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

***И.Т. Трубилина»,***

***Российская Федерация***

***Колесников Роман Юрьевич, ПИ2102***

***roman563412@gmail.com***

**Автоматизированный системно-когнитивный анализ текстов (на примере сочинений о писателях)**

**Заголовок**

Обзор системы Эйдос для интеллектуального анализа текстов: возможности, интерфейсы, демонстрация

**Резюме**

Лекция посвящена применению интеллектуальной системы Эйдос для обработки и анализа текстов. Рассматриваются основные возможности системы, включая создание обобщенных лингвистических образов классов, семантических ядер, сравнение текстов с этими образами, кластеризацию, атрибуцию текстов (определение вероятных авторов, в том числе анонимных и псевдонимных), датировку, определение жанра и смысловой направленности.

Описывается процесс получения доступа к системе через сайт разработчика, включая скачивание и установку. Демонстрируется навигация по сайту для поиска документации и подборок публикаций по АСК-анализу текстов.

Объясняются принципы работы с текстовыми данными в Эйдос через различные интерфейсы:

Интерфейс 2.3.2.1: Предназначен для ввода больших текстовых файлов (книги, статьи). Описаны стандарты именования файлов для автоматического определения классов и номеров реализаций. Упоминается возможность настройки длины анализируемых слов и проведения лемматизации.

Интерфейс 2.3.2.2: Работает с текстами, представленными в таблицах Excel, где колонки могут содержать как классификационные признаки, так и описательные текстовые данные.

Подробно разбирается процесс лемматизации – приведения слов к их первообразной форме (лемме), основанный на базе данных профессора Зализняка. Отмечается, что система Эйдос способна автоматически дополнять эту базу новыми словами и словоформами, встречающимися в анализируемых текстах.

Проводится демонстрация работы с облачным приложением Эйдос на примере анализа школьных сочинений о писателях. Показывается процесс загрузки приложения, автоматической обработки данных (включая лемматизацию), создания моделей и просмотра результатов. Демонстрируются различные режимы анализа:

Просмотр классификационных и описательных шкал.

Оценка достоверности моделей (Inf1, Хи-квадрат).

Идентификация (определение класса объекта по его описанию) и обратная задача прогнозирования (определение характерных признаков класса).

Визуализация результатов с помощью когнитивных диаграмм и дендрограмм кластеризации (хотя отмечается, что для текстовых данных результаты кластеризации могут быть не очень наглядными из-за влияния метаданных вроде года и жанра).

Подчеркивается, что Эйдос позволяет оценивать степень истинности своих выводов через уровень сходства, что является важной особенностью системы, существующей с ранних версий.

Детальная расшифровка текста

Раздел 1: Введение и обзор возможностей системы Эйдос для анализа текстов

Подраздел 1.1: Приветствие и тема занятия

00:00:02 Ну что ж, уважаемые Фёдор Олегович, уважаемые слушатели.

00:00:06 Сегодняшнее наше занятие посвящено применению системы Эйдос для обработки текстов.

Подраздел 1.2: Проверка наличия установленной системы у слушателей

00:00:12 Ну с чего начнём? С того, что я покажу вам...

00:00:15 Да, давайте мне информацию такую: у кого установлена система Эйдос и кто будет сейчас пытаться повторять за мной?

00:00:23 Плюсики, пожалуйста, в чате укажите.

00:00:28 Вот. Ну...

00:00:30 Алексей, Екатерина... Ну, вы пришли первыми, да?

00:00:35 Ну и всё, и всё, да?

00:00:37 Ну хорошо, ладно, будем иметь в виду.

Подраздел 1.3: Формат взаимодействия и демонстрация экрана

00:00:40 Значит, э-э, если какие-то будут вопросы возникать, э-э, по применению уже на ваших компьютерах, вы можете мне голосом прямо сразу говорить,

00:00:48 потому что в чате я сейчас чат не буду, наверное, видеть полностью.

00:00:52 Значит, я сейчас сделаю демонстрацию экрана.

00:00:55 Вот.

00:00:59 И начнём.

01:02 Вы должны видеть экран мой.

Раздел 2: Доступ к системе Эйдос и сопутствующим материалам

Подраздел 2.1: Навигация по сайту разработчика

01:05 Вот. Значит, что мы делаем э-э первым делом? Значит,

01:10 э-э, я вам хочу показать, это вот мой сайт lc.kubagro.ru.

01:15 Я буду в чат отсылать э-э ссылочки, как обычно.

01:19 Вот. И на сайте

01:22 мы можем увидеть э-э подборки публикаций по различной тематике.

01:27 Это на второй кнопочке: теоретические основы, технологии, инструментарий системно-когнитивного анализа.

01:33 Вот по этой ссылочке мы выходим на вторую страницу, где всё у нас про систему Эйдос, все

01:40 свидетельства Роспатента, монографии, всё касается интеллектуальных технологий на этой страничке.

Подраздел 2.2: Скачивание системы Эйдос

01:46 И э-э выбираем, там где вот круговая диаграмма, в самом начале страницы, круговая диаграмма

01:51 о типах интеллектуальных систем. Рядышком кнопочка есть скачать и запустить систему Эйдос. Вот эту кнопочку кликаем.

01:58 И здесь у нас, видите, есть красным шрифтом э-э кнопочка для скачивания самой системы.

02:05 Вот.

02:06 Ну сначала вторую кнопочку, я вам сейчас

02:09 ссылочку пошлю по второй кнопке.

02:12 И здесь у нас есть

02:18 среди э-э, ну здесь вот скачивание системы.

02:22 Ещё раз, если кто не скачал, первую скачиваем, э-э первый вариант архива

02:27 и распаковываем в корневом каталоге какого-нибудь диска, лучше всего SSD диска.

Подраздел 2.3: Поиск публикаций по АСК-анализу текстов

02:33 И вот здесь у нас среди подборок публикаций есть подборка э-э работы,

02:38 тематическая подборка публикаций по АСК-анализу текстов.

02:43 Вам, наверное, видно, да? Я так думаю.

02:47 Вот. И мы по этой... это 6.3, пункт 6.3.

02:51 Кликаем по этому пункту. Здесь, правда, надо ещё один раз потом кликнуть.

02:55 Вот, тоже по пункту 6.3.

02:58 И у нас появляется

03:01 подборка публикаций по

03:03 анализу текстов с помощью системы Эйдос.

03:06 Я вам говорил, что она позволяет обрабатывать табличные данные, тексты, причём в таблицах могут быть числа и тексты.

03:13 Э-э, также позволяет обрабатывать э-э текстовые файлы и позволяет обрабатывать графические файлы. Ну, графическому мы посвятим следующее занятие

03:22 э-э обработки графической информации. А сейчас мы посмотрим, рассмотрим

03:27 вопрос по анализу текстов в системе Эйдос.

03:31 И вот здесь сразу же в этой подборке публикаций мы видим, что тут довольно-таки много публикаций, ну сравнительно, 16 публикаций

03:38 по анализу текстов.

03:41 И э-э я э-э начинаю с чего? С того, что рассказываю вам, какие возможности системно-когнитивного анализа по анализу текстов.

03:50 Какие возможности.

03:52 Вот. Ну сначала, я не знаю, стоит ли это посылать. Ссылочку я послал.

03:57 Ну пошлю.

03:59 Пошлю э-э ссылки на все публикации по анализу текста, которые сейчас есть.

04:05 Э-э, ну, возможно, у меня ещё какие-то есть, но я не помню, где-то там потерялись.

Раздел 3: Возможности системы Эйдос для анализа текстов

Подраздел 3.1: Формирование обобщенных лингвистических образов и семантических ядер

04:11 Вот. Значит, что мы видим?

04:14 Что системно-когнитивный анализ позволяет формировать обобщённые лингвистические образы классов,

04:20 которые являются семантическими ядрами.

04:23 Сейчас я объясню все эти понятия.

04:25 На основе фрагментов текстов или примеров, относящихся к ним текстов на любом языке.

04:31 Причём на языке естественном языке или искусственном.

04:35 Вот. То есть что это значит? Что мы можем взять какие-то тексты

04:40 э-э и отнести эти тексты к каким-то обобщённым категориям, классам.

04:46 И э-э создать обобщённые образы этих классов на основе ряда текстов, создать один образ класса, на основе ряда других текстов - другой образ класса.

04:55 Причём один и тот же текст или фрагмент текста может относиться, ну текст, будем считать, может относиться к различным классам одновременно.

05:04 Ну, например, может быть э-э автор какой-то, направление науки,

05:11 год публикации, жанр.

05:15 Вот. Ну, допустим, статья, монография. И мы можем один и тот же текст э-э использовать для формирования этих обобщённых образов,

05:23 а также других,

05:25 э-э, которые я не перечислил.

Подраздел 3.2: Сравнение текстов и объектов с обобщенными образами

05:28 Э-э, количественно можно сравнивать лингвистический образ конкретного человека или описание объекта,

05:35 процесса с обобщённым лингвистическим образом.

05:38 Ну, допустим, мы можем взять э-э фотографии университета или Перми и э-э описать словами эти фотографии.

05:46 Как вот сейчас делается в системах генерации изображений типа Midjourney.

05:51 Вот. Мы описываем словами некие изображения, и система их генерирует эти изображения.

05:56 Мы можем сделать наоборот. Да, и если обратная функция, насколько я знаю, то есть можно изображение взять и будет э-э система создаст его описание.

06:05 Так вот, мы можем

06:07 описать различные места словами.

06:11 Вот. И после этого, значит, причём это место может быть описано с разных ракурсов, с разных позиций.

06:18 Вот. И с разных точек зрения, что ли, и так далее.

06:22 Вот. И после этого можно э-э решить такую задачу:

06:27 дать это конкретное описание

06:30 какое-то или новое какое-то конкретное описание этого места.

06:33 И э-э система определит, что это за место по описанию. Причём описание не будет совпадать с теми описаниями, которые там были

06:41 в качестве примеров при формировании этого обобщённого образа. Сразу вспоминается эпизод из фильма э-э "Джентльмены удачи", где искали там такое дерево,

06:52 и потом там памятник. Помните? Кто ж его посадит там и так далее.

06:56 То есть какое-то э-э словесное, вербальное описание места, э-э довольно долго они искали, не могли найти это место.

07:04 Потом случайно там оказались.

Подраздел 3.3: Кластеризация и исследование предметной области

07:07 Также можно сравнивать обобщённые лингвистические образы классов друг с другом и создавать их кластеры и конструкты. То есть можно сравнивать,

07:15 ну, скажем, писателей друг с другом по их э-э текстам

07:19 или учёных. Можно

07:24 э-э сравнивать э-э тексты разных э-э лет написания

07:30 вот, или разных жанров.

