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

***Российская Федерация***

**107 Теория информации, данные и знания. Лабораторная 4. Завершение лабораторной 3.0.3 интеллектуальная обработка табличных данных 2020-09-29**

Заголовок: Оценка ценности признаков и обработка текстов в системе Эйдос: Лабораторная работа 3.02

Резюме текста

1. Введение и организационные моменты

Семинар посвящен лабораторной работе №4 по дисциплине "Теория информации, данные, знания" для группы ИТ20-02, проводимой 29 сентября 2020 года. Занятие ведут профессора Луценко Е.В. и Аршинов Г.А. Обсуждается необходимость отключения микрофонов участниками для избежания фонового шума.

2. Обзор предыдущей работы и постановка текущей задачи

На предыдущем занятии (лабораторная работа 3.03) группа остановилась на необходимости изучения ценности признаков. Текущая задача – изучить ценность признаков для решения задач идентификации, прогнозирования и принятия решений, а также начать лабораторную работу 3.02, посвященную обработке текстов.

3. Объяснение ценности признаков и шкал

Ценность признака (градации шкалы): Объясняется через меру вариабельности (среднеквадратичное отклонение) количества информации, содержащейся в строке (объекте) относительно среднего значения по классу. Высокая вариабельность указывает на высокую информативность признака для различения классов.

Демонстрация в Excel: Показывается сортировка признаков (свойств) по убыванию их среднеквадратичного отклонения.

Принцип Парето: Иллюстрируется, что небольшая доля признаков (например, 20-30%) часто несет основную часть информации (например, 80%), что позволяет отбирать наиболее значимые признаки. Строится график нарастающим итогом для визуализации.

Ценность шкалы: Определяется как средняя ценность (вариабельность) ее градаций (признаков). Малоценные шкалы можно удалять для сокращения размерности модели без существенной потери информации.

4. Ценность (определенность) классов

Определение: Классы также оцениваются по вариабельности информации в соответствующих им векторах признаков.

Интерпретация: Классы с высокой вариабельностью (например, "вешалка") легко идентифицируются и хорошо детерминированы факторами модели. Классы с низкой вариабельностью (например, "мышка", "телефон" в данном примере) являются "аморфными", слабо детерминированными, их сложнее идентифицировать и достигать в задачах управления. Достоверность идентификации класса коррелирует с его вариабельностью.

5. Введение в Лабораторную работу 3.02: Обработка текстов

Назначение: Демонстрация возможностей системы "Эйдос" по работе с текстовой информацией.

Установка: Описывается процесс установки ЛР 3.02 в системе.

Ввод текстовых данных: Показывается, как тексты (например, из статей) вводятся в систему через Excel, где каждый абзац становится отдельной строкой (примером).

Параметры обработки текстов:

Признаки: В качестве признаков могут выступать целые текстовые поля, отдельные слова или символы.

Лематизация: Возможность приведения слов к их начальной форме (лемме) для унификации. Упоминается словарь Зализняка и возможность добавления новых слов/лемм в базу системы. Можно настроить минимальную длину слова для анализа.

Многоуровневая сеть: Аналогия: символы (рецепторы) -> слова (нейроны 1-го уровня) -> макрохарактеристики текста (автор, жанр, датировка, тема) (нейроны 2-го уровня).

Приложения: Система позволяет решать задачи атрибуции текстов (определение авторства, датировки, жанра, темы), сравнивать лингвистические образы авторов/текстов, выявлять характерные слова/термины для разных классов текстов.

6. Заключение и следующие шаги

Подводятся итоги обсуждения ценности признаков/шкал/классов и возможностей текстового анализа. Следующее занятие будет посвящено работе 3.03 (продолжение) и методике риэлторской оценки. Даются ссылки на статьи по обработке текстов.

Детальная расшифровка текста

1. Введение и организационные моменты

Ну что, ребята. У нас сегодня 29 сентября 2020 года. Четвертая лабораторная работа с группой ИТ20-02. На четвертой паре. Микрофоны выключайте все.

Вот. Четвертая пара, 13:50 - 15:20. Четвертая лабораторная работа с группой Т20-02 по дисциплине Теория информации, данные, знания.

Занятия ведут профессор Луценко Евгений Венеминович, это я, и профессор Аршинов Георгий Александрович.

Всех приветствую.

Да, здравствуйте, ребята. Здравствуйте.

Вот.

Здравствуйте.

Да, здравствуйте, здравствуйте. Вот. Так что сейчас мы посмотрим, на чем мы остановились.

2. Обзор предыдущей работы и постановка текущей задачи

Группа 202. На прошлой лабораторной работе мы остановились на том, что нам нужно теперь сейчас вот изучать ценность признаков и начать изучать лабораторную работу 3.02. То есть ценность признаков для решения задач идентификации, прогнозирования, принятия решений.

3. Объяснение ценности признаков и шкал

Ценность признака (градации шкалы):

Сейчас я объясню, что такое ценность признаков для решения задач идентификации.

Для этого мы посмотрим установленную сейчас лабораторную работу, которую мы изучали, 3.03. И на ней же, на её же примере мы посмотрим, что такое ценность признаков.

Значит, для этого мы посмотрим сначала на сами модели. Вот, допустим, смотрим на какую-то модель. Инф1, к примеру. И в этой модели справа есть колоночка такая: среднеквадратичное отклонение. Что это значит, ребята? Это величина, которая показывает степень отклонения количества информации в строке какой-то определённой от среднего по этой строке. Среднеквадратичное отклонение.

Евгений Венеминович, включите совместный доступ.

Да, да, да, да, да, точно. Я как раз догадался, что нету его.

Вот смотрите. Ну, это в режиме 5.5 делается. Мы видим модель. Смотрим эту модель и вправо двигаемся до конца. И у нас есть колоночка среднеквадратичное отклонение в данном случае количества информации от среднего. То есть рассчитывается суммарное количество информации, среднее на класс и среднее отклонение от среднего, грубо говоря.

Демонстрация в Excel:

Что у нас получается? Посмотрите, некоторые значения свойств 0,229, например, отклонения, другие 0,378, а третье 0,591. Что это значит? Это значит, что у нас разная вариабельность. Одни, если построить графики вот так вот по оси, то получится, что одни сильно отклоняются. Сейчас я прямо их построю эти графики.

Вот. Ну, допустим, вот я беру F3 модель. Сохраняю в Excel. Обращаю ваше внимание на то, что в Экселе они открываются эти базы данных.

Принцип Парето:

И сортирую в порядке убывания по вариабельности. В порядке убывания. Получается, смотрите, есть значения свойств: наличие проводов есть, такое свойство, и наличие проводов нет. У них вариабельность 1,7 количества информации по классам. Потом то же самое, наличие кнопок есть, нет, материал пластмасса, форма круглая. То есть дальше мы видим, что идёт убывание вариабельности. Значит, и потом идут признаки, у которых вообще никакой нет вариабельности. Потому что их вообще нет этих признаков. Есть диапазоны числовые, в которых нету вообще данных, наблюдений.

Что это вот означает эта вариабельность, ребят? Я вам сейчас опишу пример такой. Представьте себе, что у нас есть такие две крайние ситуации. Одна крайняя ситуация, что, ну, можно себе представить, что, допустим, у студентов и студенток есть мобильные телефоны. Мне нужно отличить студентов от студенток. Я спрашиваю: есть мобильный телефон? Мне все говорят: да, да, есть, есть. Вот. Получается что? Что этот ответ на этот вопрос мне вообще не позволяет решить, студент это или студентка. Никакой информации об этом он не содержит ответ на этот вопрос. То же самое касается других таких вопросов, которые касаются признаков, которые есть у всех.

