну сначала дам вам ссылочку на мои работы по этой тематике. Вот. По интеллектуальному анализу текстов, а потом уже...

Кстати, давайте сейчас я вам открою экран. То я что-то рассказываю, а вы не знаете, что я там рассказываю. Увидели, да? Вот я запустил запустил систему Eidos. И вот здесь вот у меня есть ссылочки на подборки публикаций моих по разной тематике. В частности, вот есть ссылка по АСК-анализу текстов. Я вам эту ссылочку сейчас отправляю в чат, для того, чтобы вы могли эти работы посмотреть, ознакомиться с ними.

B. Основные концепции: Модели, формализация, вербализация

И в двух словах вам хочу рассказать о том, что есть модели различные, и эти модели можно классифицировать по степени формализации. И, значит, начнём с того, что всегда все решения, у нас же дисциплина это всё-таки принятие решений сейчас, системный анализ и принятие решений. Все решения всегда принимаются на основе модели. Даже в том случае, если эта модель нам неизвестна и неизвестен сам процесс, сам механизм, алгоритм принятия решений.

Вот, допустим, иногда мы вроде как спонтанно принимаем какие-то решения, особенно мелкие, там повернём направо, налево, там, пойдём куда-нибудь там и тому подобное. Вот возникает вопрос: а с какого перепугу мы принимаем те или иные решения? Почему мы вот так действуем? Я могу сказать, что у нас есть некоторая модель реальности, внешней, внутренней. Есть определённые цели, ценности и мотивации, мы стремимся к их достижению по различным причинам. Ну, допустим, нам просто хочется достичь определённого состояния, определённого результата. Что это такое, когда нам что-то хочется, это трудно объяснить. Сейчас нам даже это и не ставлю такой задачи. Вот. Иногда бывает не так просто достичь этого результата, бывает нужно приложить определённые волевые усилия или там физические усилия. Ну, волевые, они в основе лежат всех наших усилий, всех видов наших усилий: физических, и умственных, и других.

Вот. И если спросить человека: "А почему ты вот решил это делать?" Вот. Он не всегда даже может и объяснить. И если спросить его: "А почему ты..." Ну, можно сказать, допустим: "Мне вот нужно добиться такого-то результата". Мы его спрашиваем: "А почему ты таким способом решил это сделать? Почему ты таким способом решил добиться этого результата?" И он скажет: "Ну как, я вот так понимаю, что вот таким способом, наверное, можно было бы этого этого это достичь, этого, ну, добиться этого". Ну, то есть люди толком не могут объяснить, как они принимали то или иное решение.

Ну вот всё это, то, что я сейчас сказал, э-э, это можно научно сформулировать так, что часто решения принимаются неформализуемым путём. Сейчас я объясню, что такое неформализуемым путём. На основе опыта, интуиции и профессиональной компетенции. Вот. Ну, то есть, что значит неформализуемым путём? Это как раз и означает, что модель используется неформализованная. То есть мы не можем описать даже, как мы это, почему мы так поступили.

И вот здесь я хочу сказать, что описание словами – это самый первый этап формализации. Это и есть первый этап формализации.

Ну, так этот выражение словами называется вербализация. То есть я бы сказал так, научно: вербализация – это первый этап формализации модели. И при этом мы иногда не можем и словами сформулировать, что мы там, нашу причину нашего поведения. Тогда есть такое высказывание хорошее, что в русских сказках народ, э-э, там тысячи лет назад, не знаю, когда, э-э, сформулировал, что ни словами сказать, ни пером описать. То есть есть такие вещи, которые даже не поддаются вербализации, то есть даже словами не опишешь. А интуитивно мы это понимаем и применяем, и действуем соответствующим образом.

Я могу вам сказать, что если так вот посмотреть на объём вообще осознаваемого, и какая доля из того, что мы осознаём, может быть описана словами, то я так думаю, что надо э-э очень критически к этому относиться и думать так, что, ну, несколько процентов, наверное. Вот. Э-э, очень многие вещи мы вообще не описываем словами, не можем описать, нет у нас терминов или нет навыков. Вот есть люди профессиональные, которые, э-э, в профессиональную компетенцию которых входит описание словами своих мыслей. Ну, это учёные, конечно, вот, э-э, проповедники, ну, священнослужители, и писатели. Вот они выражают словами то, что они чувствуют, то, что они думают. Э-э, учёные в одном стиле там аргументируют, э-э, эти в другом стиле, апеллируя к чувствам, там, к образам, э-э, ну, в общем, к духовным ценностям. То есть сам-то стиль-то разный, и цели разные, но это люди, которые обладают развитой способностью вербализации. То есть вот его спрашиваешь: "А почему?" Он может тебе объяснить. То есть у него есть соответствующие компетенции, соответствующие умения, навыки. То есть он может это делать, понимаете? Аккуратно, то есть методически правильно со своей точки зрения, в той области, где он работает.