07:32 Мы это сделаем.

07:34 И можно проводить, естественно, кластерный анализ и другие э-э формы анализа, проводить этих обобщённых образов классов.

07:43 Также можно исследовать моделируемую предметную область путём исследования её лингвистической системно-когнитивной модели.

07:50 То есть просто можно ещё ряд форм получать различных.

Подраздел 3.4: Атрибуция текстов (авторство, датировка, жанр)

07:54 Проводить интеллектуальную атрибуцию текстов.

07:57 Атрибуция текстов имеется в виду, вот этот термин атрибуция, он предполагает имеет такой смысл, что мы определяем вероятных авторов.

08:05 Причём эти вероятные авторы, они могут быть и анонимные, и псевдонимные.

08:10 Анонимные - это вообще не указан автор в тексте, а мы можем этот текст проанализировать и сказать: вот, судя по всему,

08:16 этот текст написан таким-то и таким-то автором, потому что очень э-э сходен этот текст с теми текстами, которые этот автор э-э создал.

08:26 А псевдонимные тексты - это тексты, которые подписаны

08:31 фамилией какого-то автора, но не того, который написал.

08:34 Псевдонимом подписаны.

08:36 Вот Владимир Ильич Ленин, например, написал статью

08:40 по поводу стачки и расстрела э-э расстрела демонстрации на реке Лена.

08:47 Вот. И стал известен как Ленин. Эта статья, она была очень популярна, её переписывали, читали все.

08:55 Вот. И э-э стало известно, что есть некий Ленин,

08:59 блестящий публицист,

09:02 э-э, который вот написал эту статью. И он стал известен под этим псевдонимом.

09:06 И потом, когда говорили, это вот Ульянов, он никто не знал, кто такой Ульянов. А вот такой Ленин все: "А, так это тот автор той статьи, что ли?" Ну да.

09:14 Вот. И он, видите, этот пиар получился такой, он вошёл в историю, потом стал этот псевдоним использовать.

09:21 Также можно определять вероятную датировку текстов, жанр,

09:25 смысловую направленность, э-э содержание текстов.

09:30 То есть, ну, например, можно взять сайты, проанализировать контент сайтов и э-э определить,

09:37 э-э, о чём эти сайты.

09:39 Причём э-э совершенно неважно, э-э на самом деле там написано про это или нет.

09:45 А важно, что сайты определённой смысловой направленности э-э имеют определённые

09:52 характерные для них слова, обороты слов.

09:56 И система именно эти слова и будет считать характерными. То есть там неважно, чтобы на... чтобы было написано прямо прямым текстом о чём-то и так далее.

10:05 Ну вот в этих статьях, э-э, видите, довольно-таки много разных примеров приведено. Все статьи в открытом доступе.

10:12 Можно их смотреть, читать.

Раздел 4: Демонстрация работы с системой Эйдос

Подраздел 4.1: Запуск системы и начальные настройки

10:15 Вот.

10:17 И дальше мы что сейчас сделаем? Сейчас мы запустим систему Эйдос. Да, если какие-то вопросы возникают по ходу изложения, пожалуйста, прямо сразу же и спрашивайте.

10:28 Желательно прямо голосом, чтобы я слышал.

10:31 Можно в чате, но в чате я могу не заметить, что вы спросили.

10:35 Вот. Ну, пока вот вроде как всё.

10:39 Запускаем систему Эйдос.

10:45 На моём компьютере подключён FTP, поэтому здесь она быстро запускается.

10:50 На компьютерах, на которых FTP заблокирован,

10:53 будет э-э пытаться достучаться и проверять его наличие, а потом будет писать, что его он отключён.

11:01 Вот.

11:02 Значит, в системе Эйдос есть несколько автоматизированных программных интерфейсов с внешними источниками данных.

11:08 Некоторые из которых позволяют вводить тексты.

11:11 Она при запуске сразу же запускает сайт системы.

Подраздел 4.2: Обзор интерфейсов для ввода текстовых данных

11:16 И вот мы посмотрим, где эти интерфейсы. 2 3 2

11:19 В этом режиме мы видим все эти интерфейсы: 1... 2 3 2 1, 2 3 2 2 и так далее.

11:27 Вот интерфейс 2 3 2 1 -

11:29 это интерфейс, который прямо предназначен для ввода текстовых файлов.

11:34 Вот. То есть здесь могут быть большие файлы, ну книги, книги могут вводиться

11:40 и сравниваться по своим, значит, э-э, ну, словам и сочетаниям слов.

11:47 В этом режиме сейчас пока что э-э реализовано только TXT файлы, доковские нет.

11:54 И здесь есть э-э встроенный перекодировщик,

11:58 который позволяет нужную кодировку файлов получить, найти их и получить.

12:04 Вот. Помощь есть, где описывается, вот примерно то, что я рассказываю.

12:09 Есть четыре стандарта

12:11 ввода текстов

12:13 в системе.

12:15 Э-э, где имена файлов содержат информацию о том, э-э, какой класс или классы создавать э-э на основе этого текста.

12:25 Значит, первый стандарт, мы используем, видите, ID1, ID2, IDN, потом тире, решётка, TXT.

12:38 Значит, э-э, система будет рассматривать каждый ID, отделённый запятой, как э-э

12:45 наименование класса.

12:47 То есть можно взять, допустим, монографию, у которой несколько авторов, коллективную монографию,

12:51 и э-э перечислить их через запятую,

12:54 и, значит, э-э этот текст будет использоваться для формирования обобщённых лингвистических образов всех авторов.

13:02 Вот эта вот э-э решёточка, она может быть любого размера и предназначена для номера.

13:08 Там может быть 1 2 3 написано, может быть написано 01 02 03, там, как хотите.

13:13 Главное, что система Эйдос доходит до тире в наименовании файла и считает, что дальше это уже

13:19 не информация о классах, а это информация о реализации.

13:23 То есть у этого автора или авторов может быть несколько файлов,

13:27 которые она все использует для формирования этих вот образов ID1, ID2.

13:32 А там, где решётка, там будет просто номер файла,

13:36 соответствующий этим классам.

13:38 Второй стандарт системы Эйдос для скачивания файлов, где текстовых,

13:43 проще.

13:44 Там просто имя класса, потом тире и номер реализации. Ну, например, имя класса - это может быть или автор, или какое-то направление.

13:54 Вот. Также точно и здесь вот ID - это не только авторы могут быть, это может быть автор, допустим, э-э специальность, например, научная.

14:02 И написано будет монография или статья, например,

14:05 этого автора, то есть жанр, например, может быть написан, год там же может быть.

14:11 То есть там можно указать все эти вот параметры для формирования образов которых, э-э соответствующих, используется этот файл.

14:19 Ну, второй стандарт, он значительно понятнее. Вот, можно просто написать имя автора и номер э-э текста об этом авторе.

14:28 Э-э, третий стандарт объединяет первый и второй стандарты.

14:35 А четвёртый стандарт - это стандарт для Kaggle, для для текстов, которые на Kaggle

14:41 э-э можно считывать и создавать э-э интеллектуальные приложения.

14:46 Кроме того, мы можем приложение либо э-э

14:50 создать, либо сформировать распознаваемую выборку с имеющимися шкалами уже,

14:55 если уже ранее было создано приложение.

14:58 И э-э для эту выборку создать в стандартах интерфейса 2 3 2 1, 2 3 2 2.

15:05 Вот.

15:08 Мы также можем формировать файл исходных данных

15:12 для создания модели продолжения фраз.

15:15 Это что касается интерфейса 2 3 2 1. Это интерфейс специально создан для

15:20 обработки текстовых файлов, и он работает. Есть статьи, где описан этот интерфейс.

15:25 Мы можем использовать в качестве признаков э-э классов

15:29 э-э

15:33 то есть, когда мы будем эти тексты анализировать в системе, то мы можем

15:37 дать ограничения на длину слов, которые мы будем использовать.

15:42 Ну, допустим, здесь вот у нас такое условие: больше четырёх символов, чтобы было или равно, и меньше 35 или равно.

15:50 Больше или равно, меньше или равно.

15:52 Для чего? Для того, чтобы очень короткие слова не использовать и очень длинные.

15:56 Также можно проводить лемматизацию.

15:59 Можно использовать слова в качестве признаков текста, а можно использовать мемы,

16:04 то есть сочетания слов.

16:07 Вот.

16:10 Ну сейчас мы, я вижу здесь накладку при разработке.

16:14 Не задавал эту опцию при разработке и увидел при разработке нового интерфейса,

16:19 новой версии интерфейса. Буду знать. Вот так вот, когда вот рассказываешь, то и находишь там что-то.

16:25 Вот. И есть у нас э-э учебные приложения для разработки

16:30 моделей, связанных с обработкой текстов.

16:33 И, значит, есть э-э лабораторная работа третьего типа 3.02.

16:40 Э-э, очень простая лабораторная работа,

16:43 такая предельно примитивная, я бы сказал.

16:48 Э-э, и есть э-э облачные приложения,

16:51 тоже предназначенные для

16:54 демонстрации возможностей обработки текстов, которые представлены в таблицах.

17:00 Вот это облачное приложение я его сделал на основе работы 3.02.

Подраздел 4.3: Лемматизация и семантические ядра

17:08 Значит, вот когда мы запускаем Help этой лабораторной работы, то нам как раз и выводятся вот эта информация о том, какие есть

17:16 публикации по АСК-анализу текстов.

17:20 И э-э что ещё я хотел бы сказать? Что такое семантические ядра?

17:24 Это те слова, которые отличают какой-то лингвистический образ класса

17:31 от других классов, то есть те, которые наиболее характерны для этого класса

17:35 и э-э отличают его э-э от других классов.

17:39 Это и есть семантическое ядро.

17:42 Ну, что-то вроде, можно так сказать, ключевых слов, если взять научные тексты, то это напоминает то, что мы называем ключевые слова обычно.

17:50 Какие-то слова,

17:52 извините, наиболее характерные для этого текста.

17:57 И мы будем использовать

17:59 приложение облачное.

18:02 Скачать приложение из облака.

18:04 Э-э, значит, я недавно разместил

18:07 специальное приложение для сегодняшнего занятия.

18:11 Восьмого числа,

18:13 то есть 3 дня назад,

18:16 разместил такое приложение: АСК-анализ текстов на примере сочинений о писателях.

18:23 И вот это приложение,

18:25 приложение вот это, которое я упоминал,

18:29 работа 3.02, тоже здесь есть

18:32 в самом начале.

18:33 Работа третьего типа, они могут э-э загружаться либо локально, либо

18:39 могут загружаться э-э онлайн с хостинга, с того, на котором мой сайт находится.