Теперь представьте себе, что я спрашиваю: а брюки есть? Вот. Мне говорят студенты, что есть, и студентки говорят, что есть. Тогда мы получаем уже некоторое количество информации о том, что это студенты или или не студентка из этого признака, что есть брюки. Но и о том, что студент, и о том, что не студентка, получаем очень мало информации, потому что вероятность встретить этот признак в группе студентов 100%, а в группе студентки 98%, а всего 90, э-э, сколько там получается, 99% вероятность этот признак встретить. То есть единица от 99 отличается очень мало, и 98 от 99 отличается очень мало. Поэтому об этом признаке очень мало информации о том, студент это или студентка.

И есть признаки, у которых большое количество информации о том, что это студент и о том, что это студентка, или о том, что это не студент, и о том, что это студентка. Вот, допустим, этот признак длинные волосы. Вот что этот признак, какую информацию он несёт? Он несёт очень много информации о том, что это студентка, и очень много информации о том, что это не студент.

То есть есть признаки, которые вообще не несут информации, есть, которые несут, но мало, а есть, которые несут много. Чем они отличаются? Если построить график изменения информативности в строке, то получится у нас, э-э, вот я сейчас построил семейство графиков. И мы видим, что некоторые графики, они мало отклоняются от нуля. Чуть-чуть информация о принадлежности к чему-то содержится, чуть-чуть информация о непринадлежности содержится. Вот. А есть признаки, которые много содержат информации о принадлежности и непринадлежности. Вот. Вот я сейчас, допустим, вот этот график, он мало от нуля отклоняется, а этот много отклоняется.

Ценность шкалы:

И вот количественно мы можем оценить… Поняли, в общем. Количественно можно оценить степень вариабельности значений в строке этой вот мерой среднеквадратичное отклонение. Можно другие меры использовать: или дисперсию, или среднее отклонение от среднего. Много таких мер существует, которые можно было бы использовать.

Но суть в чём? Если признак содержит много информации о принадлежности классу, много информации о непринадлежности, то у него вариабельность высокая, количество информации в этой строке, соответствующей этому признаку, значению свойства или значению фактора. Если же он почти бесполезен, то там мало информации о принадлежности и мало информации о непринадлежности содержится. Ну, соответственно, признак бесполезен. Мы можем их рассортировать в порядке убывания их степени полезности. Вот. И можем построить график нарастающим итогом. Здесь у нас сначала самые ценные признаки, потом менее ценные, а потом вообще бесполезные идут. А нарастающим итогом как сделать? Берём, значит, эту это значение перетаскиваем сюда, а потом суммируем предыдущее значение плюс текущее значение. Вот. И дальше это тянем. У нас получается некая кривулечка уже нарастающим итогом.

Видите как? То есть здесь сначала самые ценные идут признаки, которые больше всего информации о принадлежности и непринадлежности содержат. Потом идут признаки, которые меньше такой информации содержат. А потом признаки, которые вообще такой информации не содержат полезной. Вот я сейчас вам показал, как это рассчитывается. В системе Эйдос тоже это есть. То есть мы можем перейти в режим 375 и посмотреть, какова ценность. Ну, сразу же появляется сообщение, что будет графическая форма записана.

Вот. И мы видим такую же вот кривую, как я вам показывал по модели ABS. И так вот по всем моделям, их не так много, 10 моделей. Нажимаем Escape, переходим на следующую, на следующую модель. И вот видим, допустим, модель F3. В чём здесь смысл вообще вот этого графика? Вот смотрите, здесь написано: 50% наиболее значимых признаков обеспечивают почти 80%, 79% суммарной значимости. Вот 50% признаков, 80% суммарной значимости. А 30% наиболее значимых признаков обеспечивают 50% суммарной значимости. То есть, когда у нас статистика побольше, чем вот в этой маленькой модели, тогда эта кривулька, она вообще вот такой вид имеет. Она резко растёт вверх, а потом почти горизонтально идёт. Это называется Парето-подмножество. Парето когда-то математик итальянский, он предложил такую идею, что где-то, он так условно сформулировал это, ну, скажем, при этом и числа были. В общем, небольшое число факторов в основном определяют ситуацию, а остальные все факторы, большое количество факторов, особой роли не играют. Ну он сказал это так примерно, что 5% наиболее сильно влияющих факторов обуславливают ситуацию, а 95% факторов практически роли не играют. То есть есть какое-то небольшое число факторов, которые в основном влияют на ситуацию.

И вот мы сейчас посмотрим. Вот это тоже ещё более ярко выражено, эта модель более достоверная. Вообще есть такая закономерность, что чем более достоверная модель, тем сильнее выражена вот эта кривулька, логистическая называется. Сначала быстро растёт, потом почти не изменяется. Вот здесь 18% наиболее значимых признаков обеспечивают 50% значимости, а 50% обеспечивают 85% суммарной значимости. То есть эта кривая более выгнутая, чем модели F3. Это и коррелирует с её достоверностью модели. Чем сильнее она выгнута, тем более достоверная модель.

А вот здесь вот у нас табличные приведены формы, экселевские причём, которые прямо эту информацию содержат, но в табличном виде. Вот, значит, если мы сейчас перейдём вот сюда, то мы увидим. Вот, допустим, берём модель и видим, что здесь у нас наиболее ценным признаком является наличие проводов и отсутствие. Имеется вот такая вот значимость этого в натуральном выражении, в процентном отношении. И мы это всё уже видели в виде графиков и в системе Эйдос, и в экселевском файле. Вот.

Какой всего этого смысл? Такой, что можно удалить менее значимые признаки, которые основной, в общем-то, особого вклада не вносят. Но дело в том, что признаки являются градациями шкал. То есть мы не можем удалить признак, мы можем именно градацию шкалы удалить. Поэтому давайте посмотрим, как это делается. Значит, для этого нам нужно оценить ценность не признаков для решения задачи идентификации, прогнозирования, решения, а именно определить ценность шкал. Это делается очень быстро на основе той информации, которая имеется, даже для больших размерностей моделей тоже быстро. И формируются формы выходные по значимости шкал уже. Вот, допустим, значимость описательной шкалы F3. Берём, смотрим, что там у нас? Значимость описательной шкалы F3. В модели F3, я имею в виду. Значит, смотрите, значит, у нас получается, что наиболее ценными являются шкалы наличие проводов, кнопок, экрана, а почти бесполезные размер два, цвет, наименее ценными шкалами. То есть если мы хотим сократить размерность модели и не потерять при этом ничего ценного, то, что есть в ней, то нужно удалять шкалы целые. То есть мы берём вот эту таблицу экселевскую и вот эту колоночку размер два, цвет, материал, просто берём и удаляем. Получаем в результате модель гораздо меньшей размерности, которая содержит наиболее ценную информацию. Ну, на такой маленькой модели они все имеют ценность определённую, эти шкалы. Но когда большая модель, большое количество данных, шкал, объектов, тогда получается ярко выраженная логистическая кривая, вот такая изогнутая. И там видно уже, что действительно можно удалить и будет нормально. То есть она вот так идёт. Вот. И ясно, что вот эти вот последние можно удалить без ущерба для модели. Ну, здесь в данном случае пока нет. Но показать можно, как это выглядит.

Это всё, что касается значимости шкал.

4. Ценность (определенность) классов

Определение:

Теперь есть нечто аналогичное по классам. Мы можем сказать, что некоторые классы характеризуются высокой вариабельностью количества информации по признакам, которые их характеризуют эти классы. А некоторые классы наоборот, вариабельность очень низкая по ним.