C. Язык как система моделирования и роль Eidos

Так вот, язык, ребята, является очень универсальной системой моделирования. По сути дела, мы очень многие вещи всё-таки можем описать словами. И когда мы это делаем, то мы можем сказать так: строим вербальную модель того, что мы описываем. И система Eidos обладает такой способностью, у неё есть такие возможности, реализованные в ней, что даже такого вот очень невысокого уровня формализации предметной области хватает для того, чтобы она эту, эти данные на таком уровне формализации подхватила, повысила их степень формализации и создала модели, на основе которых можно потом решать задачи.

D. Системность и принцип Эшби

Сейчас я сошлюсь на статью, э-э, дам вам ссылочку на статью, в которой как раз описывается процедура повышения формализации модели. Система Eidos – это система, которая последовательно повышает степень формализации модели и уровень системности модели. А уровень системности, когда повышается, то это означает, что модель становится более пригодной, более адекватной для отражения более сложных объектов моделирования.

Я вам рассказывал про принцип Эшби, наверное, нет? Связанный с познанием сложных систем.

Эшби? Нет.

Что?

Ну его как-то назвали, но...

Я Уильям Росс Эшби, принцип Эшби. Что уровень системности модели должен быть выше, чем уровень системности объекта познания. Иначе она неадекватно его отражает этот объект познания. Ну, допустим, если модель у нас двумерная, а объект познания трёхмерный, то модель познания, модель объекта познания будет отражать только его проекцию, а не трёхмерную модель, 3D. Про это рассказывал, нет?

Наверное, не помните, да?

E. Потеря информации при формализации

Ну ладно, всё ж не расскажешь. Хотя по очереди постепенно, может быть, можно и попытаться. Так вот, э-э, система Eidos, она обеспечивает э-э обработку текстовых моделей объекта познания и повышение степени формализации модели до э-э статистических, информационных, аналитических моделей, более высокого уровня системности, и позволяет на этой основе решать ряд задач.

Значит, э-э, прежде всего, э-э, хочу сказать вот что, что когда мы вербализуем какую-то идею, которая словами не выражена или интуитивная модель, то при этом происходит необратимая потеря информации. При каждом повышении степени формализации происходит необратимая потеря информации.

Вопрос возникает такой: зачем мы вербализуем модели интуитивные? Вот я что-то понимаю, ну и понимаю, и хорошо. Зачем мне это описывать словами? Вот зачем сейчас я описываю словами то, что вам собираюсь рассказать на лекции? Вот я понимаю то, что хочу рассказать? Да, я понимаю. Оно у меня на интуитивном уровне всё это реализовано.

F. Необходимость вербализации для общения

Кстати, в какой форме и в каком месте в пространстве и времени, локализация пространства и времени, и в какой форме представления реализованы или интуитивные модели, никто не знает. Но обычно говорят, что это что-то связано с головой, но я не очень в этом уверен, даже я скажу так, я вообще уверен, что это не так. Потому что есть внетелесный опыт Out of Body, и есть э-э изменённые формы сознания, в том числе высшие формы сознания. Есть опыт клинической смерти. И вот весь этот опыт, и то, и то, и то, вот эти три направления, они подтверждают, что человек может мыслить, чувствовать, прекрасно, адекватно всё осознавать, и очень чётко, гораздо чётче, чем сейчас, кстати, при этой форме сознания, когда у него или вообще нет тела, или когда оно есть, но не действует. Ну, скажем, в состоянии клинической смерти. Или оно есть, но не используется для того, чтобы осознавать окружающую реальность. Ну есть там и спит, например, лежит на кровати там, ну и себе лежит. А в это время человек может осознанно действовать на других планах реальности, и у него огромные возможности информационные при этом, на порядке превосходящие возможности в обычной форме сознания.

IV. Телефонный разговор (возможность лекции, личные новости)

Профессор отвечает на телефонный звонок.

Здравствуйте, Валерий Евгеньевич, здравствуйте.

Здравствуйте, Евгеньевич.

Здравствуйте.

Я здесь.

Меня хорошо слышно? Это Коржикова.

Да я вижу прекрасно. Здравствуйте, Валерий Евгеньевич. Я сейчас занятие веду вообще-то, ну так, если коротко.

Хорошо. Когда с вами можно поговорить?

Ну сейчас я скажу конкретно. После, ну в общем, около 5 часов. У меня подряд ещё две пары.

Я понял, около 5 часов. Ну, суть дела: хочу предложить вам выступить лектором на курсах по интеллектуальным информационным технологиям, которые, ну, в составе преподавателей нашей кафедры АСУ и О проводятся. Ну, там участвует человек где-то М8, чтобы подключить. Ну, уже объявил.

Ну это интересно.

Интересно?

Да. Это интересно.