18:48 Вот. Значит, сейчас мы э-э переходим э-э

18:53 вниз и загружаем вот это приложение, и на нём, на его примере будем рассматривать этот вопрос сегодняшний.

19:01 Для того, чтобы его загрузить, достаточно нажать на кнопочку Установка Эйдос приложения из облака.

19:08 Сразу появляется окошко. Э-э, язык программирования, который использовался SBS++, он позволяет

19:15 получить атрибуты файлов на хостинге без скачивания этих файлов. То есть он работает с файлами на хостинге примерно так, как будто они находятся локально на компьютере.

19:25 Вот. То есть интересные такие э-э возможности. То есть независимо от того, какого размера там файлы и сколько их, обращение происходит мгновенно,

19:34 получение атрибутов.

19:36 Ну примерно, как по локальной сети, наверное, можно сравнить.

19:40 Здесь мы видим перечень файлов, стандартные их э-э характеристики: имена,

19:46 размер э-э файлов в байтах,

19:48 дата создания и время создания.

19:51 Этот и суммарный объём ещё я написал, просуммировал.

19:55 Когда файлы созданы на хостинге. Там дата и время создания на хостинге именно.

20:02 Это окошко просто информационное, закрываем его.

20:06 Начинается процесс скачивания

20:08 этих файлов на э-э компьютер локальный.

20:13 Они все скачались.

20:15 Появляется...

20:17 потом была небольшая пауза, бывает иногда некоторая пауза

20:21 из-за того, что там эти файлы начинают обрабатываться, анализироваться. И потом вот вывелась информация нам, информационное окошко

20:28 о том, что в папку, прямо конкретный путь на реальный, на на тот, который на данном компьютере, в том месте, где запущена система.

20:36 Значит, э-э информация, что в эту папку скачан файл

20:41 API Type,

20:43 э-э, который э-э, в котором содержится информация о том, какой автоматизированный программный интерфейс использовался

20:49 при записи этого приложения в облако.

20:53 То есть, когда система записывает приложение в облако, то она записывает ещё и информацию о том, какой автоматизированный программный интерфейс использовался при создании этого

21:02 приложения. Там один из шести указывается.

21:06 Вот.

21:07 Второе.

21:09 Ну здесь об этом как раз и написано, что это делается.

21:11 Среди этих файлов есть доковские файлы, TXT файлы, PDF файлы, которые, вероятно, содержат описание

21:18 данного приложения, что так и есть.

21:21 Дальше нажимаем О'кей.

21:24 Появляется э-э окно уже того интерфейса, который мы с вами видели на первом занятии,

21:29 содержательном, где приложение мы рассматривали.

21:33 Значит, этот интерфейс имеет э-э... это интерфейс 2 3 2 2. То есть здесь у нас тексты находятся в экселевской таблице.

21:40 Сейчас мы на неё посмотрим. Значит, какие параметры мы здесь видим, указанные? Значит, э-э файл тип данных новый Excel.

21:49 Э-э, колоночки классификационные со второй по четвёртую, описательная пятая.

21:54 Нули и пробелы считать отсутствием данных.

21:57 Создание модели,

21:59 э-э равные интервалы в числовых шкалах.

22:02 Не применять э-э сценарный метод АСК-анализа.

22:05 Не применять специальную интерпретацию текстовых полей классов.

22:10 То есть здесь совершенно стандартные параметры, которые по умолчанию.

22:14 Да, кстати, все параметры этого вот интерфейса тоже считанны сейчас с хостинга, и они поэтому здесь вот мы видим их.

22:20 А вот э-э

22:22 текстовые поля

22:24 описательных шкал, э-э используется специальная э-э интерпретация текстовых полей признаков описательных шкал. Видите, птичка стоит.

22:33 Э-э, ребят, пожалуйста, дайте мне подтверждение, что вы всё это видите и слушаете

22:39 в чате.

22:43 Угу.

22:44 Вот. Хорошо.

22:47 Вот. Значит, здесь мы э-э видим такие параметры, что мы можем э-э значение поля целиком использовать как значение поля,

22:54 а можем использовать слова.

22:57 И то есть э-э какие-то элементы текстов, которые разделены разделителем такими, как пробел, запятая и знаки, другие знаки препинания.

23:07 И можем сделать ограничение вот здесь на минимальную длину символов.

23:12 Почему? Потому что опыт показывает, что при обработке

23:17 текстов, э-э чаще всего очень короткие э-э слова из одного, двух, трёх символов, они обычно не играют существенной роли,

23:27 а сильно э-э увеличивают количество э-э слов в справочниках, в справочниках.

23:33 Поэтому можно их не учитывать.

23:35 И мы видим, вот сейчас у нас э-э задан вариант использования в качестве признаков этих файлов самих слов.

23:43 Вот. А также в системе, видите, поддерживается вариант, когда в качестве признаков рассматриваются символы.

23:50 И здесь вот есть ниже, мы видим опцию такую:

23:54 проводить лемматизацию или не проводить лемматизацию.

23:58 Значит, лемматизация - это операция преобразования слова

24:02 к первообразному, из которого оно создано, это слово, путём словообразования.

24:10 Сейчас мы конкретно на это посмотрим. Нажимаем О'кей.

24:15 Сейчас э-э выдаются сообщения о том, что там конвертация файла, анализ файла.

24:21 И сейчас мы видим, ребята, что у нас просит система немножко подождать,

24:26 что она работает и проводит вот ту самую лемматизацию. То есть она берёт э-э слова, которые там есть в тексте, и находит для них первообразные слова.

24:36 Примерно так же э-э поступают разработчики поисковых систем в интернете.

24:43 То есть, когда мы делаем запрос в интернете, то там склонения слов, э-э количество, значит, множественное там и единственное числа, они роли не играют.

24:53 То есть тоже слова в запросе сначала слова лемматизируются, а потом уже делается запрос э-э поиск по индексным массивам.

25:01 Значит, сейчас у нас э-э, видите, потрачено 32-33 секунды на лемматизацию,

25:08 2 сотых секунды на одно слово.

25:11 И, значит, 1424 слов, значит, было проанализировано.

25:17 1424 слова. То есть это не очень много. То есть там маленькая э-э маленькие тексты для того, чтобы быстро можно было всё это обрабатывать и смотреть.

25:26 Но в результате лемматизации число их сократится.

25:29 Приведу вам пример лемматизации, которую я всегда студентам привожу.

25:33 Э-э, ну, например, слово э-э стол.

25:36 Какие слова можно путём э-э словообразования, методов различных словообразования,

25:42 э-э создать из основываясь на этом слове стол как базовом?

25:47 То есть слово стол - это лемма,

25:49 а слова, сформированные на его основе: столица,

25:53 столешница, столяр, столовая

25:57 и другие слова.

25:59 Кто из вас ещё придумает, тот выиграл.

26:02 Вот.

26:06 И вот все слова подобные, они будут преобразованы к слову стол, к единственному числу, именительного падежа.

26:13 Вот. Лемматизация есть двух уровней: когда преобразуется к единственному числу именительного падежа,

26:19 а есть ещё преобразование второго уровня лемматизации к существительному.

26:24 То есть есть некоторые э-э другие э-э формы речи,

26:28 которые тоже можно преобразовать к существительным.

26:32 И тогда вот эта лемматизация второго уровня.

26:35 Если вы возьмём, найдём сервис онлайн, э-э обеспечивающий лемматизацию, и посмотрим, что получается, когда мы возьмём некий текст, поместим туда в окошко или лемматизируем его,

26:47 то получается так, как будто немцы только-только начинают осваивать русский язык.

26:51 Вот примерно так пишет она. Попробуйте для интереса э-э онлайн-сервис лемматизации какой-нибудь и туда какую-нибудь фразу там закиньте Пушкина, Толстого, посмотрите, что получится.

27:04 Нажимаем О'кей.

27:05 Здесь мы видим э-э внутренний калькулятор, э-э конвертера,

27:09 э-э и интерфейса, который преобразует текстовые файлы в

27:14 внутренние стандарты системы Эйдос, новую базу данных.

27:18 И мы видим, что у нас три текстовых классификационных шкалы, одна описательная.

27:24 И выходим на создание... Здесь нет числовых шкал, поэтому нет вопроса о том,

27:28 э-э насколько числовых диапазонов разделять эти шкалы.

27:32 Потому что этот вопрос касается только э-э числовых шкал.

27:36 Текстовые шкалы, градации формируются на основе

27:39 такого подхода, что просто сортируются по алфавиту градации шкалы

27:44 э-э с признаком unique,

27:46 и выбираются уникальные значения и формируются соответствующие справочники.

27:51 Выйти на создание моделей.

27:53 Сейчас созданы справочники э-э описательной классификационной шкалы и градации.

27:59 И потом с их помощью закодированы исходные данные, в результате чего получается обучающая выборка.

28:05 То есть мы выполняем этапы преобразования данных в информацию, а её в знания.

28:10 Первые этапы у нас уже выполнены.

28:13 То есть мы взяли исходные данные.

28:15 Э-э, здесь вот у нас рекомендации, что делать дальше.

28:20 Значит, давайте посмотрим на сам файл исходных данных,

28:23 а потом выполним режимы соответствующие, которые здесь перечислены.

28:27 Но и другие. Значит, сейчас скажу, что мы преобразуем данные в информацию, её в знания, и потом применяем эти знания для решения задач, как обычно.

28:36 Только у нас данные теперь текстового типа.

28:39 И задачи решаются связанные с обработкой, интеллектуальной обработкой текстов.

28:44 Э-э, чтобы преобразовать данные в информацию, мы что делаем?

28:49 Информация - это данные, в которых выявлены причинно-следственные связи. Причинно-следственные связи э-э выявляются только, то есть существуют реально только в реальной области, то есть между событиями. Между данными причинно-следственных связей нет.

29:05 Поэтому,

29:08 я последовал рекомендациям, закрыл окно 1.3, и таким образом приложение у нас установлено, о чём мы видим информацию.

29:15 АСК-анализ текстов на примере сочинений

29:18 о писателях. То есть это не сами тексты этих писателей, а это сочинения об этих писателях, школьные.

29:25 И выходим в главное окно.

29:27 Так вот, э-э,

29:30 информация - это данные, в которых выявлены причинно-следственные зависимости. Причинно-следственные зависимости существуют только между э-э событиями реальными, в реальной предметной области.

29:42 Это режим 6.4 я показываю.

29:46 Вот. Поэтому, чтобы эти события выявить в данных, в данных, мы сначала и выявить потом между ними зависимости, мы, значит, что делаем? Мы сначала должны разработать справочники событий прошлых и будущих

29:56 и потом с их помощью закодировать исходные данные, а потом уже найти зависимости причинно-следственные между прошлыми событиями и будущими.