Интерпретация:

Что это значит, если перевести на русский язык? Как это можно интерпретировать? Значит, э-э, классы, по которым высокая вариабельность количества информации в моделях, то есть берём колонку и строим график по колонке. Получаем график либо близкий к нулю, либо сильно отклоняющийся от нуля или от среднего. Выходные формы, где у нас эта информация содержится. И то же самое по классификационным шкалам есть информация. Какие классификационные шкалы оправданы, а какие нет. Здесь просто формируются выходные формы экселевские, которые можно посмотреть.

И мы видим, что некоторые классы характеризуются высокой степенью вариабельности. Вот. А некоторые нет. Вот обобщающий класс, например, здесь он более выражен, чем конкретные.

Теперь, что считается? Давайте я вам сейчас скажу, что считается ценностью шкалы. Ценность градации я вам объяснил, вполне, так сказать, наглядно. То ценность градации - это вариабельность количества информации в модели, в строке, соответствующей значению свойства или значению фактора. Вот. А как относительно шкалы? Если взять шкалу, там много градаций, каждая имеет какую-то свою ценность, выше, ниже. Я посчитал, недолго думая, грубо говоря, так, честно, если сказать, что ценность шкалы является средней ценностью её градаций. Если все градации шкалы имеют высокую ценность, то и шкала будет иметь высокую ценность. Если в основном градации имеют низкую ценность, только одна какая-то градация более-менее ценная, тогда, значит, шкала не очень ценная. Ну, примерно так.

То есть мы видим результаты такого анализа. Значит, я могу вам сказать, что те классы, которые описываются малыми значениями информативности, то есть мало информации о принадлежности классу, мало информации о непринадлежности. Вот берём, смотрим какую-то вот класс, например. И видим, что в каких-то какие-то классы описаны… Вот у нас тут среднеквадратичное отклонение по классам. Вот класс вешалка, он является хорошо описанным, 0,537 вариабельность. А вот класс клавиатура в меньшей степени 308, а вот класс мышка вообще низкая вариабельность у него этого класса. То есть мало отличаются информация, которая в нём содержится вот в этих координатах. То есть это это среднеквадратичное отклонение координат этого класса. Вот, вектора этого. То получается, что оно вообще мало отличается от среднего. У других классов значительно больше. Ну вот у класса 235 телефон тоже низкая вариабельность.

То есть я могу вот что сказать, что чем выше вариабельность координат вектора, тем этот вектор легче идентифицировать. То есть для системы распознавания меньше проблем с идентификацией этого класса, описанного таким вектором, координаты которого имеют высокую вариабельность. Вот и всё. То есть если мы посмотрим на степень достоверности идентификации различных классов, то она коррелирует с этой вариабельностью координат по классу. 4 1 3 Достоверность идентификации классов по F-мере рассортируем. Вот вешалка лучше всего идентифицируется. Ну у неё и вариабельность самая высокая координат этого вектора класса соответствующего. А клавиатура, монитор, мышка плохо идентифицируются. Вот. По F-крит... по F-мере. То есть это там, это не означает, что они не идентифицируются, а означает, что больше ошибок, ложноположительных решений. Ну, ложноотрицательных здесь нет. То есть больше ложноположительных решений в данном случае. То есть общая достоверность идентификации этого класса ниже с учётом того, что есть ложные решения. А эти однозначно идентифицируются. Здесь ложных решений практически нет. Видите, достоверность 100%.

Вот вам и всё, так сказать, всё, что отсюда можно извлечь. То есть мы можем сказать, что если бы, что если мы эти свойства, значения свойств интерпретируем как факторы, которые влияют на объект моделирования, то можно сказать так: есть факторы, которые жёстко обуславливают поведение объекта моделирования. То есть сильно влияют на его переход в те или иные состояния. Это координаты вектора класса, которые имеют большие значения. Они сильно влияют на принадлежность или непринадлежность к этому классу. А есть координаты или значения свойств, значения факторов, которые мало несут информации о принадлежности или непринадлежности к этому классу. Так вот, если класс описан вот только такими слабо, ну, скажем так, координатами такими описан, которые мало отличаются от нуля, то такой класс трудно идентифицировать и трудно его получить в системах управления. Если он поставлен будет как целевой, то непонятно, как его перевести объект управления в это состояние, соответствующее этому классу. Потому что факторы влияют на него слабо на переход объекта моделирования в это состояние, соответствующее этому классу. Это слабо детерминированный класс, я так называю. То есть у него система детерминации, факторы, которые обуславливают переход в это состояние, соответствующее классу, они не оказывают существенного влияния на этот процесс перехода, они слабо влияют на этот процесс. Это называются классы аморфные, статистические, слабо детерминированные. Вот. А есть классы, которые жёстко детерминированы, которые есть факторы, которые прямо жёстко обуславливают переход в соответствующее состояние с высокими очень модулем силы влияния, положительным или отрицательным. Ну, положительным обуславливает, отрицательным препятствует. То есть чётко дифференцированные классы. И мы это видим. И при идентификации мы видим, что это связано с достоверностью. Хорошо идентифицируются классы те, которые жёстко детерминированы, почти как в поисковой системе, информационно-поисковой системе. А плохо те, которые слабо детерминированы.

5. Введение в Лабораторную работу 3.02: Обработка текстов

Назначение и установка:

Всё, значит, мы рассмотрели вопрос, ребята, о том, что мы на предыдущем занятии должны остановились. Это ценность признаков и степень сформированности классов. И начинаем изучать новую лабораторную работу 3.02, которая иллюстрирует возможности системы по работе с текстами. Вот, работа 3.02. Начинаем её изучать.

Для этого нужно её установить эту работу. Для того, чтобы её установить, выбираем все ранее созданные приложения. Переходим в режим 1.3. Выбираем работу 3.02. И нажимаем О'кей. Ну, она предупреждает нас система, что сейчас там вот в этой папочке, где исходные данные, все данные будут удалены. Это и происходит, они удаляются.

Ввод текстовых данных:

Потом туда из папочки лабораторных работ переписывается из папочки соответствующей лабораторной работы, переписывается файл исходных данных в папочку исходных данных. И вот этот файл параметров окошка программного интерфейса. Вот переписывается файл исходных данных.

Параметры обработки текстов:

И запускается программный интерфейс ввода данных из табличных из таблиц. Ну а в таблицах может быть и тексты, могут быть и числа, а могут быть и тексты. Если там тексты, ребята, тогда задаётся вот в этой экранной форме опция ставится: применять специальную интерпретацию текстовых полей классов, применять специальную интерпретацию текстовых полей признаков. Что это обеспечивает эта опция? Она обеспечивает… И здесь вот обратите внимание, не проводить лематизацию. Она обеспечивает рассмотрение в качестве признаков либо значения поля целиком текстового, либо рассмотрение в качестве признаков слов. Но здесь вот я пришёл к выводу, что нет смысла рассматривать как признаки слова, состоящие из одной-двух букв. Допустим, такие как "не" там или "а". Такие нет смысла рассматривать. И здесь вот больше трёх букв рассматриваются, больше трёх символов. Можно это поставить ноль, например, больше нуля. Вот. И есть ещё вариант рассматривать значения полей как символы. Тогда у нас получается как бы многоуровневая сеть. На нижнем уровне получается рецепторы - это символы, а нейроны - это слова. Мы, кстати, у нас есть такая работа лабораторная 3.01 распознавание слов по входящим в них символам, символах. Это, значит, первый слой сети. На нижнем уровне рецепторы - символы, нейроны - слова. На втором уровне сети в качестве рецепторов мы рассматриваем нейроны предыдущего слоя, то есть слова. Вот. А в качестве нейронов рассматриваем, ну, допустим, какие-то характеристики текста. Я их называю макрохарактеристики текста. Это могут быть такие характеристики текста. Значит, автор, авторство, датировка текста, смысловая направленность текста, содержательная, жанр текста. Ну, жанр, например, может быть какой там, трагедия там, оперетта. Вот, если там театр. Исторический роман, детектив, фантастика, фэнтези. Ну там фантастику можно разделить на техническую, там фэнтези, сказочную, да, допустим, фантастику. Вот. Можно разделить там ещё какие-то жанры упомянуть и направленности текстов. Содержательная сторона направленности текстов, можно написать там текст о там том-то, том-то. Ну, допустим, текст о мотоциклах ралли, например, там или там какие ещё там есть супермотоциклы? Ну, например, там Триумф там, например, да?