Ну всё, я... Бучацкий сейчас заведующий кафедрой, я ему сказал. Вот Евгения Вениаминовича я его спрошу, как он согласится. Ну, он говорит: "Хорошо, я согласен. Поговорите".

Ну он меня знает вообще-то. Он же был аспирантом у Симанкова. Поэтому он меня...

Да-да-да. Он меня видел.

Да он так и сказал: "Я его хорошо знаю". Ну а что, говорит, поговорите.

Ну, насколько хорошо, настолько хорошо я не знаю, но бегали там два парня, вот он и ещё один друг его там.

Ну да, ещё Шопин. Он тоже будет, в принципе, сейчас там на этих как бы курсах.

Ну, ну, конечно, мы друг друга видели и так здоровались там, вот так вот, на таком уровне. Ну хорошо, Валерий Евгеньевич, мне это приятно, что вы мне это предложили, спасибо.

Там у вас так слухи прошли, я не знаю, как Слоика, что там, какие проблемы?

Он умер.

Ой, мои соболезнования. Я тоже такое слышал.

Он 19 октября умер.

Ай-яй-яй. У него что, связано с вирусом, да?

Ну, никто не говорит толком, но так вот, как мне сказали, вроде бы у него было сердце и задыхаться он стал, и попал в реанимацию без сознания. 3 дня пролежал и умер.

Понятно.

Так мне объяснили. Это Татьяна Петровна. Татьяна Петровна так сказала. Так что очень жаль, конечно, это человек был эпоха, эпоха, наша эпоха закончилась с ним.

Да, это верно. Ну что, ещё раз соболезную, конечно. Хорошо, Геннадьевич. Значит, на 17 часов мы как разговор откладываем или состоится с вами?

Ну я позвоню, потому что ж детали ж важно уточнить. Что, как...

Хорошо, договорились.

Что, как, сколько и за сколько?

Хорошо. Давайте.

Хорошо. Договорились. В 17 часов, значит, я жду звонка вашего. По Ватсапу можно, да?

Ну, можно просто по телефону.

Можно просто. Хорошо.

WhatsApp - это если где-нибудь за рубежом, а так вот здесь какая разница?

Ну, я понял. Хорошо.

Угу, ну давайте.

Жду тогда от вас звонка. Хорошо?

Хорошо, спасибо.

До встречи, до свидания.

До свидания.

Профессор возвращается к лекции.

Вот. Пригласили работать в университет ещё один. Ну хоть ненадолго, правда. Значит, там провести курс.

V. Возобновление обсуждения: Проблемы вербализации

Ребят, напомните, на чём мы остановились, пожалуйста. А то сбили меня немножко с толку.

Какой-то Эшби... понял. Какой-то... Забыл уже, как опять называется.

Господи. Вот так вот во время занятия отвлекают преподавателя.

А до этого что я говорил? Скажите, напомните ход мысли.

Ну вы нам скинули ссылку АСК анализа, потом ссылку...

На... повышение уровня формализации модели. Вот что я хотел сказать, ребята, что каждое повышение уровня формализации сопровождается необратимой потерей информации. Вот что касается выражения в словах наших идей, то здесь есть слова Тютчева, что мысль произнесённая есть ложь. Есть слова э-э этих дзен-философов и дао-философов. То есть это дзен - это японская, а дао - китайская философия. Э-э, у дао говорят так: что такое дао описать словами нельзя, э-э, произнесённое дао есть не есть дао. То есть если мы начнём описывать дао, то тогда мы его исказим. То есть наши слова непригодны для адекватного описания дао, того, что такое дао. Вот я не знаю, когда я сейчас вот это сказал, то я исказил дао или нет? Наверное, уже начал искажать. Подкрадываться к тому, чтобы его исказить, потому что я уже о нём довольно много сказал. Вот. Ну, сказал, что нельзя сказать там и так далее. Понятно, да, это такой немножко юмор в этом есть какой-то. Вот. Ну я скажу так, что э-э чем выше объект познания, чем он сложнее, э-э, чем чем больше он от нас отличается, скажем так, тем он сложнее его описать. И наш язык, он вполне неплохо себя проявляет, когда мы описываем, так сказать, бытовые всякие вещи вокруг нас, которые. Но когда мы начинаем познавать там какие-то глубокие истины, то он начинает буксовать прямо явным образом. И возникают парадоксы различные, и не можем мы выразить мысль там и так далее, и так далее. Хотя есть люди, которые, ну, мастера в том, чтобы выражать свою мысль. Это я сказал, учёные, священнослужители и писатели, и поэты особенно. Ну и писатели вообще. Вот. Конечно, есть гении в этой области. Ну как вот любая способность, способность вербализации, выразить мысль в словах - это способность есть у всех, но она у всех на таком среднем уровне, ну как обычном, скажем так. А есть выдающиеся представители, у которых она блестяще развита эта способность. Вот этих вот людей, у этих направлений деятельности, ну, как тоже, у кого лучше, у кого хуже, но в общем, вот как-то выражено это заметно.