30:05 Вот. И тогда мы преобразуем данные в информацию.

30:08 А потом уже сможем эту информацию использовать для достижения целей, решения различных задач. И тем самым преобразуем её в знания. То есть стандартно совершенно

30:18 процесс для системы Эйдос. Даже вот эти же самые интерфейсы используются.

30:22 Значит, теперь мы смотрим, посмотрим на то, какие модели созданы. А нет, не модели, а какие созданы справочники

30:28 моделей

30:30 и обучающая выборка. А потом создадим модели и исследуем их на достоверность,

30:35 как и описано в режиме э-э 6.4.

30:39 Значит, классификационные шкалы, режим 2.1.

30:43 Это у нас, видите, три шкалы: автор, жанр и год.

30:48 Посмотрим на файл исходных данных. Какая была у нас рекомендация дана?

30:52 Посмотреть на файл исходных данных. Посмотрим.

30:56 Значит, приложение у нас находится в папочке EidosX,

31:04 Aidata, Indata.

31:07 Кстати, мы видим здесь А001 - это уже папочка созданного приложения.

31:11 Indata. И здесь мы видим экселевский файл.

31:15 Вот. Indata XLSX.

31:19 Здесь мы что видим в этом файле? Что у нас есть информация об источнике данных.

31:25 Это первая колонка, она не является шкалой.

31:29 И здесь, видите, интересно написано: Достоевский 1, Достоевский 2, Достоевский 3. Что это такое?

31:35 Это взято сочинение школьное, причём не полностью оно взято,

31:39 э-э о Достоевском,

31:41 э-э в вордовский файл, э-э вставлен в Excel.

31:45 Э-э, это не все знают, я вам скажу, что э-э когда мы вставляем вордовский файл в Excel,

31:51 то каждый абзац

31:53 вставляется в отдельную строку.

31:56 То есть очень удобно э-э фрагменты текста получить.

32:00 Каждый фрагмент текста - это абзац исходного текста.

32:04 И, значит, если мы здесь напишем Достоевский, протянем вверх, вниз,

32:08 то Excel сам пронумерует Достоевский-1, напишем, и протянем вниз, то он напишет нам 1 2 3 4 5 и так далее.

32:16 Вот. Если мы напишем два раза Достоевский вот здесь в двух клеточках, возьмём две этих клеточки, протянем вниз, то не будет меняться э-э название клеточки.

32:26 То есть то есть содержимое этой клеточки не будет меняться.

32:29 И здесь мы видим, видите, что у нас есть э-э классификационные шкалы, выделенные жёлтым фоном.

32:35 Э-э, автор, жанр и год

32:38 публикации.

32:39 И вот мы, видите, э-э создали модель сейчас, где все эти три шкалы используются со второй по четвёртую.

32:48 А вот в этой шкале, где у нас сам текст или фрагмент текста, э-э здесь, значит, абзацы мы видим этих сочинений.

32:56 Значит, в одной клеточке Экселя можно помещать до

32:59 32К символов, то есть довольно много.

33:03 И в этом режиме, где мы вводим в систему Эйдос, используется такой метод,

33:08 мемополя используются, где мы можем до 64К использовать эти тексты. То есть вполне система Эйдос сможет обрабатывать такие тексты, ну примерно шесть страниц текста.

33:19 Это когда они в таком вот э-э интерфейсе экселевском.

33:23 Но если честно, я не... не проверял сам в последнее время, делает она это или нет. Я когда разрабатывал, она это делала.

33:30 Конечно, я вёл отладку. А потом я прошло много времени, я сейчас, честно сказать, не уверен, что она 32К поддерживает.

33:37 Но 255 символов точно поддерживает, это уж 100%.

33:41 И вот здесь мы видим э-э известных писателей и поэтов: Достоевский, Гоголь,

33:47 Горький, Грибоедов,

33:49 Лермонтов, Пушкин,

33:52 э-э Тургенев и Толстой.

33:55 Ну вот я взял этих писателей.

33:58 И у нас сейчас задача - сформировать вот эти обобщённые лингвистические образы на основе ряда фрагментов текстов об этих писателях.

34:06 Мы сформируем образ Гоголя, лингвистический Достоевского.

34:11 И также мы сформируем обобщённые образы классов проза, стихи

34:16 и обобщённый образ э-э текстов, написанных в определённом году.

34:21 Чем это интересно, ребят? Значит, я вам скажу, что, допустим, какой-то текст может у нас быть э-э исторический, например.

34:30 Или, ну тот, который историки, так сказать, спорят о его датировке.

34:34 А у нас в системе уже много текстов, это э-э, о которых есть датировка.

34:40 И мы знаем, какие э-э слова и словосочетания были в ходу в этот период.

34:46 Вот. И каких там не было. Ну, например, если мы анализируем тексты XVI века,

34:51 то там ничего не будет про э-э плазму, про лазеры, про космические корабли. Хотя про космические корабли может и быть. Вот в Махабхарате, например, там есть про космические корабли.

35:03 Вот. И там даже есть и про ядерное оружие.

35:07 То есть такое впечатление, что Махабхарату написали в будущем, а не в прошлом.

35:11 Ну это отдельные такие бывают тексты удивительные, странные. А вообще, то есть если серьёзно говорить, то мы можем, значит, по таким словам, которые характерны для современной науки, например, понять, что это написано не в XVIII веке, а в XX, XXI.

35:27 Вот. Если мы возьмём тексты, которые написаны на территории э-э России,

35:32 Советского Союза,

35:34 и там будет всё время про КПСС написано, ну, наверное, это Советского Союза тексты будут.

35:39 Вот. Про э-э, кто, я не знаю, знаете уж, что такое КПСС или нет?

35:44 Э-э, Коммунистическая партия Советского Союза. То есть очень многие тексты содержали это сочетание, эту аббревиатуру.

35:51 То есть можно по частоте встречи этой аббревиатуры понять, какая датировка текстов.

35:58 Ну примерно так вот.

36:00 Вот суть такая.

36:01 Значит, э-э на файл посмотрели, посмотрели теперь на э-э градации классификационной шкалы автор.

36:08 Здесь уникальные, видите, уже э-э

36:11 имена авторов, фамилии, имена авторов.

36:14 Значит, те, которые были во второй колоночке.

36:19 Жанры, те, которые в третьей колоночке, и года э-э в четвёртой колоночке. Значит, теперь посмотрите на года, как я их написал.

36:27 Я написал, видите, год, а потом пробел и буковка Г с точечкой. Для чего?

36:33 Для того, что система Эйдос работает таким образом, что если у нас

36:39 в колонке какой-то есть э-э текстовое значение, а потом дальше все числа выше и ниже числа,

36:47 то всё, что там в этой колонке есть, рассматривается как текст, если там хотя бы одно текстовое значение есть.

36:54 Ну и я единообразно набрал года с буковкой год.

36:59 Вот. Э-э для того, чтобы они как тексты рассматривались эти года, а не как числа.

37:04 Если они были бы как числа без буковки год в экселевском файле, то система спросила бы, на сколько числовых диапазонов эту шкалу делить.

37:12 И я бы мог указать три, например, или 10, там, сколько хотите.

37:16 Вот. И э-э у нас было бы тогда определён не сам год э-э при датировке текста,

37:23 а диапазон лет, к которому относится этот текст.

37:27 Ну сейчас мы проведём потом несколько экспериментов. Сейчас я э-э начальный вариант я показываю.

37:32 Значит, можно вот таким образом э-э вводить э-э года,

37:37 то есть вводить данные, которые мы хотели бы, как чтобы они анализировались как текст, а не как числа.

37:42 Теперь смотрите, вот у нас текст, э-э это уже описательная шкала,

37:47 режим 2.2.

37:49 Значит, здесь мы видим э-э слова.

37:52 И слова после лемматизации.

37:55 То есть здесь, если слово э-э есть, которое может быть преобразовано к первообразному, то здесь в основном-то первообразные слова и есть.

38:04 Обратите внимание, 730 слов, 731 слово, видите, да?

38:09 А у нас, когда мы проводили лемматизацию, было 1454 слова, если вы запомнили. Я вам несколько раз вам обратил ваше внимание,

38:17 что при лемматизации 1450 слов анализировалось.

38:21 Вот таким образом мы видим, что в результате лемматизации число слов сократилось, ну, практически в два раза.

38:28 Из-за чего? Из-за того, что там многие слова были образованы из одного слова, которое здесь вот у нас

38:34 э-э и есть как градация.

38:36 Значит, проводить лемматизацию или нет - это вопрос такой спорный.

38:40 В принципе, в первообразных словах тоже информация содержится об авторах,

38:45 об жанрах, периодах.

Подраздел 4.4: Просмотр обучающей выборки и создание моделей

38:49 Значит, смотрим теперь на обучающую выборку.

38:52 Значит, здесь мы видим э-э фрагменты текстов вверху. Это первая колоночка, которая не является шкалой.

39:01 43 примера.

39:03 А внизу, слева, мы видим коды классов,

39:05 которым относятся. У нас было три классификационных шкалы, поэтому здесь три кода класса.

39:11 А здесь коды слов

39:14 из справочника слов.

39:18 И в системе Эйдос э-э вот эта обучающая выборка, она представлена в виде трёх таблиц.

39:23 Каждая из которых э-э не больше 2 ГБ должна быть.

39:27 То есть суммарно 6 ГБ.

39:29 Но самая большая таблица, обычно третья вот эта вот, где признаки перечисляются.

39:34 Ну я вам скажу так, что 2 ГБ - это так, в общем-то, немало, но это не Big Data, конечно, но это немало, и можно вполне обрабатывать книги

39:43 и сравнивать книги между собой и статьи. То есть объёмы текстов достаточно большие можно обрабатывать. То есть здесь вот может быть прокрутка,

39:52 и там будет э-э можно много прокручивать, будет много результатов этих кодов слов.

39:59 И дальше у нас по нашей схеме обработки данных, информации и знаний,

40:04 дальше у нас создание моделей и проверка их на достоверность. Это следующий этап системно-когнитивного анализа. Первый - это

40:12 постановка задачи, выбор объекта моделирования,

40:15 факторов на него влияющих и результатов их влияния. Второе - это

40:19 получение степени формализации исходных данных

40:22 путём э-э разработки справочников прошлых и будущих событий

40:26 и кодирования исходных данных с помощью этих справочников и получения базы событий, которая является обучающей выборкой.

40:33 Здесь указано, какие таблицы в этих базах данных

40:36 создаются.

40:38 То есть э-э по две, по три таблицы создаётся на базу данных

40:43 с отношениями один ко многим. Теперь следующий этап системно-когнитивного анализа -

40:47 это создание моделей, отражающих взаимосвязи причинно-следственные между событиями. И тем самым мы

40:53 э-э завершаем преобразование исходных данных в информацию.