Приложения:

И вот, значит, что мы, как мы вводим такой текст, ребята, это в качестве источника информации для модели. Это файл исходных данных. Значит, если я возьму, открою файл какой-нибудь, вот давайте сейчас я возьму, открою файл свою статью открою. Вот открыл свою статью. Вот смотрите, я беру вот так и какие-то абзацы выделяю блоком. Вот так беру и выделяю блоком абзацы. Помещаю в буфер обмена. Запускаю Excel. Ставлю курсор вот сюда, к примеру. Первая строка - это у нас будет строка наименований. Здесь о чём это описание фрагмента текста, это авторство текста, например. А здесь вот я просто беру и вставляю. Что я получаю в результате? Я получаю абзацы в виде отдельных в отдельных строках. То есть у меня каждый абзац, ребята, текста, посмотрите, вот видите, абзацы, вот они: первый, второй, третий, четвёртый, да? Символ конца абзаца срабатывает, когда мы помещаем текст в Excel, как э-э, команда перехода на следующее поле, следующую строку. У нас получилось каждый абзац в отдельном, в отдельной строке. То есть это очень удобно и просто вводить тексты в экселевские таблицы. Что это даёт? Вот такое, такой способ ввода текста.

Смотрите, значит, у нас здесь указано источник, автор, текст. И в качестве источника указаны абзацы: первый, второй, третий, четвёртый, пятый. Тексты, ребята, это тексты не самих этих авторов, это тексты об этих авторах из школьных сочинений. И этот текст о Достоевском, этот о Достоевском, этот о Достоевском, этот о Достоевском. Аналогично вот о Гоголе. О Гоголе у нас тут несколько фрагментов, и они переименованы, ну, пронумерованы эти фрагменты. О Горьком, Грибоедове, Лермонтове, Пушкине, Тургеневе, Толстом. Ну, вы понимаете прекрасно, что без проблем можно было сделать гораздо больше. Я сделал столько именно, чтобы просто продемонстрировать вам возможность обработки текстов.

Потом я сейчас вам дам ссылочку, сразу не отходя от кассы, раз я вспомнил, то сейчас прямо сразу и покажу вам. Вот. Значит, вот ссылочка на статьи, в основном мои статьи и соавторов, посвящённые обработке текстов. Я вам просто кидаю ссылочку. Могу скинуть всю полностью эту страничку. Ну давайте всю страничку тоже. Прикольно будет. Здесь есть статьи разных периодов времени. И более новые статьи, они более такие крутые, более развитые. Ну, похоже, что этот чат, он ограничен по объёму. Туда слишком много текста не поместишь. Поэтому вот ссылочка вам, пожалуйста. Открываете ссылочку, там работы. Вот они. Они тоже со ссылочками. Есть работы самые старые 2004 года где-то, вот 2003 года. Атрибуция текстов как обобщённая задача идентификации и прогнозирования. Четвёртого, четырнадцатого, четырнадцатого, семнадцатого, восемнадцатого года, девятнадцатого и двадцатого. То есть на протяжении многих лет я к этой теме возвращался и писал статьи по анализу текстов. И недавно ещё одну написал.

Вот. Значит, что можно сказать об этом, о том, как система Эйдос позволяет обрабатывать тексты? Вот в чат я вам послал информацию об этом. Значит, автоматизированная системно-когнитивный анализ текстов позволяет, слушайте дальше внимательно, формировать обобщённые лингвистические образы классов, семантические ядра, на основе фрагментов или примеров, относящихся к ним текстов, причём на любом языке. То есть я беру, привожу примеры текстов определённого вида, определённой датировки, определённого автора, определённого жанра, определённой смысловой направленности. И один и тот же фрагмент текста, он одновременно является текстом от определённого автора, определённого жанра, определённого года и определённой датировки, э-э, смысловой направленности. Одновременно, то есть он одновременно относится ко всем этим классам. И образы, соответствующие этим классам, формируются на основе этих примеров текстов. А так, значит, формируются обобщённые лингвистические образы этих классов, соответствующих лингвистическим, на основе этих примеров. После этого мы можем количественно сравнивать лингвистический образ конкретного человека или текста или описание объекта с обобщёнными лингвистическими образами групп, классов. Значит, тут требуется пояснение некоторое, ребят. Значит, дело в том, что описание словами является самым начальным уровнем вербализации. То есть то, что у нас там интуитивно мы понимаем, мы можем описать словами. Ну, конечно, при этом происходит необратимая потеря информации. Не все могут одинаково хорошо описывать словами свои интуитивные эти субъективные образы. У одних это получается круто, это обычно писатели, поэты. У других так себе получается, а третьи вообще, они понимают всё как бы, а сказать толком не могут. То есть они не обладают таким даром или у них не отработаны навыки вербализации их собственных мыслей, они не могут выразить свою мысль словами.

Так вот, ребята, если мы всё-таки можем выразить мысль словами, то мы можем описать, скажем, различные объекты словами просто. Вот те же самые объекты, которые мы сейчас в работе 3.03 рассматривали, мы можем словами описать. Ту же самую мышку там, компьютер, телефон, монитор там, клавиатуру, мы можем описать словами, что это такое. После этого, ребята, система Эйдос уже может формировать обобщённые образы этих объектов, описанных словами. Ей этого достаточно. Она слова рассматривает как признаки объектов. То есть можно словами описать всё, что угодно: место какое-то на территории университета или какую-то дисциплину, или какие-то объекты, ситуации. И всё это можно обобщать и сравнивать. Потом, когда мы сформировали обобщённые образы, ну, допустим, мы можем взять фабулу преступлений по определённым статьям административного кодекса, например. И взять фабулы по определённой статье, потом взять фабулы по другой статье. И сформировать обобщённое описание, лингвистическое, то есть создать семантическое ядро, это называется, или лингвистический образ класса этой статьи административного кодекса. То есть потом после этого мы можем взять любую фабулу реального, допустим, преступления и сравнить с этими обобщёнными образами и сказать, что похоже больше всего, что вот по этой статье идёт дело. Понимаете? Или по этим вот двум статьям идут дела, если они имеют одинаково высокий уровень сходства, близкий.

Потом мы можем сравнить обобщённые лингвистические образы классов сами друг с другом и создать их кластеры и конструкты. То есть определить, что такие-то вот писатели больше похожи друг на друга, такие-то меньше. Работы, которые написаны в такие-то годы, они сходны. Работы, которые такого-то жанра, они сходны. То есть вот мы можем вот такие проводить исследования с текстами. Вот.

Потом мы можем исследовать моделируемую предметную область путём исследования её лингвистической модели. То есть мы можем сказать, какие слова наиболее характерными являются для описания того или иного класса. Вот, допустим, такой жанр - комедия, например. Какие слова наиболее характерны для комедии, какие для трагедии, какие для водевиля, какие там для исторического романа? Какие слова наиболее характерные для этих видов текстов? Мы можем сказать, что стихи, допустим, Горького похожи на стихи Пушкина, например. Больше, чем на прозу Тургенева. То есть мы можем сравнивать тексты друг с другом.