Вот. И для чего мы вообще это делаем? Для чего мы выражаем мысль в словах? Вот сейчас, для чего я выражаю свою мысль в словах? Как вы считаете? Вот мог бы я провести занятие, не выражая свои мысли в словах?

Нет, мы не можем услышать ваши мысли.

А жаль. Хотя, может быть, наоборот, к счастью. А то бы сейчас такое услышали. Значит, ну это я так подшучиваю немножко. Вот. Ну, э-э, конечно, значит, я вам могу сказать, что э-э наше сознание сейчас таково, и ваше, и моё, что мы можем осознавать как объективно только физический уровень реальности. И, соответственно, мы можем осознавать слова, которые каким-то образом на этом физическом уровне реальности воплощены. То есть мы мысль нашу преобразуем в такую форму материальную, я бы сказал так. Вот мысль, я не знаю, в какой там она форме, внутренняя форма представления. Ну, знаю, но не скажу. Вот. Значит, мы осознаём только этот физический мир практически. Поэтому, э-э, как объективно, я хочу подчеркнуть. То есть вне себя, вне нашего сознания, мы видим объекты, да? Причём, э-э, вас я тоже вижу как объекты. Как это ни парадоксально. Хотя я понимаю, что вы личность, что у вас там психические свойства есть, там сознание, самосознание. Это всё, конечно, понимаю. Но я это понимаю интеллектуально, а вижу-то я объект. Я вижу перед собой объект симпатичный. Ну я так смотрю на некоторых, очень даже. Вот. Ну я хочу сказать так, тоже так, это небольшой комплимент девочкам. Вот. Так вот, э-э, вы понимаете прекрасно, что сейчас мы не осознаём мысли друг друга, да? То есть у нас нет телепатии, способности к телепатии. Вот Александр это чётко совершенно выразил. Вот. А была бы, тогда бы не надо было вот это всё рассказывать, понимаете? Вы бы прекрасно бы восприняли мои мысли, мгновенно причём. Весь объём сразу раз и перекачали туда, там, как вот с флешки там, да? Есть довольно быстродействующие. И на этом бы обучение закончилось. Вот. То есть вы бы имели прямой доступ к мыслям, э-э, к знаниям. А так получается, что это опосредованный доступ. То есть мы должны выразить слова. Вот я должен в словах выразить. При этом у меня получается так себе. Ну, иногда так более-менее получается, иногда не очень. Это связано с тем, ну вы и сами тоже это замечали, наверное. Ну, допустим, я в нормальном состоянии или я уже замученный? Если у меня, допустим, вот сейчас вот у вас четвёртая пара, для меня это, для меня это третья пара. То есть я провёл вторую и третью. Ну что, то есть, ну да, то есть это для меня третья пара. Ну, третью пару я ещё так более-менее. Четвёртую мне уже тяжелее. А пятая, шестая, седьмая, восьмая пара - это вообще полный ужас какой-то, понимаете? Вот смотрите, вот у меня, допустим, вторник, э-э, первая пара, да, вторая, вторая, третья, четвёртая, пятая, шестая - это всё вторник. А потом смотрите, здесь, правда, нету, слава богу. А здесь ещё есть седьмая и восьмая пара. Вот первая, вторая, шестая, а потом смотрите, седьмая и восьмая. Каков я на восьмой паре? Как вы считаете? Как я вообще себя чувствую? Да я спать хочу. Откровенно вам скажу, я просто хочу спать. Я устал уже, понимаете, за день. Ну и как я буду рассказывать? Конечно, я буду хуже рассказывать, понимаете? Ну как, я буду стараться, конечно, там всё, но у меня будет получаться хуже.