40:58 У нас в результате появляются, кроме данных, кроме базы исходных данных, у нас появляются базы справочников прошлых и будущих событий,

41:06 э-э сами исходные данные, закодированные с помощью этих справочников, то есть э-э база событий или

41:13 обучающая выборка.

41:15 И базы, отражающие причинно-следственные связи между этими событиями. Вот все вместе эти базы данных представляют собой информационные базы.

41:24 Потом посмотрим на сами модели созданные и будем с помощью них решать различные задачи.

41:29 Создаём эти модели в режиме 3.5.

41:33 Э-э с параметрами по умолчанию нажимаем О'кей.

41:38 Создаются модели. Ну, насколько это быстро или медленно? Значит, ну я вам скажу так, что здесь э-э модель получается э-э 23 класса, по-моему, на 700 признаков.

41:49 То есть достаточно небольшая размерности.

41:52 Если мы будем брать книги, анализировать, то там размерности будут значительно больше размерности, то есть по числу признаков, по числу слов.

41:59 Ну могут быть десятки, сотни тысяч слов.

42:03 Ну сотни тысяч едва ли.

42:05 Вот. Ну хотя сейчас э-э в современном словаре

42:08 английского языка больше миллиарда слов уже, мне кажется, насколько я помню.

42:14 Да, очень важный момент, ребята. Значит, для лемматизации использовалась база лемматизации профессора Залезняка,

42:22 разработанная профессором Залезняком, академиком Академии наук России.

42:28 Значит, я про него хочу рассказать в двух словах. Это замечательный учёный, он, к сожалению, умер

42:34 уже. Значит, замечательный лингвист, э-э уникальный совершенно человек, энтузиаст.

42:42 Значит, он очень многое сделал э-э в своей области,

42:47 значит, э-э лингвистики.

42:50 Заслуженный учёный такой, который всеобщим уважением пользуется.

42:54 Значит, э-э я... У него много достижений, но одно из его достижений - это э-э перевод, обнаружение и переводы новгородских рукописей

43:05 на современный русский язык.

43:07 И там очень интересные получены сведения. Ну, например, о том, что была поголовная грамотность,

43:12 что дети писали друг другу письма, дети, не только взрослые. Там есть тексты, которые связаны, ну, бухгалтерского характера, там учёт там и э-э контроль. Есть тексты такие, что там кто у кого брал взаймы.

43:26 Есть там э-э любовные письма, есть э-э дневники школьника. Это такого юноши, который учился в школе, именно в школе он учился.

43:36 И у него была коробочка, в которой он держал рукописи.

43:40 И там были э-э рукописи, переписка из друзьями,

43:44 а также его дневник,

43:46 где он просто описал о своей жизни, о событиях,

43:49 о том, кем бы он хотел стать, когда вырастет. Я не знаю, стал он им или нет.

43:54 Вот. Это остаётся, так сказать, загадкой.

43:58 Но это удивительно, что вот такие тексты э-э были найдены, и, значит, настолько высокий был уровень культуры, образования в Древнем Новгороде.

44:09 Вот. И это его достижение профессора Залезняка, что он эти рукописи обнаружил, перевёл.

44:16 И вторая его работа, имеющая огромное для нас значение,

44:20 что он разработал базу данных лемматизации русского языка,

44:24 которая, слушайте, ребят, дальше внимательно,

44:27 содержит около 2 млн слов.

44:29 Вот здесь я даю ссылочки на эти базы,

44:32 где я взял источник информации

44:35 на Хабре была статья, я нашёл эту базу на Хабре.

44:39 Она была в экселевском виде. Ну, в Экселе мы не можем больше миллиона строк в таблице иметь. Поэтому там было э-э в одном листе э-э

44:47 в одном листе Экселя было несколько колонок, и там были вот эти м-м базы

44:54 по две колонки. Слева база э-э слева колоночка лемма, а справа словоформа соответствующая. И так вот несколько было колоночек

45:03 таких штук шесть, наверное, насколько я помню. Я сделал небольшую программку, которая всё это конвертировала в нормальную базу данных.

45:11 И эта база данных используется в системе Эйдос. Она у нас... эта база данных, она размер имеет 270 МБ.

45:18 И э-э исходный свой размер, но она хорошо сжимается до 10 МБ.

45:24 Сейчас я попробую вам показать эту базу.

45:28 Вот эта база лемма.

45:30 Вот.

45:37 Значит, я открываю э-э просмотрщик э-э DBS файлов.

45:41 Указываю кодировку,

45:43 которая используется.

45:46 Э-э 866 кодировка

45:49 OEM или или CP 866.

45:53 И, ребят, скажите, пожалуйста, вам видно, что тут написано, нет?

45:58 Читабельно?

46:05 Вот. Вот так идут слова

46:07 и словоформы. Слева идёт словоформа,

46:10 а справа лемма.

46:14 Как вы думали, сколько слов в русском языке?

46:17 Ну, обычно в разговорном языке около 2.000 слов используется.

46:21 Значит, э-э писатели в своих произведениях использовали э-э

46:26 порядка там 10-20-30.000 слов.

46:31 Больше всех разных слов использовал Пушкин, Александр Сергеевич, около 25.000 слов.

46:37 А профессор Залезняк разработал базу лемматизации, в которой 2 млн слов, ребята.

46:43 Просто-напросто.

46:45 То есть в русском языке гораздо больше, чем 2 млн слов. Почему я говорю гораздо больше? Потому что он э-э разрабатывал эту базу традиционным путём.

46:54 Ну то есть без использования компьютерных технологий, насколько я понимаю.

46:58 То есть он просто брал э-э сами тексты и как вот Даль написал свой словарь, также вот и он.

47:05 И э-э там э-э огромное количество э-э слов, взятых из различных произведений.

47:12 Система Эйдос, она сделана таким образом. Вот слушайте дальше внимательно.

47:16 Если она не обнаруживает в базе лемматизации какого-то э-э какой-то леммы,

47:23 и, соответственно, то есть не леммы, а словоформы не обнаруживает, и, соответственно, и леммы не обнаруживает,

47:29 то она, знаете, что делает? Она добавляет в базу лемматизации эту словоформу.

47:34 И у неё есть интерфейс, она потом пишет, что были добавлены слова в базу, и указывает э-э в сообщении информационном

47:41 номер режима, в котором можно скорректировать базу лемматизации удобно.

47:46 Что значит удобно? Значит, представляем, мы видим эту базу в отсортированном виде.

47:51 Вот. И отмечаем э-э блоком, э-э то есть ставим лему на одной позиции, одной словоформы,

47:58 а потом протягиваем её, как в Экселе, вниз на все словоформы, которые этой леме соответствуют.

48:04 То есть очень удобный способ дополнения базы Залезняка.

48:08 Опыт показывает, что когда мы вводим э-э любую информацию, научную, там статьи, монографии, то там сотни, тысячи слов добавляются, даже когда мы отдельное приложение делаем. То есть мы делаем сейчас приложение, у нас там 1.500 слов добавилось в базу лемматизации.

48:24 Потому что в базе лемматизации отсутствуют научные термины

48:28 и современные слова практически полностью отсутствуют,

48:32 которые имеют часто иноязычное происхождение.

48:37 Вот. То есть система Эйдос является прекрасным средством дополнения базы лемматизации. Если бы она была у профессора Залезняка,

48:44 то эта база лемматизации была бы не 2 млн, а я не знаю, сколько там, три или четыре, или 10 млн было бы

48:50 размером.

Подраздел 4.5: Оценка достоверности моделей и визуализация результатов

48:52 Вот. Теперь, значит, смотрим на сами э-э модели, которые созданы. Они уже созданы

48:58 и проверены на достоверность.

49:01 Уже мы видим, что модели эти преобразуются для отображения в системе достаточно долго,

49:08 потому что они размерность их довольно большая уже.

49:11 И здесь мы видим, э-э что представляют собой эти модели? Колоночки

49:16 представляют собой классы.

49:19 То есть не не классификационные шкалы, а именно классы, градации классификационных шкал.

49:25 Мы здесь видим сначала одну шкалу классификационную - авторы.

49:29 Автор.

49:30 Потом видим вторую классификационную шкалу - жанр. И градации этой шкалы, классы: проза, стихи.

49:37 А потом видим годы.

49:41 И потом сумма

49:43 и средняя там и так далее. И вот здесь вот у нас э-э 700 строк.

49:49 И мы видим, что по первому классу Грибоедов у нас пять фрагментов.

49:54 По второму - семь, по третьему - шесть.

49:57 Ну это всё значит что? Что мы не можем эти числа сравнивать друг с другом, потому что они несопоставимы.

50:04 Из-за того, что тут написано единичка и 10 единичка, это, значит, числа на самом деле сравнивать нельзя, потому что разное количество примеров, вообще говоря, в классах.

50:14 Вот. Ну, например, проза - 24 фрагмента, стихи там 19 фрагментов. Если мы тут и тут видим единичку,

50:21 то эта единичка, она вообще разный смысл имеет. Поэтому мы должны использовать либо

50:26 хи-квадрат для сравнения фактических частот и теоретических,

50:29 либо мы должны перейти к относительным величинам.

50:34 Значит, мы можем перейти к относительным величинам двумя способами: либо э-э делить э-э

50:39 абсолютную частоту на суммарную

50:42 абсолютную частоту,

50:45 либо, значит, можем разделить на число примеров обучающей выборки.

50:50 И мы э-э переведено в проценты, мы видим, что эти числа уже можно сравнивать, они уже являются сопоставимыми.

50:58 Но сравнивать их неохота, особенно учитывая размерность матриц.

51:02 Это причём маленькая такая задачка совершенно э-э учебная сейчас рассматривается.

51:09 И всё равно неохота, потому что достаточно большая размерность уже матриц.

51:15 Поэтому в системе Эйдос автоматизировано сравнение, оно автоматизировано несколькими разными способами сравнения осуществляется.

51:22 В хелпе режим 5.5 описывается, как это делается.

51:26 Матрица абсолютных частот, матрица относительных частот, сравнение

51:30 фактических и теоретических абсолютных частот путём вычитания и путём деления.

51:36 И нормирование к нулю, в случае, если нет зависимости, путём логарифмирования, вычитания единицы. Получается 10 моделей в системе Эйдос. Э-э, получается 14, из них четыре совпадают, получается 10.

51:48 Все они рассчитываются. Смотрим теперь на модель Inf1 Харкевича.