Мы можем определять вероятное авторство. Вот эта атрибуция текстов - это имеется в виду, что определяется вероятный автор текста. То есть иногда бывает, что тексты подписываются не фамилией автора, а фамилией каких-то людей, которые заказали этот текст, например, автору. Он его написал по каким-то причинам. И в качестве авторов его там нет. И можно определить, кто на самом деле автор этого текста по специфике этого текста, по специфике терминологии, по специфике сочетания слов. Они называются мемы, сочетания слов.

Ну что, ребята? Значит, я вам пришлю просто название работы на чат. А вы сами её найдёте без труда эту работу. В наших условиях современных это не представляет никаких проблем. Эта работа, она очень интересная. Вы, наверное, слышали про Павлова, вторую сигнальную систему. Значит, вот я вам могу сказать так, что Павлов со второй сигнальной системой - это старшая детсадовская группа по сравнению с тем, что вот написано в книге Ричарда Броди "Психические вирусы". А это уже, так сказать, докторская диссертация, вот если сравнить с Павловым. То есть это далёкое-далёкое развитие того направления, которое, ну, скажем так, основы которого заложил Павлов.

И вот, можно определять вероятность авторства текстов не только анонимных, у которых нет автора, но и псевдонимных, то есть у которых есть авторы, но есть подозрение, что это не авторы, а просто там они написаны как авторы. Можно также определять датировку текста, жанр, смысловую направленность и другие аспекты текстов. Ну, например, датировку, как можно определять? Если мы возьмём, исследуем научные статьи разных лет и составим лингвистический образ статей определённого года, вообще всех статей во всех областях науки, то обнаружится, что в определённый момент вспышка происходит числа статей, скачкообразное увеличение числа статей определённой тематики. Например, взяли, открыли геном, сразу пошёл вал статей по геному. Открыли, значит, интеллектуальные технологии, статья Тюринга пошла, пошёл вал статей по интеллектуальным технологиям после статей Тюринга. И так вот точно так же и потом по ядерным технологиям, по космическим технологиям. Запустили спутник в пятьдесят седьмом году Советский Союз. Слово "спутник" стало международным словом. Сразу же его стали использовать на всех языках мира. Оно так и звучит "спутник" по всех языках мира. И означает именно то же самое, что означает в нашем языке. И сразу пошёл вал статей по космическим технологиям, по спутникам, станциям, перелётам, двигателям, выходам на орбиту, с посадкам и тому подобное, тому подобное и так далее, и тому подобное.

Так вот, фактически, что это значит? Что в текстах есть слова, которые характерны для определённого времени. И можно, это слово не одно, таких слов много, характерных для того или иного времени. И можно, анализируя, узнать, во-первых, анализируя тексты разных времён, узнать, какие слова характерны для того или иного времени. То есть составить семантические ядра соответствующих периодов времени, которые есть в модели. И потом можно датировать тексты. То есть вот, допустим, есть у нас некий древний текст, древний. Говорят: "Ну, интересно, когда же этот текст написан?" По некоторым предположениям, ему около 4.000 лет назад он написан. Берём мы этот текст, в систему загоняем, а там у нас есть примеры текстов, написанных и 4.000 лет назад, и 3.000, и 2.000, и вот так вот до нашего времени. Ну, не через 1.000 лет, а через, там, скажем, 100 лет хотя бы, период, через 100 лет они идут тексты, датируются. До 100 лет можно датировать, грубо говоря. И определяем датировку этого текста, который предположительно написан около 4.000 лет назад. А у него датировка получается 1930 год или тридцать пятый год по той терминологии, которая там применяется, ребята.

Вот и все дела, понимаете? То есть можно сразу сказать, что это не может быть 4.000 лет назад, потому что там написано про кривизну пространства, про замедление времени, про большой взрыв, откуда возникла Вселенная, про геном и так далее, и так далее, про космические технологии написано. Об этом, ребята, 4.000 лет назад никто не писал, понимаете? За исключением Виманы Шастрика и Махабхараты и Бхагавадгиты, где описывались эти виманы там, как они вели войну Пандавы и Шауравы там, Кауравы. И при этом использовали летательные аппараты. И царевич Арджуна выбросил у верховного главнокомандующего божественного войска Индры, ну это что-то такой персонаж в индийской, я не знаю, как это, пантеоне богов, похож на Архангела Михаила. И Индра дал ему такие всё-таки виману боевую, но оружия не дал, стрелы богов. И царевич Арджуна со своими слугами нагрузили её брёвнами, камнями и летали над войсками противника и сбрасывали эти брёвна и камни, чем приводили их в ужас. Но потом они быстро поняли, что это не стрелы богов, и ничего страшного, ничего такого какого-то колоссального урона не произойдёт. Дело в том, что если бы мы дали стрелы богов, он моментально бы уничтожил цивилизацию. Поэтому Индре хватило ума не давать ему эти стрелы богов. Как обезьяне дать пистолет. А не то что автомат Калашникова.

Теперь, что интересно ещё, что те методы, которые применяются, они совершенно не связаны с природой языка и может быть использован абсолютно любой язык, как естественный, так и искусственный, ребята. То есть можно точно так же анализировать языки программирования, системы кодирования. Вот, допустим, можно взять, какие есть там UTF-8 там, система кодирования там, э-э, там 866, Windows 1251 и так далее, куча всяких систем кодирования, да, символов. Таблиц ASCII, крутых таблиц разных языков. Можно узнать, какой таблицей закодирован текст. Каким образом? Берём мы тексты, закодированные в разных системах кодировках, в разных системах кодирования, и определяем, какие коды, с какой частотой там встречаются, какое лингвистическое семантическое ядро соответствующего текста. И определяем, что семантическое ядро такое, а это ядро соответствует UTF-8 или там э-э, какой-то другой кодировке. Вот и все дела. То есть сейчас вот могу эти кодировки не будем смотреть. Ну, в общем, понятно, что я сказал.

Вот. И это можно использовать в ряде областей, в широком круге областей, не только в лингвистике, но и в других областях можно это использовать то, что я сейчас рассказал. То, что я сейчас вот рассказываю, всё это относится к такому направлению науки, которое называется математическая лингвистика или применение интеллектуальных технологий в лингвистике, в обработке текстов. Но на самом деле, я вам сказал, что можно это использовать, скажем, в юриспруденции. То есть можно определять статью кодекса по фабуле преступления. Это уже похоже не на лингвистику, а на юриспруденцию это похоже.

Так вот, смотрите, ребят. Что здесь делается? Здесь просто задаются параметры, что у нас вот в этом экселевском файле, который мы видели, в качестве колонки с классами мы рассматриваем вторую колонку. Вот, видите, вторая. Сам тип файла у нас старый Excel. А в качестве колонки с описательными шкалами мы рассматриваем колоночку третью. Но сам аппарат, который здесь применяется, и программная система, и алгоритмы работы, они совершенно не ограничены тем, что одна колонка классов, одна колонка описательных шкал. Может быть несколько колонок классов, ну, скажем, автор, жанр, датировка, смысловая направленность. Вот. И даже можно написать статья Уголовного кодекса, статья Гражданского кодекса. А потом тексты. Текст один, текст два, текст три, текст четыре и так далее. Текст один - это, допустим, название произведения или сайта. Текст два - это, допустим, аннотация. Текст три - это, например, там сам текст этого произведения. Если текст не помещается в ячейку Excel, ну, ясное дело, что вы знаете, что 32 КБ ячейка Excel, не каждый текст туда влезет, то нет проблем, так по абзацам его разбить просто и написать здесь один и тот же класс, к которому он относится, и всё. То есть просто этот класс протянуть, и он протянется и размножится. Никаких проблем технических нет. То есть, в принципе, это делается всё очень просто и быстро.