Вот. Так вот, э-э, это значит, что вот эта способность вербализации, она у нас не постоянная, она у нас то выше, то ниже этот уровень наших возможностей в этой области. Иногда получается получше, иногда похуже, иногда более чётко, иногда менее чётко, иногда, так сказать, вам интереснее слушать, иногда как-то скучновато вроде как слушать даже. Но мы вынуждены вербализовать наши мысли для того, чтобы передать их друг другу. При этом мы можем это сделать в звуковой форме, а также и в других формах языковых, формах представления. Хотя вот мы говорим, что слова, да? Но слова мы можем просто вот произнести, причём на разных языках это можно сделать. Вот. А потом мы можем и, в общем-то, их и записать эти слова. Вот, с помощью какого-то алфавита или каких-то символов типа иероглифов. Вот. Причём иероглиф он неоднозначное выражение в словах имеет. А, вот я про Дао говорил, что есть вещи, которые вообще нельзя описать словами. Ни словами сказать, ни пером описать. Э-э, то есть наш язык, он не как, имеет ограниченные возможности тоже. Как и, в общем-то, и средства изобразительные. То есть если мы, допустим, э-э, хотим изобразить что-то, вот, то, попробуйте вот сейчас взять и изобразить портрет э-э кого-нибудь. Ну, допустим, там, даже вот самих себя там портрет или ещё кого-нибудь. Я вам скажу, получится так себе. Ну, не так, как у выдающихся художников. А если мы попытаемся изобразить что-то такое более сложное, то у нас получится хуже. Степень адекватности нашего изображения будет ниже. Оно хуже будет отражать оригинал, понимаете? Потому что это сложный объект для изображения. Вот. А если он будет ещё сложнее, ещё выше и непонятнее для нас, тогда мы ещё хуже его изобразим. Ну, в том смысле, что нам-то, может быть, будет казаться, что нормально изображено, а вообще-то он будет, это изображение будет хуже отображать объект, который мы изображали. То есть степень адекватности ещё ниже будет. А если мы возьмём самый высокий объект, какой-то сверхвысокого уровня, э-э, духовного какого-то, попытаемся его изобразить, то у нас практически ничего не получится. То есть мы нечто изобразим, но оно будет плохо изображать эту духовную сущность высшего порядка. Э-э, и даже можно даже так сказать, что возможно, оно даже будет его как-то и дискредитировать или искажать даже, понимаете? То есть есть некоторые вещи, которые даже, может быть, и не надо делать. Потому что это не хорошо получится, а плохо. И вот, скажем, иконы, например, иконы. Вот иконы, их там целые школы есть этих иконописцев, традиции определённые, каноны, в которых они это делают. А вот взял кто-то там нарисовал сам э-э икону и, не умея рисовать, не зная канонов, он такое может нарисовать, что это вообще-то будет даже, ну, своего рода такое святотатство даже, понимаете? Какой-то шарж получится, может быть. Ну я просто так говорю, что не всякое изображение будет адекватным, и в некоторых случаях эти изображения являются оскорбительными, например, для кого-то. Понимаете? Хотя человек, может быть, и не хотел оскорбить, он просто он такой вот, он неважный, не очень художник, он талантлив, не очень у него хорошо это получается. Ну тогда, значит, не надо и браться за это, если не получается, понимаете? Чтобы никого не обижать. Это вот я сейчас рассказываю то, что в сказке про Незнайку было. Помните, когда Незнайка э-э стихи написал про всех, а потом, а потом взял, стал художником и нарисовал портреты всех э-э этих э-э человечков этих маленьких. И все обиделись на него. Помните, нет, такую сказку?

Ну я помню только Незнайку на Луне и всё.

Ну вот. Ну, тоже хорошая. А до этого там было, что он решил стать поэтом, решил стать художником и сочинил стихи такие обидные про всех, примитивные. А потом взял и нарисовал ещё такие картинки про всех, что все на него обиделись вообще. Потому что он не умел этого делать так, чтобы это людей могло, ну, как бы они нормально это восприняли, вот так скажем. Ну и ему сказали: "Лучше ты этим не занимайся, лучше там занимайся тем, чем ты можешь заниматься".

VI. Возможности системы Eidos для анализа текстов

Так вот, ребята, когда мы что-то вербализуем, то есть описываем словами, всё равно мы как-то это делаем ущербно. Вот. То есть есть потеря информации необратимая. Потом следующие уровни формализации, более высокие, они тоже вызывают необратимую потерю информации. Чем меньше формализована модель, тем более она содержательна. Но тем сложнее её передать. Вот я, допустим, интуитивно что-то понимаю, а вам это передать тяжело мне. Чтобы мне это сделать, мне нужно, вот видите, словами всё описать. Потом вы эти слова воспринимаете, потом у вас восстанавливается понимание своё. А своё понимание, которое у вас возникает, оно основано на том, как вы понимаете мои слова, которые я говорю, какое у вас э-э мировоззрение, подготовка, э-э опыт жизни, может быть, там и так далее, и так далее, понимаете? Я думаю, что вы, конечно, меня понимаете, так, в принципе, но не совсем, понимаете? Так вот честно. Ну не на 100%. И я так думаю, что то, что я хочу до вас донести, тоже мне удаётся это сделать не на 100%.

Вот. Ну теперь давайте вернёмся к нашим баранам и посмотрим, а что же мы можем, э-э, какие задачи мы можем решать э-э путём анализа текстов.