51:53 Значит, э-э здесь мы видим количество информации в битах,

51:56 которое мы получаем о принадлежности того или иного текста тому или иному автору или классу, о принадлежности или непринадлежности тому или иному классу,

52:06 если в этом тексте есть определённое слово.

52:12 В битах нормировано.

52:14 То есть вот слово, например,

52:16 жизненный, оно несёт информацию о том, что это проза, о том, что это не стихи.

52:22 Причём в стихах тоже встречается,

52:24 но э-э реже, чем в среднем по всей выборке. Поэтому его количество информации отрицательное.

52:30 Точно так же мы видим слово жизнь, которое и встречается у Грибоедова и Пушкина, но реже, чем в среднем. А вот у Тургенева и Толстого особенно, оно встречается чаще, чем в среднем.

52:42 Вот. У Достоевского примерно так же, как в среднем.

52:46 Вот. Это уже количество информации, полученное путём сравнения абсолютных

52:51 частот фактических и теоретических, но не путём вычитания, как хи-квадрат, а путём деления.

52:56 И нормирование путём логарифмирования. Получается мера информации Харкевича.

53:02 Смотрим Inf3 - это хи-квадрат. Замечательная

53:06 модель Карла Пирсона. Она очень хорошо сбалансирована.

53:11 Э-э, то есть у неё по строкам, по строкам сумма всех э-э коэффициентов взаимосвязи в строках, по всех строках равна нулю.

53:20 Ну и средняя, соответственно.

53:22 И то же самое и по колонкам.

53:25 То есть сумма э-э коэффициентов взаимосвязи по колонкам тоже равна нулю.

53:31 Это очень хорошо сбалансированная модель, где встречаются модули

53:35 коэффициентов взаимосвязи положительные, отрицательные,

53:39 такие у этих модулей этих коэффициентов взаимосвязи, что вот эти суммы по строкам и колонкам равны нулю. Обычно эта модель даёт очень хорошие формы для анализа.

53:49 Дальше мы здесь не будем другие модели смотреть. Выходим из этого режима

53:53 5.5, и система вываливается. Это, ну, скажем так, э-э сознательная недоработка некоторая. То есть

54:02 я там э-э пытался сделать, чтобы она не вываливалась, но получается всегда ошибка возникает, связанная с

54:09 проблемами восстановления э-э визуальной среды, рефреш.

54:14 Вот. И я там и так, и так, и так пробовал. В конце концов, думаю, да ладно. В общем, сделал просто выход из системы

54:21 без возникновения ошибки, по крайней мере.

54:25 Вот.

54:28 Значит, теперь э-э дальше по плану у нас э-э оценка достоверности моделей.

54:34 То есть модели мы создали, посмотрели.

54:37 И теперь надо выбрать наиболее достоверную модель.

54:41 Это делается в режиме 3.4.

54:43 Значит, если мы наиболее достоверную модель не выберем, а будем использовать какую-нибудь модель, которую мы создали, то надо понимать, что

54:52 решение задач может быть не совсем удачным. Ну то есть идентификация может быть ошибочная,

54:57 прогнозирование может быть ошибочное, принятие решений может быть ошибочное, если модель недостоверна.

55:03 Но при обработке текстов такое практически исключено.

55:07 Значит, почему? Э-э, я это думаю по той причине, что у нас все объекты, все классы переописанные. То есть у нас э-э

55:15 э-э объекты, которые мы использовали для формирования обобщённых образов, они э-э содержат огромное количество признаков,

55:22 очень большое количество признаков.

55:25 И оно является избыточным, это количество признаков, э-э

55:29 более чем достаточным для создания достоверных моделей. То есть мы видим, что все модели, которые здесь есть, как правило, они имеют очень высокую достоверность.

55:39 Это стандартный критерий хи-квадрат. А нет, извините, не хи-квадрат, а э-э коэффициент F-мера Ван Рейсбергена, F-мисури.

55:48 А это вот уже её обобщение, нечёткое мультиклассовое обобщение, а это ещё и норми... э-э инвариантно относительно объёма выборки.

55:56 И мы видим, что все вот эти вот э-э оценки достоверности модели очень высокие

56:03 во всех этих вот э-э методах оценки.

56:07 Которые представляют собой разные интегральные, разные критерии достоверности: F-мера и её обобщение.

56:16 Что за F-мера описано в хелпе, что за обобщение тоже там описано. Что значит инвариантно относительно объёма, всё это описано, всё это я рассказывал.

56:26 Значит, возьмём модель, которая рекомендуется по классическому критерию Ван Рейсбергена, голубым цветом выделена колоночка и строка, где максимальное значение.

56:36 И смотрим, как распределены... Да, вот это сообщение информационное, это и подобные появляются тогда, когда впервые э-э приложение только создано,

56:45 идёт анализ какой-то, и впервые создаются какие-то папочки. В системе Эйдос все графические формы обязательно записываются в свои папочки, поддиректории.

56:56 То есть они там есть, и потом можно их использовать. А можно скриншоты сделать.

57:01 Значит, здесь мы видим э-э положительные решения вправо от оси э-э вертикальной оси Y,

57:07 отрицательные решения о непринадлежности классам слева.

57:11 И видим, что количество положительных решений истинных э-э всегда больше, превосходит, чем количество ложных,

57:19 за исключением очень низких уровней сходства, около 5% есть ошибки.

57:25 Смотрим хи-квадрат. Там по-другому это выглядит.

57:28 Мне больше нравится, как хи-квадрат выглядит.

57:31 То есть при низких уровнях сходства есть ошибки, ложноположительные решения.

57:35 Ложноотрицательные решения тоже есть, но только при очень низких уровнях сходства и очень мало.

57:40 Это решение о непринадлежности классу.

57:43 Значит, если мы посмотрим долю истинных решений, это третья кнопочка, то мы увидим, что быстро возрастает с ростом э-э

57:52 с ростом уровня сходства, быстро растёт доля истинных решений.

57:57 Таким образом, э-э уровень сходства является адекватной мерой степени истинности решения.

58:03 И система Эйдос это была всегда.

58:06 Э-э, первое свидетельство Роспатента на систему Эйдос в девяносто четвёртом году получено.

58:11 Где уже это всё было.

58:15 Вот. То есть уже ей 30 лет в этом году, да? Или сколько там, я не могу посчитать даже.

58:20 24 здесь, и там ещё шесть. Ну да, 30 лет

58:25 назад получено свидетельство Роспатента

58:28 э-э на систему Эйдос.

58:32 А почему раньше не получено? Я она раньше была? Была, ребят, но не получена раньше, потому что раньше их не выдавали. В России не было э-э регистрации авторских прав на программное обеспечение.

58:43 Видно, да?

58:50 И почему я про это так подробно говорю? Потому что буквально несколько дней назад я прочитал, что модель Chat GPT O1, э-э очень э-э прогрессивная, мощная модель, э-э предобученная, генеративная языковая модель

59:05 фирмы OpenAI.

59:07 Ну, наверное, я бы не побоялся слов сказать, гениальная. То есть это выдающееся достижение, безусловно.

59:13 Как и 4 и 4 Omni,

59:16 4О. Ну О1 - это вообще шедевр, то есть ещё более мощная э-э система интеллектуальная.

59:23 Так вот, э-э в статье было написано, что она отличается от предыдущих. Ну это, между прочим, не совсем так.

59:29 не от всех предыдущих. А, да, именно О1 этим отличается,

59:33 что она э-э не просто пишет ответы,

59:37 а, извините, у меня здесь сейчас э-э VPN э-э VPN это э-э

59:43 не включён, поэтому О1 не запустится. Значит, э-э модель О1, ребята, значит, она обладает э-э одним, одна из её отличительных, одно из очень важных её отличительных особенностей этой модели является то, что она способна оценивать э-э адекватность своих ответов. И там, когда даётся ответ, то её внутренний мир демонстрируется, то есть демонстрируется на экране ход рассуждений

00:08 с оценкой степени уверенности в истинности этих этапов рассуждений.

00:14 Э-э, и является огромным достижением, что она оценивает степень истинности своего решения, того, которое она сообщает,

00:23 и говорит, что...

00:24 Да, извините, вас звук пропал, микрофон отключился.

00:29 Ой, извините. Извините, да. Значит, э-э огромным достижением э-э модели О1 является то, что когда вот она пишет ход рассуждений, то там она оценивает ещё и достоверность этих этапов рассуждений.

00:45 И э-э потом в конце она оценивает э-э истинность, степень истинности своего своего ответа.

00:52 То есть она может оценивать критически э-э степень истинности своих решений.

00:57 И это считается огромным достижением этой модели О1. Я хотел сказать, что в системе Эйдос это было всегда.

01:03 И в той версии, которая 30 лет назад была, тоже это было.

01:07 А там был тот же самый, те же самые критерии сходства,

01:11 которые являются адекватной мерой степени истинности решения, потому что мы видим,

01:15 что при увеличении уровня сходства растёт доля истинных решений.

01:20 То есть он является адекватной мерой степени истинности решения. И система Эйдос даёт, порождает нечёткие решения

01:28 в хорошем смысле этого слова, в смысле нечёткой логики.

01:32 Вот. И э-э оценивает степень их истинности. И мы видим,

01:36 как она это делает. В результатах решения задачи, допустим, идентификации, мы видим,

01:43 что здесь вот у нас есть уровни сходства.

01:46 И э-э показана степень истинности решения. Вот смотрите, фрагмент Достоевский 2.

01:53 Э-э она относит его к классу Достоевский, год шестьдесят шестой, 1866, то, что это проза.

02:01 А потом пишет, что это возможно Лермонтов 10%, возможно 840 год.

02:06 А возможно даже и Горький, но 3,6%, возможно 1906 год.

02:11 Вы понимаете, да? То есть она даёт ложноположительные решения, но она в них сама сомневается, отдаёт очень низкий уровень сходства.

02:18 Всего 3% там.

02:21 Вот. Это что значит? Что она даёт нам решение и сама оценивает степень их истинности. А мы видели при анализе достоверности модели, что система э-э оценивает распределение истинных и ложных решений

02:35 в зависимости от уровня сходства. И мы видим, что чем выше уровень сходства, тем больше можно доверять этому решению.

02:42 Что, в общем-то, э-э основано на фактической работе, на фактических результатах работы системы.

02:48 То есть если мы не знаем этих вот, да, птички стоят там, где соответствует истине

02:53 решение.

02:55 Но если мы, значит, э-э не знаем, какое истинное решение, то мы можем предполагать,

03:01 что те решения, где высокий уровень сходства, они обладают более высокой степенью истинности, а те, где низкий уровень сходства, они более низкой степенью истинности, или, может быть, даже являются ложными.

03:12 То есть, э-э то, что сейчас в модели О1 появилось буквально неделю назад там или месяц назад,

03:18 э-э, в статье я прочитал неделю назад, а появилось оно, конечно, немножко раньше, месяц назад.