Вот. Теперь, что такое лематизация, ребят? Значит, э-э, слова в русском языке есть исходные, ну, в других языках тоже нечто подобное, но, по-моему, в русском это особенно развито. Вот, скажем, стол. Какие слова произошли от слова стол? Кто знает? Давайте. Говорите. Будем играть в такую игру. Кто скажет последнее, получит самоэкзамен. Кто скажет последнее слово, которое произошло от слова стол? Юрий Александрович, какое слово от слова стол? Столешница. Столица. Кто ещё может что-нибудь сказать? Давайте. Столб, да, наверное? Столешница, столик. Вот. Столяр. О, классно! Столяр, правильно. Ещё кто? Похоже, то, что столяр, похоже, он уже идёт на самоэкзамен. Ну, в общем, вы поняли, да? Вот. И, значит, есть такой словарь лематизации. Наиболее полный словарь лематизации русского языка разработан академиком РАН Зализняком, который не так давно скончался. Это очень интересный человек, блестящий учёный, человек замечательный сам. То есть про него вот я читал просто, знаю. В общем, это просто, э-э, как умница, понимаете? Вот он действительно сделал огромную, ценную работу. Вот его статья про него. Ну, это та статья, где я нашёл его базу лематизации, представленную в виде экселевского файла. Значит, он провёл такую работу, ребята, он нашёл лемы для 2 млн русских слов. Представляете? Теперь я включил базу лематизации, сделал небольшую программку, которая этот файл экселевский преобразовала в базу данных нормальную. Там ведь 2 млн, а в Экселе самое больше миллион, да? Поэтому они что сделали? Они взяли в одном листе две колоночки: первый миллион и ещё две колоночки справа - второй миллион. Ну я взял это всё, преобразовал в ДБС, вытащил всё с левой колоночки, а потом взял то, что справа, продолжил дальше. Вот. Так вот в системе Эйдос есть режим лематизации. Вот. Значит, э-э, когда мы обрабатываем слова, то есть мы когда вот находим слова вот здесь вот, то они могут быть лематизированы. Давай, раз я про это говорю, то и покажу. Вот. Ну, в принципе, вот всё. Нажимаем, файл текстовый преобразуется, считывается, начинается лематизация. Лематизация - процесс поиска с помощью индексного массива в базе лематизации, поиск леммы, ну то есть самого слова, сначала самого слова, а потом леммы. Лемма - это первообразное слово, исходное, из которого оно образовано. Вот. Вот, допустим, мы слово столяр найдём, и она нам напишет слева стол. Вот. И вытащит это слово стол. И заменит слово столяр на слово стол. Слово столица заменит на слово стол. Ну, может быть, это не всегда оправдано, не знаю. Потом есть лематизация двух уровней. Есть примитивная лематизация, есть глубокая лематизация. Примитивная просто первообразное слово вытаскивается из базы. А глубокая лематизация ещё есть слова, глаголы, ребята, которые происходят от существительных. Вот. Ну я не лингвист, мне сейчас трудно так на лету привести пример, но я могу вам сказать, что вы можете зайти в интернете написать онлайн лематизация. Вот. То получается, знаете как? Как в немецком переводе получается. Вернее, как вот немцы говорили, которые не знают русского языка, с очень сильным немецким акцентом и без падежов. Или без падежей, как правильно? Вот. И получается что? Что вы можете это изучить всё. Так вот, глубокая лематизация, то есть поверхностная, она приводит слово к исходному, а глубокая, она ещё переводит глаголы в существительные соответствующие исходные.

Значит, теперь смотрите, вот на лематизацию текста, текст состоит, как вот у нас там есть, включает 1243 слова. Затрачено 48 секунд. Ну, 4 секунды на одно слово. Базу… Теперь дальше слушайте внимательно. В базу лематизации Лема ДБФ добавлено 144 новых слов, новых слов. Необходимо указать для них лемы. В системе Эйдос есть сервисный режим, ну, в смысле, обычная экранная форма, удобная для добавления этих вот лем. И теперь смотрите, значит, вот у нас новые слова, вот здесь вот отмечены признаком New. Вот. Надо указать лему для каждого слова вот здесь. Своего - свой, качествами - качество, собою - не знаю даже, товарища - товарищ. Понятно, да? То есть вот здесь справа надо написать лемы. Ну это можно и в Экселе сделать, кстати, без проблем.

Ну, дальше я хочу вот что сказать. Если бы у академика Залезняка была система Эйдос, то в базе Залезняка было бы не 2 млн слов, а уже миллионов пять, наверное, было бы. Потому что любую, я вам хочу подчеркнуть, любую, какую бы задачу бы не решали в системе Эйдос, связанную с текстами, если мы зададим лематизацию, добавляются сотни тысяч слов в базу лематизации Залезняка. Это что означает? Это означает, что они там не были найдены, они там не были найдены, это новые слова, их там нет в базе Залезняка. То есть мы можем активно развивать эту базу и сделать, чтобы там было не 2 млн, а 20 млн, понимаете? Просто если этим заниматься. Просто вот порешать эти задачки с обработкой текстов, книжки позагонять туда в систему сейчас. Вот. Кстати, есть другие интерфейсы, где тоже лематизация используется, и тоже это же самое делается. И у вас база лематизации будет уже совершенно другой. Ну, правда, надо будет филологов приглашать, чтобы они писали эти вот лемы справа там или слева, где там они.

В общем, выходим теперь на создание модели. Создаётся классификационная описательная шкала и градации и генерируется обучающая выборка. Предлагается нам посмотреть на это всё. Мы посмотрим сейчас. Установлена работа 3.02 атрибут анонимных псевдонимных текстов. Теперь дальше, ребята, изумительное, я хочу вам вещь сказать, изумительную мысль. Вот мы сейчас создали обобщён, ещё мы не создали, но написали, создали справочник классов. Вот. В качестве классов у нас выступают сами писатели, о которых эти тексты. И создали базу описательных шкал, где у нас градациями являются сами слова, а точнее лемы этих слов. Если бы мы здесь лемы не поставили признак лемы, а просто самые слова бы использовали, то здесь бы были бы слова в разных склонениях. Одно и то же слово в разных склонениях. И было бы около 800 слов, там, по-моему, чуть меньше 900. Вот. То есть лематизация уменьшает количество слов. И мы видим обучающую выборку. Смотрите, ребята, сейчас вот я хочу показать вам вот справа. Обратите внимание на правое нижнее окошко. А там у нас коды слов в соответствии с этим вот справочником. Есть фрагменты текста маленькие, где там всего несколько слов, а есть большие. И вот здесь вот у нас в системе Эйдос, как вы понимаете прекрасно, нет никакого принципиального ограничения на количество кодов в этих таблицах и количество обучающих объектов обучающей выборки. Ограничение есть одно, оно не жёсткое. Вот эта вот база данных этих кодов признаков должна быть меньше 2 ГБ. И то же самое классов, и то же самое наименование объектов выборки. Это всё вместе должно быть где-то в пределах 6 ГБ. Много это или мало? Ну я оценивал, ребят, когда вот я же пытался обрабатывать задачи очень большой размерности. Это где-то около 5 млн примеров. Вот если у нас будет не 43 фрагмента текста, а где-то около 5 млн, тогда мы будем находиться вблизи границ возможностей системы Эйдос по вот этим базам данных обучающей выборки.

И теперь нам нужно что сделать? Значит, что здесь изумительного во всём этом? Что обработка числовой информации и текстовой информации дальше совершается одинаково практически. Вот то, что мы проходили с вами на работе 3.03. В начале вроде как по-другому, и таблицы исходных данных сильно отличаются, и параметры ввода этой информации в систему отличаются. А потом получается практически то же самое: справочник классов получился, справочник значений факторов или значений свойств. А потом получил и обучающая выборка получилась. В общем-то, которая принципиально особо ничем не отличается от тех, что там были, да, вообще такая же практически.