A. Формирование обобщенных лингвистических образов

Значит, вот я нажал, если вы заметили, в системе Eidos на кнопочку, которая появляется, когда мы заходим в систему, первое окошко запуска системы, у нас уже здесь есть кнопочки подборки публикаций по разной тематике. И вот здесь вот есть подборка публикаций по анализу текстов. Мы можем формировать лингвистические обобщённые лингвистические образы классов. То есть у нас есть на основе примеров текстов, относящихся э-э к этим классам. Причём эти тексты могут быть на любом языке, на котором они могут быть записаны с помощью э-э латиницы и кириллицы. Вот это я подчёркиваю. То есть мы не можем на арабском языке обрабатывать, на китайском, японском, на хинди, э-э, суахили, там, где вот используются эти пиктограммы какие-то, иероглифы, мы не можем эти языки использовать. На грузинском не можем. Э-э, то есть те языки, которые используют э-э другие алфавиты, мы их не можем использовать. Почему? А потому что их нет TrueType fonts. Национальной страницы такой нету, понимаете, TrueType fonts. В этом, в стандарте OEM 866 кодирования. Так вот, значит, этим только и ограничено. Так вот таких языков, ребят, которые мы можем использовать, 50, кроме русского. То есть русский и ещё 50 разных языков. То есть это все основные языки мира. Ну, кроме китайского, пожалуй.

B. Сравнение текстов с образами

Второе, что мы можем сделать, мы можем сравнить количественно э-э конкретный э-э текст э-э или описание объекта или процесса с помощью слов с обобщёнными лингвистическими образами классов.

C. Кластеризация

То есть очень интересный момент, ребят. Вот если я описываю некую местность, помните, там были "Джентльмены удачи", и там Крамаров, они там искали место какое-то определённое, где там кого-то он встретил, я не помню уже. И там было дерево, вот такое вот дерево вот с такими вот ветвями. Вот они искали, искали по городу, ездили на машине, потом говорят: "Вот это дерево!" и показывает э-э витрину магазина, а там пальма. И, короче, в конце концов нашли всё-таки это место. Но я хочу что сказать? Что можно описать различные объекты, процессы, явления словами, и система Eidos способна это уже использовать для создания образов, анализа, поиска, э-э, ну, и, в общем, выполнения различных, решения различных задач.

D. Исследование предметной области

Вот я сказал, сравнивать обобщённые лингвистические образы друг с другом можно сравнивать. Значит, конкретные можно сравнивать с обобщёнными, а также обобщённые друг с другом. И формировать кластеры и конструкты текстов, лингвистических образов. Потом можно исследовать моделируемую предметную область путём исследования её лингвистической системно-когнитивной модели.

E. Атрибуция текстов

Проводить интеллектуальную атрибуцию текстов, то есть определять вероятное авторство текстов, анонимных и псевдонимных, датировку, жанр, смысловую направленность. Что такое анонимные, псевдонимные, э-э, атрибуция вот эта? Атрибуция - это определение авторства. Э-э, вот. Э-э, что это под этим понимается? Значит, текст написан э-э тем человеком, который там подписан в качестве автора или нет? Или некий анонимный текст есть, вообще автора там никакого не написано внизу, а или вверху. А, значит, нужно определить, а кто же этот текст написал? Или написано там, допустим, Ленин, э-э, подпись. А мы этот текст анализируем, а написал его не Ленин, а Ульянов. То есть это что значит? Это ошибка или нет, ребята? Это не ошибка, просто текст псевдонимный. То есть он подписал не Ульянов, а Ленин. Э-э, и всё. И потом стал известен как Ленин больше, чем как Ульянов. Также можно определить, э-э, когда примерно написан текст, в каком жанре он написан по этому тексту. И всё это можно делать для любого естественного или искусственного языка или системы кодирования. Что значит искусственного языка? Ну, берём, скажем, эсперанто. Можно для эсперанто это сделать. Точно так же, как для польского, там, украинского, русского, латинского, там, какого угодно, немецкого, французского, испанского, какого угодно языка. И для эсперанто это можно сделать. Это искусственный язык, разработанный учёными. А можно сделать для э-э C++, C#, э-э, просто C языка. Вот, а также для других языков программирования. И определить, а чем отличается C от C++ и от C#? Какими командами? Вот это можно определить с помощью системы Eidos, понимаете, ребята? То есть мы можем ввести эти тексты на этих языках, э-э, то есть исходные тексты программ. Она их проанализирует и решит эти вопросы.

VII. Практический пример: Настройка лаборатории Eidos

Значит, здесь вот, ребята, описаны статьи, приведены ссылки на статьи, э-э, довольно-таки есть серьёзные статьи у меня по этой тематике. Я длительное время этим занимаюсь, но статей написал не очень много, но с 2003 года у меня статья первая, 2003 года. Э-э, здесь смысл этой статьи, знаете какой? Что обработка текстов как решение обобщённой задачи идентификации и прогнозирования. Всё дело в том, что сам язык содержит уже средства моделирования окружаю... э-э, объекта познания, э-э, очень довольно развитые, в общем-то. То есть вот наша грамматика языка, она позволяет описать, что там, описывать объекты позволяет существительными описывать, процессы глаголами позволяет описывать, то есть динамику изменения этих объектов. Позволяет описывать их свойства. И позволяет это всё использовать для обобщения, например. Вот мы, допустим, можем сказать: вот по улице проехала машина. Да? А можно сказать, а вот это, то есть машина - это существительное, проехала - это глагол, и всё. А потом меня спрашивает Казбек: "А что за машина?" А я говорю: "Ну как, ты что, не видишь, что ли? Вон красная Мазда проехала". Он говорит: "Тройка, что ли?" Я говорю: "Ну да". Тройка. То есть мы определили марку и модель. И ещё признак один указали, что это красного цвета или такого тёмно-малинового цвета машина проехала. То есть наш язык вполне приспособлен для того, чтобы моделировать окружающую реальность и также и внутреннюю нашу реальность, психическую. И вот всё это дело мы можем использовать для того, чтобы описывать объекты, процессы и, в общем, э-э, потом это... То есть я хочу сказать, что это тоже моделирование и тоже позволяет решать ряд задач.