03:24 Это было всегда в системе Эйдос. Это интересный, между прочим, момент.

Раздел 5: Решение задач идентификации и прогнозирования

Подраздел 5.1: Идентификация (атрибуция) текстов

03:30 Теперь, значит, смотрим, что у нас э-э, когда мы заходим в результат... Теперь начинаем решать задачи в моделях.

03:37 Сначала мы решаем задачу идентификации.

03:41 То есть у нас модель... Будем решать задачи Inf3 или Inf4, которые... Я хочу сказать, что особой роли не играет, какой именно модели, потому что они все обладают очень высокой достоверностью. 0,9 по Рейсбергену - это, ну, практически они высокодостоверные все модели получаются лингвистические.

04:01 Вот. Можно решать задачи идентификации, то есть атрибуции текстов, то есть определение вероятных авторов, определение э-э жанра, ну, проза или стихи, и определение э-э датировки текста.

04:14 А потом и остальные задачи тоже посмотрим.

Подраздел 5.2: Обратная задача прогнозирования (семантические ядра классов)

04:16 Значит, э-э ещё я хочу вам показать, ребята, значит, э-э статью,

04:23 где э-э это всё сделано на значительно больших объёмах данных. Вот такой подобный анализ проведён.

04:30 Это мы выйдем э-э в этот режим, где я показывал.

04:35 И найдём статью.

04:37 Э-э

04:44 Диана Сергеевна Луценко, Диана Сергеевна, которая уже не Луценко.

04:49 Вот, это внучка. Она закончила

04:53 Романо-германскую филологию,

04:56 университет Кубанский государственный.

04:59 И мы с ней написали статью по математической лингвистике, по анализу текстов

05:06 довольно давно это было, я смотрю.

05:13 Вот.

05:17 Когда же это было? Ну здесь описываются модели, методы. А потом идут э-э исходные данные. Видите,

05:23 автор, э-э название, жанр, период,

05:28 год

05:29 и фрагменты текстов. Ну здесь э-э выборка-то ограниченная показана в статье, а вот лемматизация,

05:36 э-э 56.000 слов,

05:39 видите?

05:41 Проанализировано было.

05:43 И потом, значит, другие формы.

05:46 Вот, то, что примерно я вам показываю сейчас, но ещё не всё показал.

05:50 А в конце вот здесь, здесь есть ссылка на интеллектуальное облачное Эйдос-приложение,

05:57 в котором э-э мы можем почитать

06:00 про это исследование.

06:03 И, значит, э-э

06:07 есть ещё здесь статьи.

06:11 Луценко Диана Сергеевна.

06:16 Здесь только одна видна на на русском языке, одна на английском языке.

06:23 Вот.

06:28 Есть статья, где э-э меня и нету. Или я тоже я есть там?

06:34 Есть.

06:39 Значит, э-э почему я это показываю? Потому что здесь э-э в полном варианте это сделано.

06:46 То есть здесь показано больше э-э возможностей различных.

06:51 То есть показано, что есть э-э возможность э-э больше классов, больше классификационных шкал использовать.

06:59 Вот это я хотел показать. И сейчас ещё э-э попробую вам в облачных приложениях показать,

07:05 что там есть интересного. Значит, вот выходим в этот второй пункт, э-э скачать систему Эйдос. А, я можно было этого не делать.

07:15 Можно было прямо сразу. И вот здесь вот у нас есть датасеты.

07:19 Датасеты.

07:21 И здесь вот можно посмотреть текст, слово текст поискать.

07:26 Текст. Он находит семь раз.

07:29 Видите? 193-е приложение.

07:33 Атрибуция текстов песен с применением АСК-анализа.

07:37 Интеллектуальная атрибуция литературных текстов. Это как раз вот сделано. 199-е приложение.

07:44 Новая вкладки откроем его.

07:47 Э-э, анализ АСК-анализ паспортов научных специальностей ВАК. Это мною сделано, написана статья.

07:54 Интересно очень, потому что семантические ядра научных специальностей сформированы.

07:59 И можно взять любую статью или вообще любой текст

08:03 и э-э определить научную специальность, автоматизированно. На какую, то есть какая вот терминология используется, сходная с той, которая в паспорте специальности или нет.

08:14 Атрибуция анонимных псевдонимных текстов с тиражом.

08:19 Вот. И также э-э по научным АСК-анализ классификации текстов по научным специальностям ВАК.

08:25 Новая номенклатура.

08:28 Вот. И потом вот. То есть довольно много есть приложений, э-э иллюстрирующих работу с текстом.

08:35 То есть я это вам показываю, потому что у нас сегодняшняя тема - работа с текстами. То есть можно, э-э, в общем, э-э

08:44 посмотреть, как это сделано.

08:46 И там не только вот этот интерфейс 2 3 2 2 используется, там и другие интерфейсы тоже есть статьи с другими интерфейсами.

08:53 Дальше, значит, мы э-э должны решить задачу идентификации наиболее достоверной модели. Для этого мы

08:59 задаём наиболее достоверную модель Inf3, хи-квадрат.

09:05 И э-э потом в ней и проводим распознавание.

09:12 И смотрим на его результаты.

09:21 Ну вот на такой размерности баз данных мы видим, что это происходит несколько медленнее.

09:35 Присвоение модели статуса текущий. И потом распознавание именно в текущей модели проводится, идентификация.

09:46 Аж 40 секунд на моём компьютере. Само распознавание проводится в режиме 4.1.2.

09:52 У нас была модель Inf7, а теперь мы хотим в модели Inf3.

09:56 Параметры можно использовать по умолчанию.

09:59 Это она сравнила все э-э тексты, которые у нас есть, сравнила с обобщёнными образами классов.

10:08 И получились результаты, которые мы можем посмотреть. Это уже результаты решения задачи идентификации.

10:16 Вот результаты решения этой задачи в модели F3.

10:20 То есть э-э вполне такие разумные результаты. То есть она видит, мы видим, что

10:25 она определяет

10:28 и автора, и год написания, и жанр.

10:34 И обратный результат мы можем в обратной форме посмотреть 4 1 3 2,

10:40 где мы слева видим классы, а справа видим э-э фрагменты текстов.

10:45 И видим, насколько правильно они отнесены системой э-э к классам.

10:49 Мы видим, что вот Пушкин у нас, да, класс, например. И его тексты Пушкин 3, Пушкин 1 в порядке убывания степени сходства.

10:56 Меньше всего похож Пушкин 6 на обобщённый образ Пушкина.

11:02 Всего 56% сходства.

11:05 И мы видим здесь не только э-э э-э

11:11 на э-э какие, значит, э-э

11:15 фрагменты текстов наиболее похожи, но и какие наиболее не похожи.

11:20 Это интересная информация. Вот, допустим, Грибоедов

11:23 4 больше всего не похож на Лермонтов 2,

11:28 на Тургенева, на Достоевского. Видите, вот максимальное различие.

11:32 Это называется конструкт.

11:34 То есть это смысловые полюса конструктов: Грибоедов и

11:38 Лермонтов.

11:42 А Пушкин больше всего на кого не похож? Ну, похож он больше всего на самого себя.

11:47 А не похож больше всего на Лермонтова

11:50 по совокупности всех этих показателей, которые мы использовали.

11:55 Тургенев немножко похож на Достоевского,

11:58 очень незначительно.

12:00 Вот, значит, э-э Достоевский 4 похож на обобщённый образ Тургенева на 8%.

12:07 А Достоевский 5

12:11 похож на 0,5%.

12:13 Остальные просто не похожи.

12:16 И больше всего Тургенев не похож на Грибоедова. Вот такая вот информация мы здесь видим.

12:21 Также мы видим, какие фрагменты текстов, которые относятся к прозе.

12:26 И видим, что больше всего на прозу не похож Пушкин и Грибоедов, и Лермонтов,

12:32 что правильно.

12:34 Ну и, соответственно, на стихи тоже мы видим, кто больше всего похож на стихи.

12:38 Пушкин 3 больше всего похож на стихи.

12:41 А из всех, кто у нас писал стихи, больше всего не похож на стихи Лермонтов 4.

12:47 Он он похож всего на 11-12%.

12:51 Перерыв делаем? На обобщённый образ. А мы сейчас спросим. Э-э, уважаемые слушатели, как у нас насчёт перерыва? Будем делать перерыв или нет? Или продолжить?

13:00 В принципе, мне тут осталось, я не знаю, полчаса, может быть, чтобы закончить.

13:06 Как вы смотрите на этот вопрос?

13:12 А кто меня слушает, поднимите руку там или плюсик наберите.

13:18 Вот.

13:22 Давайте продолжать.

13:24 Э-э, хорошо, хорошо.

13:28 Значит, мы задачу идентификации рассмотрели. Теперь нам интересно узнать, а что ж там сформировала система, какие обобщённые образы классов?

13:37 Э-э, это обратная задача прогнозирования. То есть мы не по факторам определяем э-э будущее состояние, а по будущему состоянию целевому определяем факторы,

13:48 какие необходимо, чтобы получить это состояние.

13:51 То есть смотрим, какие у нас сформировались классы.

13:54 И это у нас, ребята, и есть э-э семантическое ядро автора. То есть выбираем вверху автора.

14:03 А потом нажимаем модель, выбираем.

14:06 И видим, какие слова наиболее характерны для этого автора,

14:11 в отличие его от других авторов.

14:17 Вот. Смотрим на Пушкина, может быть, получше получится.

14:21 Онегин, Людмила, Руслан,

14:23 геройский, Евгений, произведение.

14:26 Смотрим в виде графическом это.

14:29 Здесь только семь указывается наиболее э-э значимых,

14:33 наиболее характерных для Пушкина слов.

14:36 Ну, Евгений Онегин, Руслан и Людмила

14:39 и так далее.

14:42 А больше всего не похожи на Пушкина герой, стихотворение, проблема, поколение, отец,

14:48 Базаров. Ну, про Базарова Пушкин ничего не написал.

14:52 А кто написал? А давай посмотрим.

14:56 Тургенев.

14:58 Базаров, родители, отец, проблема, ребёнок, поколение.

15:04 Отцы и дети.

15:05 Понимаете?

15:07 Вот.

15:13 То есть мы получили семантические ядра и антиядра писателей. То есть мы видим, что слева наиболее характерные слова для того или иного писателя, а справа наименее характерные для него слова.

15:27 Вот Толстой, жизнь, свои, глубокоуха.

15:31 Странно. Нравственный, война,

15:34 человек.

15:36 Вот.

15:38 А Горький что?

15:42 Мать, мать.

15:48 Актуальность, читатель.