А теперь смотрим, ребята, синтез модели. Вот здесь вот уже не надо убирать птичку с графического процессора, иначе мы будем ждать до конца занятия. И, значит, вот так запускаем просто и смотрим. Вот у нас создаются эти модели и проверяются на достоверность. Вы, наверное, замечаете, что она работает довольно шустро. Значит, из-за того, что используется графический процессор. Теперь смотрите, что у нас получилось? У нас получилась достоверность моделей. Значит, по критерию Ван Рисбергена модель F4 семантический резонанс знаний 0,989 достоверность. А по критерию L1 эта модель имеет достоверность, знаете какую? Один. То есть нет ни одного ложноположительного, ложноотрицательного решения. Вот. И 986, 0,986. Это что означает вообще? Это означает, что так оно, кстати, обычно и получается, что модель получилась исключительно хорошая. Значит, я хочу вам сказать, ребята, чтобы вы просто это были в курсе. Значит, что модели лингвистические получаются очень высокой достоверности всегда практически. Сколько вот я видел, всегда получается очень высокая достоверность. Почему вопрос возникает? Значит, ну у меня такая гипотеза. Представьте себе, что почему система может неправильно относить объекты к классам? Ну, потому что не хватает информации у неё об этих объектах. Не хватает описания. Описание слишком короткое, недостаточное, чтобы принять однозначное безошибочное решение. А когда мы описываем словами тексты, там, ребята, этих слов, ну я так мягко скажу, в десятки раз больше, чем необходимо, чтобы однозначно безошибочно эти тексты идентифицировать.

Смотрим результаты идентификации. То есть результаты достоверности очень высокие. Ну, допустим, я вот беру модель и показываю вам, как это выглядит. Значит, ну здесь, может быть, не очень наглядно, но, в общем, здесь решение ложных очень мало, до 9%. А потом идут, начиная с тридцати, идут только истинные решения. Отрицательных решений ложных нет вообще. Но сейчас посмотрим на квадрат, она обычно более такая понятная. Вот. Значит, начиная с 35% идут преобладающие истинные решения. Вот. Смотрим на это. Начиная с 12%, намного превосходит число истинных решений число ложных решений. Ну это, возможно, не очень понятно. Сейчас я покажу формы сами 4 1 3 1.

Вот смотрите. Вот я показываю Достоевский. Видите? 93% фрагмент Достоевский 1 на 93% похож на обобщённый образ класса Достоевский. И на 1,5% похож на Горького. На остальных не похож. И вот мы смотрим, ребята, очень получается красивая картина. Причём это не самая достоверная модель. Сейчас я вот возьму сейчас самую достоверную модель запущу. То есть система совершенно безошибочно относит тексты тем классам, которым они относятся. Потому что задачи, связанные с обработкой текстов, являются переописанными задачами, переопределёнными. То есть там избыточное описание. То есть там описание значительно более подробное, чем необходимо для достоверного решения.

Вот смотрите, что мы видим? Видите? Картина. Ну вы уже видели первую лабораторную работу 3.03. И теперь вам есть с чем сравнивать. Вы видите, что эта модель, она гораздо лучше, чем модель работы 3.03. Здесь просто практически нет ошибок. И явно, так сказать, правильно решаются задачи. Смотрим теперь обратную форму 4 1 5. Какая задача, кстати, решается? Задача атрибуции текстов, определение авторства. Вот, допустим, фрагменты текста Грибоедова. Один похож на 93%, а другой самый не похожий на 66%. Остальные просто не похожи, понимаете? Вот. Пушкин. Значит, смотрим на Пушкина. На Пушкина то же самое. Все тексты Пушкина похожи от 87 до 55%. А Толстой самый похожий на Пушкина - минус 3%. Тургенев. А вот Тургенев, ребят, смотрите, он все фрагменты текста о Тургеневе от 89 до 62% сходства. И ещё есть Лермонтов 6, 1,6% сходства. Ложноположительное решение вылетело, видите? И ещё есть Достоевский и Достоевский Достоевский 4 и Достоевский 5 тоже на ложноположительное решение. Но каков уровень сходства? Достоевский 4 на 0,6% похож на Тургенева, а Достоевский 5 на 0,9% похож, а Лермонтов на 1,5% похож, на 1,6 на Тургенева.

Вот вы сейчас бы, представьте себе, проверяемая статья в антиплагиате, и она оказывается похожая на статью какого-то автора на 0,6%. Это плагиат или нет, ребят? Да, конечно, нет. Вот если там процентов 70-80, то тогда уже возникает вопрос. И вот мы видим вот так же по другим авторам. Автор Горький, обратите внимание, у него нет такой большой чёткости в разделении с другими авторами. У него Толстой похож на 8% на Горького, Тургенев на 1,5%. Лермонтов похож от 85 до 26%. То есть есть фрагменты текста Лермонтова, которые не очень похожи на обобщающий класс Лермонтов. И есть, смотрите, фрагменты текста Достоевского, Тургенева, которые на 9 и на 3% сходны. То есть мы можем, и вот мы видим, насколько эти авторы похожи друг на друга по своим фрагментам текстов. Но, в общем, мы видим, что задача решена очень хорошо. То есть если мы берём уровень сходства, начиная там с 40%, как вот мы видели в графике, то вообще вопросов не возникает. Все все решения безошибочные.

Теперь давайте попробуем этих авторов сравнить. То есть мы сравнили фрагменты текста с обобщёнными образами авторов. У нас получилось очень хорошо. Теперь давайте мы сравним эти обобщённые эти фрагменты текста, то есть обобщённые образы классов, сформированные на основе фрагментов текстов различных авторов, сравним эти обобщённые образы классов друг с другом. Мы уже это сделали. Теперь мы только запускаем выходную форму. И смотрим, насколько эти авторы похожи друг на друга. У нас получается интересная картина, такая немножко такая разочаровывающая, что они все активно не похожи друг на друга. Одни чуть-чуть не похожи, другие сильно не похожи. Вот, допустим, Горький чуть-чуть не похож на Толстого на 8%. А вот, допустим, Пушкин сильно не похож на Достоевского на 23%. Понятно, да? О чём речь идёт?

Но всё-таки они, скажем так, все не похожи, но но в разной степени не похожи. Поэтому можем вывести дендрограмму и посмотреть вот эту степень непохожести. В дендрограмме это видно.

Значит, что мы получаем? Смотрите, они все друг на друга не похожи. То есть у них объединение кластера происходит на очень высоком уровне различия. Видите, здесь? То есть не здесь вот происходит объединение, а далеко по шкале различий. Но всё-таки одни ближе объединяются, другие ещё дальше. Вот Толстой и Горький, они сильно не похожи друг на друга, но они всё-таки больше похожи, чем ещё на Пушкина. Или ещё на Гоголя. И всё-таки, вот если взять всех этих писателей, мы видим, что они образуют два больших кластера, обозначенные синим цветом и красным. Значит, в синем цвете Лермонтов, Тургенев, Достоевский. Особенно Тургенев и Достоевский похожи друг на друга сравнительно, если сравнивать с другими. И к ним примкнувший Лермонтов, видите, да? А с другая группа - это Грибоедов, Гоголь, Пушкин, Толстой и Горький. Причём такая вот интересная картина, что Грибоедов больше похож на Гоголя, Пушкин сам по себе, Толстой на Горького похож. И всё-таки Пушкин тяготеет к Толстому и Горькому. Понятно? Так вот получается.