Вот сейчас мы решим с вами лабораторную работу э-э третьего типа, вторую лабораторную работу, которая предназначена именно вот для того, чтобы в элементарной форме вам это показать. Вот мы эту работу устанавливаем. Эту экранную форму вы, наверное, уже видели, ребята. Эту экранную форму. Сейчас я буду вам в чат посылать информацию. Надо было даже раньше послать. Вот. Берём вот так прямо и начинаем посылать. Правда, это время занимает. Вот. Ну ладно.

И э-э переходим мы в режим установки лабораторной работы 1.3. Устанавливаем работу работа с текстами.

Кстати, на Кагле есть интересное приложение э-э анализ э-э детских книжек э-э для разных возрастов, детей разных возрастов. Вот, скачиваем набор данных. Он попадает у меня в папочку Downloads. Что-то он не попадает, вот попал. Вот. Э-э, берём в папочку Input Data, размещаем этот набор данных. Преобразуем его имя в indata, потому что иначе система не будет воспринимать его. Потом идём, знаете куда? Идём, ищем онлайн-конвертер. Вообще в системе Eidos есть свой CSV конвертер. Вот, но э-э я рекомендую... Я, честно говоря, знаю про то, что этот конвертер в системе Eidos, он в данном случае не совсем подходит. Иногда он очень хорошо себя проявляет. И мы заходим в папочку Input Data и конвертируем. Вообще он в любом случае, чем в любом случае лучше, чем экселевский, потому что экселевский дата, то есть числа заменяет на даты некоторые. По крайней мере, мой конвертер этого не делает.

VIII. Запуск анализа Eidos и интерпретация результатов

Смотрим, ребята, что у нас получилось. У нас получился экселевский файл, в котором возрастная группа, текст и само произведение. И всё. Вот. Причём сам текст может быть довольно большим, ну, сравнительно так. Вот. Это как такая аннотированная описание этой книжки. Что мы делаем? Мы делаем то же самое, ребята. Только мы теперь уже не как лабораторную работу это устанавливаем, а конвертер запускаем э-э API, автоматизированный программный интерфейс, и запускаем уже прямо из режима. И здесь вот мы смотрим, что у нас те же самые параметры, видите, заданы, ничего мы не меняем. А вот это CSV конвертер так работает. И здесь я рекомендую, в том числе онлайн, если не получается. Причём конкретно онлайн, который я уже много их видел и изучил их возможности. И мы дальше осуществляем то, что мы уже делали в лабораторной работе, но уже мы это делаем не в лабораторной работе, а уже мы это делаем на реальном примере с Кагла. Который значительно большего объёма, значительно большего. Вот. Книжки, значит, возраста 59 у нас возрастных категорий. Вот, а слов у нас уже не 767, а 3.585. Видите, да? И слова на английском языке. Дальше мы осуществляем синтез и верификацию моделей на графическом процессоре. И у нас, ребята, не восемь, то есть у нас там было не восемь, а было, сейчас я не помню сколько. У нас э-э фрагментов текста было восемь писателей, а фрагментов текста было штук 30-40, по-моему. А здесь 430. То есть здесь на порядок больше и классов, и объектов обучающей выборки. То есть, в принципе, вот эту модель с этими вот книжками детскими можно использовать вместо той лабораторной работы, но я этого не делаю. Знаете почему? Расчёты дольше будут длиться, и так быстро вот не покажешь ничего. Но сразу после неё можно показывать уже реальный пример, совершенно реальный пример, который покажет нам, какие слова, каким образом связаны с теми или иными возрастными категориями. Для каких возрастных категорий характерны эти слова, для каких другие слова. То есть мы найдём э-э семантические ядра и антиядра каждой семантической, каждой возрастной категории. Вот. Ну она тоже считает помедленнее, конечно, но всё равно видно, что не очень долго. То есть вот сейчас у нас, ну, где-то минуты три будет она считать. Вот здесь вот у нас 2 минуты осталось, а здесь прошло полторы минуты. Ну где-то минуты три, три с половиной, может быть. Ну то есть вполне приемлемые, ребята, сроки расчётов, даже на таких уже довольно большого, большой размерности моделях, довольно приличных. И не такой вот маленький учебный примерчик, который я там придумал, а уже взято с солидного сайта. Фирма Kaggle, она существовала сначала самостоятельно длительное время, потом стала лидером в области порталов с данными для машинного обучения, и потом её купила фирма Google. Сейчас это, можно сказать, так, гугловский, гугловский портал. Значит, э-э, как она этого добилась, фирма Kaggle? Как она добилась такой популярности? Если раньше на сайтах просто размещались материалы для машинного обучения, э-э, и, UCI, например, просто там данные размещены и всё, берёшь, а дальше делаешь с ними что хочешь, и, в общем, никакой обратной связи с порталом нет. То есть ты больше никак с ним не взаимодействуешь. То на Kaggle там есть конкурсы, есть конкурсы платные, в которых есть призовой фонд. Вот 100.000 долларов призовой фонд. Его распределяют между первыми э-э наиболее эффективными решениями, ну, авторами первых наиболее эффективных решений. Э-э, пять призовых мест рассматривается. Ну, в разных конкурсах по-разному. Вот 100.000, вот 91.000, вот 30.000 долларов. Вот. А есть э-э конкурсы уже закрытые, уже которые закончились. Есть также просто много наборов данных, которые э-э тоже ты посылаешь результаты решения задачи на Kaggle и видишь своё место в рейтинге э-э других игроков. Вот лидеры. Ну здесь видно, видите, пять призовых мест. Это вот линия, она обозначает уровень э-э этого игрока. Уровень этого товарища. Это великий мастер, ребята, гроссмейстер из Цюриха. Он несколько раз занимал первые места в мире. То есть он неоднократный чемпион мира. Двукратный чемпион мира. По одной из задач, которая была размещена на Кагле. Вот. Значит, это продолжалось у нас 4 минуты, что вполне приемлемо. Я принимал участие в двух конкурсах, но это надо времени для этого иметь много, а у меня его нету этого времени. И э-э получалось так, что я где-то вхожу, один раз вошёл в первую десятку по рейтингу, один раз в первую тройку по рейтингу. А в России я принимал участие в чемпионате, открытом чемпионате России в семнадцатом году, и из трёх номинаций занял два первых места по оценке экспертного сообщества. Вот. То есть и не всё время, был период, когда я там был на этой позиции.