15:53 Мы также можем посмотреть, на то, какой смысл у каждого слова. Смысл э-э, извините за каламбурчик, в том смысле,

16:06 о принадлежности какому автору несёт информацию

16:10 то или иное слово.

16:12 Вот, допустим, слово вдумчиво, оно содержит информацию о принадлежности Горькому,

16:18 шестьдесят шестой год и к прозе.

16:21 И похоже, что это не стихи, не Пушкин.

16:25 То есть мы знаем смысл каждого слова теперь, э-э тот, который нас интересует в модели, то есть количество информации об этом слове, о принадлежности текста с этим словом тому или иному автору,

16:38 тому или иному году, тому или иному жанру.

16:44 Вот, допустим, слово былинный - это похоже, что стихи, не проза. Год 840. Это Лермонтов, но и не Пушкин.

16:53 То есть она знает смысл каждого слова. Значит, ну, конечно, можно и в форме нейросети это вывести.

17:01 Но здесь будет у нас сейчас э-э юмор такой своеобразный.

17:07 Специфический юмор.

17:10 Да. Ну, вот это зря, может быть, я даже и сделал.

17:14 Ну, хотел показать, что

17:17 надо э-э задавать параметры, при которых э-э формы имеют э-э осмысленную визуализацию. То есть можно понять смысл из изображения.

17:28 Вот. А сейчас она пытается нарисовать нелокальный нейрон, у которого 700 рецепторов

17:34 с силой влияния этих рецепторов, положительным, отрицательным знаком влияния на, то есть на активацию и торможение нейрона, соответствующего классу.

17:46 Ну, значит, нарисовала в конце концов. Но здесь, значит, я так понял, что ничего здесь не поймёшь.

17:54 Вот. Есть параметры. Вот здесь вот можно задать максимальное число рецепторов 24.

17:59 И визуализация. Тогда и, конечно, быстрее нарисуется и будет более понятным.

18:05 Здесь будут у нас слова, которые наибольшую имеют э-э силу влияния на активацию нейрона, соответствующего Пушкину.

18:13 И которые положительное влияние красным, отрицательное тормозящее синим.

18:19 И толщина линии соответствует весовому коэффициенту, значению весового коэффициента.

18:25 И мы можем посмотреть в сокращённом виде слой нейросети,

18:29 полносвязанная нейросеть.

18:33 Вот.

18:36 Здесь у нас ограничено шестнадцатью нейронами и шестнадцатью рецепторами. Мы можем задавать другое количество.

18:44 Вот. Ну тогда читабельность э-э уменьшится. Значит, вверху мы видим нейроны, соответствующие писателям.

18:51 Да, ребята, вот здесь вот мы видим 5% нейросети, но наиболее значимых связей.

18:56 То есть для каждого автора указаны наиболее характерные и наиболее нехарактерные слова.

19:01 Всего лишь 5%. Но наиболее значимых.

19:05 Сортировка по степени значимости слева направо, по модулю силы связи.

19:09 Можем, можем сделать побольше.

19:13 Э-э, ну, допустим, 50 можем сделать рецепторов.

19:18 Что там у нас блынкает?

19:26 Угу. Ну здесь мы видим уже э-э названия нечитабельные практически, но есть коды.

19:33 И здесь уже 10%.

19:37 Ну и если здесь задать 100,

19:39 вот, а их 700

19:42 рецепторов, а мы 100 задали из 700,

19:46 то тогда мы увидим

19:51 ну, как это выглядит уже, ничего тут не поймёшь. То есть уже и коды не видны. И это 14%

19:59 модели, ребята, отображено.

20:02 14% всего лишь.

20:05 Наиболее значимая информация, которая есть в модели.

20:08 Теперь, э-э, посмотрим, э-э, как распределяются... Да, следующие задачи.

20:12 Задача, э-э, обратная задача идентификации мы посмотрели.

20:17 Теперь, э-э, можно посмотреть задачу, э-э, связанные с исследованием моделируемой предметной области.

20:25 К ним относятся задачи

20:29 кластерного анализа относятся.

20:32 Вот. Ну, давайте, э-э,

20:36 значит, здесь результаты будут, наверное, не особо хорошие,

20:39 потому что, значит, я выбираю модель, запускаю модуль на Питоне.

20:45 Почему? Потому что у нас года есть, и

20:51 кроме э-э слов, которые есть в текстах,

20:55 для классификации этих текстов будет будут играть роль э-э жанр и годы.

21:01 Будут сказываться на группировок, группировках.

21:09 Значит, здесь можно задавать по Y, по X размеры э-э полотна,

21:15 толщину линии.

21:19 Я беру по умолчанию. Здесь показано, какие модели у нас будут использоваться для кластеризации. Это стандартные популярные модели кластеризации. Мне больше всего модель Варда нравится.

21:31 Вот. Ну, в общем, так по в общем,

21:35 но иногда бывают другие лучше дают результаты. И вот здесь вот мы видим, что результаты находятся в папке

21:43 кластеринг classes.

21:46 Вот, модели, то есть папки приложения.

21:50 Вот здесь мы видим матрицу,

21:52 э-э круговую когнитивную диаграмму.

22:01 Круговая когнитивная диаграмма авторов, э-э жанров и лет написания.

22:07 Здесь вверху написано, что это за диаграмма, в какой модели, какое приложение, когда получено, какая всё тут расписывается.

22:15 Какие параметры изображения тоже расписываются внизу здесь написано.

22:20 И мы видим, какие авторы сходны, отличаются друг от друга в виде линий. Э-э, толщина которых означает степень сходства различия, а цвет красный - сходство, синий - различие.

22:31 Также мы видим э-э

22:35 графики увеличения межкластерных расстояний

22:39 и видим э-э результаты кластеризации.

22:43 Ну, например, вот сейчас один из результатов посмотрим.

22:47 Вот она считает, что Лермонтов - это стихи, э-э шестьдесят девятый год Толстой, проза.

22:53 Вот. Ну и дальше других авторов она классифицирует.

22:57 Значит, здесь э-э имеет смысл вот сейчас э-э

23:02 ввести ещё раз данные

23:05 без э-э года и без э-э

23:14 жанра. То есть вторую колоночку мы берём классификационную, только по словам

23:21 будем анализировать авторов, только по словам и всё.

23:25 Вот.

23:39 Я посчитал, что это не совсем правильно, потому что э-э, ну если примерно одинаково у нас или довольно много представлено примеров по разным классам, то это, наверное, приемлемо.

23:34 Но если у нас э-э в разных классах сильно отличается число примеров,

23:39 то тогда э-э более правильным было бы э-э не среднее брать, а а больше с учётом веса этих вот классов по числу примеров.

23:49 Э-э, сейчас я вам покажу когнитивную кластеризацию, результат когнитивной кластеризации.

23:55 Вот. Что я предложил, собственно говоря? Значит, я предложил э-э при объединении классов на основе матрицы сходства, не корректировать матрицу сходства, а пересчитать её. Вот, собственно, ключевая идея.

24:08 Не корректировать, а пересчитать. А как пересчитать? Заново создать модель

24:13 с уже вот такими классами.

24:15 А как это сделать? Я сделал это следующим образом. Я просто в исходных данных скорректировал информацию о том, каким классам принадлежат наблюдения.

24:24 Вот эти вот классы, которые объединены в кластер новый,

24:28 они теперь после этого вот э-э объединения классов, они теперь кодируются как принадлежащие эти наблюдения не тем классам, которые объединились, а новому классу, который уже является кластером.

24:39 И потом заново формирование модели, но не всех, а только вот этой модели, в которой мы кластеризацию проводим.

24:46 И расчёт в этой модели матрицы сходства.

24:51 В результате чего получается у нас э-э корректная новая матрица сходства,

24:58 полученная путём переформирования модели с этими новыми классами,

25:02 соответствующими кластерам.

25:05 То есть всё очень убедительно, логично. Я на статью написал об этом, сейчас покажу ссылку на неё.

25:11 Свидетельство Роспатента оформил

25:13 когда-то давно на это дело.

25:17 Вот. Э-э, чем это интересно? Тем, что результаты кластеризации получаются очень убедительные.

25:22 Я показывал результаты кластеризации с помощью стандартных методов, статистика SPSS, IBM, э-э

25:28 27, например, там десятки кластерограмм этих агглюмеративных дендрограмм

25:35 получал э-э по определённым исследованиям.

25:38 Своё получал, своим методом получал и давал в тёмную э-э профессорам.

25:43 И они говорят так: "Вот это вот не очень похоже, это тоже как-то сомнительно. А вот это хорошая".

25:49 И всегда вот эту дендрограмму агглюмеративную, полученную методом когнитивной кластеризации, всегда они отмечали как очень адекватную, разумную э-э классификацию.

25:59 Эксперты, я имею в виду.

26:02 Вот. То есть это что значит? Что, значит, мы можем очень быстро посчитать много э-э методов кластеризации, используя классических, и получить дендрограммы, которые не особо похожи на правду

26:14 с точки зрения экспертов. А можем чуть подольше посчитать когнитивную э-э кластеризацию

26:21 и получить дендрограмму, которую эксперты оценивают очень положительно.

26:26 То есть я хочу сказать, что метод когнитивной кластеризации даёт результаты, которые экспертам, экспертами оцениваются как наиболее достоверные.

26:37 Вот здесь вот у нас матрица сходства приводится. Она сейчас долго считала, потому что во всех моделях считала.

26:44 Вот, круговая диаграмма. Берём, допустим, Пушкина, модель F3.

26:49 Она нам формирует э-э когнитивную, круговую когнитивную диаграмму,

26:54 где мы видим, что все авторы э-э очень сильно друг от друга отличаются. Кто-то побольше отличается, кто-то поменьше.

27:02 То есть здесь это не очень наглядно мы видим.

27:05 А если мы посмотрим дендрограмму когнитивную, там мы увидим, насколько они друг от друга отличаются. Они все отличаются сильно

27:12 друг от друга. Но всё-таки некоторые больше похожи друг на друга, чем на других.

27:19 Ну то есть они отличаются в меньшей степени, скажем так. Отличаются, но в меньшей степени.

27:25 И вот сейчас мы проводим когнитивную кластеризацию и смотрим

27:29 на результат.

27:39 Всё записывается в папочки, информация об этом была.

27:44 Вот, э-э

27:50 что-то как-то не особо

27:52 мне нравится.

27:59 Вот по тем словам, которые были э-э использованы этими э-э авторами, мы видим

28:57 вот такой результат.

29:03 Кластеризация. Ну, немножко я перехвалил, может быть.

29:07 Значит, теперь смотрим э-э статью

29:11 о когнитивной кластеризации и свидетельство Роспатента.

29:17 Вот.