A. Проверка достоверности модели

Вот, смотрим достоверность модели. Значит, насколько высока достоверность. Вот. Ну мы видим, что здесь, что при повышении уровня сходства, э-э, закономерно растёт доля истинных решений, и начиная с там где-то 35, там 30% уровня сходства, доля истинных решений преобладает. По другим критериям. Но этот уровень достоверности, он сравнительно невысокий, но вот по L2 критерию 956 - это довольно высокий уровень достоверности. Это высокий уровень достоверности уже.

B. Анализ результатов атрибуции и семантического ядра

Вот. То есть мы эту задачу решили, у нас получилось. Мы, они, собственно говоря, что хотели? Они хотели узнать, можно ли по э-э получить лингвистические эти э-э ядра, семантические ядра этих вот возрастных групп. Пожалуйста. Вот вам возрастная группа 0-3. Вот её семантическое ядро отсутствует, значит, здесь вообще. Так. Вот слова наиболее характерные и наиболее нехарактерные для этой возрастной группы. Вот, пожалуйста, получили. Можно это посмотреть не в такой ограниченной форме, а целиком. Есть таблица соответствующая, которая создана э-э и находится э-э там же, там, где база данных этого приложения. Значит, вот сейчас я вам покажу. Значит, мы заходим в базу данных приложения, сортируем по времени. Вот. И вот у нас таблица. Вот семантическое ядро, ребята, этой возрастной группы. Правила, которые просили выяснить на Кагле. Мы его выяснили. Итак, мы выяснили правила для каждой возрастной группы, и можем сказать, что если будет у нас какие-то слова в книжке, то мы по этим словам можем определить возрастную группу, которой эту книжку можно рекомендовать.

IX. Заключение и следующие шаги

Всё, ребята, у нас на этом конец занятия. Вопросов, я так понял, у вас нету, вы уже сказали про это. Поэтому, наверное, на этом до свидания.

До свидания.

До свидания.

До свидания.

До свидания.

До свидания.

До свидания.

А на следующем занятии мы рассмотрим реальную задачу разработки методики риэлторской оценки автомобилей с пробегом по методу аналогии или по методу сравнительных продаж, которая является просто задачей распознавания образов и элементарно решается в системе Eidos. То есть в системе Eidos разработка методики риэлторской оценки - это лабораторная работа. Ну это вполне реальная работа. В Плехановском университете такую работу провели, написали статью про это хорошую. Значит, ну, в общем, это вполне реальные интересные вещи. Всё, на этом, ребята, конец занятия. До свидания.