То есть мы видим вполне, как эти авторы, так сказать, сходны, отличаются друг от друга. Если такую же точно анализ провести по статьям в научном журнале, ребят, то получается совершенно иная картина, совершенно. Очень такая прикольная. Вот это график межкластерных расстояний. Здесь ещё кластеры можно сказать, сходные объединяют класс, а здесь уже явно отличающиеся.

Значит, сейчас я вам даже может покажу дендрограмму статей из новой из новой статьи. Вот смотрите, ребята, вот это дендрограмма статей. Есть различие между писателями и авторами статей? Как вы считаете? Какое различие в этих дендрограммах бросается в глаза сразу же моментально? Можете сказать, что сразу бросается в этой дендрограмме в глаза, в отличие от дендрограммы сравнения классических классиков литературы русской от авторов статей в журнале? Жду 3 секунды. Кто-нибудь скажет или нет? Статьи слишком похожи. Правильно, правильно, молодец! Есть статьи разных авторов, которые очень, ну прямо очень похожи друг на друга, вот очень похожи. То есть главная особенность классиков русской литературы - это их индивидуальность, ребята. Они узнаваемы, они очень не похожи друг на друга. А если они вот похожи, ну тогда возникает вопрос: а с какой стати они похожи, друг? Вот один был классиком, вот он взял, создал целое направление в литературе новое и жанры там, и всё, и стили и так далее. А другой посмотрел, думает: "О, классно у него получилось. Давай-ка я, пожалуй, тоже возьму и точно так же буду писать". И стал писать, ну, стараться стал писать, как Пушкин. Но, во-первых, не получится, как Пушкин. Но будет видно, что ты пытаешься быть похожим на Пушкина. Какую это вызовет реакцию в литературном сообществе? Что тебе скажут? Скажут: "Молодой человек, опомнись, ты ж не Пушкин. И внешне не похож, и потому что у тебя получается, вообще не похоже". И что делать? А делать вот что: нужно просто быть личностью. Вот взять Пелевина, например. Вот я я читал Пелевина, я вообще в шоке, если так между нами. Допустим, "Чапаев и Пустота", вот такой вот. Вот просто смотришь, вот это, ребята, личность, понимаете, индивидуальность. Ну здорово вообще выглядит. Виктор Владимирович, у меня занятие. У меня не у меня занятие непрерывно с 8:00 до 9:30, непрерывно. Всё.

Вот. Короче, если вы напишете что-то своё, это будет очень ценно и гораздо более ценно, чем если вы будете пытаться кого-то пародировать или копировать. Это вот моё личное мнение. Это будет для людей гораздо интереснее, чем если вы кого-то попытаетесь скопировать, кто, в общем-то, может быть, молодец, в общем-то, Пушкин, конечно, да, Пушкин, да, говорят. Но смысла его копировать нет, понимаете? Вот в чём дело. Он бы и сам бы посмеялся, если бы кто-то попытался бы это сделать.

Вот это вот и есть, собственно говоря… Да, здесь другой интерфейс ввода данных использовался. А нет, здесь этот интерфейс использовался, извините. Вот, этот же. Ну и что мы можем ещё сказать? А мы можем посмотреть, а что характерно для того или иного писателя? Какие слова или термины научные характерны, скажем, для Пушкина? Как вы думаете? Онегин. Ожидаемо, да? Людмила, Руслан, Евгений. Очень даже ожидаемо. Горький. Мать. Ну, во-первых, это тексты не самих этих авторов, а о них. Поэтому здесь сама фамилия автора тоже встречается. Мать, ребята. Вот. А Грибоедов? Комедия. У него Софья там у него есть. Годы, много очень указано разных лет. То есть он годы указывает, прямо привязывается к годам. Его особенность, вот он годы прямо привязывается к годам. Вот. А Тургенев? Базаров. Понимаете? Родители, проблемы, поколения. Как вам это нравится, а, ребят?

Я хочу о чём сказать? О том, что были тексты школьных сочинений. Возникает вопрос: какие наиболее характерные особенности текста, допустим, Горького, Толстого, Пушкина? Вот система искусственного интеллекта в один момент, ребята, правильно это определяет, понимаете? Достоевский, до преступления, раскольник, раскольников. Понимаете? Родион совершаем бедный, гуманный там и так далее, замечательно, злодеяния. То есть моментально мы видим, что является наиболее характерным для того или иного автора. Вот эти слова, которые наиболее характерны для того или иного автора, они называются семантическим ядром этого автора. То есть какие слова он наиболее часто использует в сравнении с другими авторами? То есть вероятность встретить этих слов у этого автора намного выше, чем в среднем по всей выборке. Если туда меня включить в этот перечень, то там тоже вполне однозначно сразу же мы определим. Сейчас. Смотрите, ребята. Сейчас будет вам интересно. Если я вообще не вижу. А, вот. Вот смотрите, ребят. Я не знаю, видно вам или нет? Луценко вот здесь вот написано. Сейчас я попробую пошире сделать. Это сколько у нас процентов? Это у нас 220. Ну давайте 250 сделаем. Вот, видите, что здесь Луценко, вот здесь вот написано, ребят. Вот Луценко. Видно, нет? Читабельно? Видно. Вот смотрите, на кого похожа статья Луценко? Есть ли какая-нибудь статья, на которую статья Луценко очень похожа? Ну это меня имеется в виду, Луценко ЕВ. Есть ли статья, на которую она очень похожа? Как вот здесь вот? А? Нету. Вот мы видим вот здесь вот авторы, вот очень похожи. Вот этот автор на этого, этот на этого, понимаете? А вот этот автор, он не похож на других авторов. Вот самый похожий на него вот этот вот, он похож на очень высоком уровне различия, то есть он не похож на самом деле. Вот. Это что означает? Эта статья сильно отличается от других статей, которые есть в этой выборке. То есть она очень оригинальная. Вот если бы можно было рекомендовать на основе такого анализа, обратить внимание на те или иные публикации, то можно было бы сказать: вот эти статьи, они все похожи друг на друга. Вот это похоже, это похоже. Это вы уже читали такое, вы это уже слышали. А вот такого вы ещё не видели. Вот почитайте, посмотрите, что там вообще, о чём речь идёт. Вот. Потому что статья сильно отличается по своей идеологии от остальных.

6. Заключение и следующие шаги

Вот. Ну, у нас и дело идёт к концу занятия. Вверху мы видим авторов нейроны, внизу слова наиболее характерные, наиболее нехарактерные для этих авторов. Вот. И всего мы здесь видим 7% нейронной сети, наиболее значимые связи, которые здесь у нас были в этом приложении. 7% мы сейчас отобразили. Можем отобразить больше, но там тогда ничего не поймёшь. А так, в принципе, достаточно.

Теперь давайте, ребята, какие у вас вопросы есть по изложенному материалу? То есть мы рассмотрели вам решение простых задач, связанных с обработкой текстов. Я вам советую вот эти работы, которые здесь по ссылочке, посмотреть, если вы немножко заинтересовались хотя бы. Там есть работы, в которых в развитой форме показано, как это можно применять в разных областях. Прямо вот так, знаете, не скупясь, я взял и написал, и описал, как это можно делать всё. Вот. Ещё 5 минут, Люсь. 5 минут. Вот. 5 минут, не знаю, что делать теперь, потому что я прервал изложение. Ну… Можно сказать ещё что-то. Кластерный анализ слов мы здесь не проводим, потому что очень много их, и большой размерности будут матрицы сходства. То есть это уже сложно будет интерпретировать. Вот. Ну, единственное, что могу сказать, что вот ещё есть один интерфейс ввода текстовой информации. Это файловый интерфейс. То есть здесь можно книжки вводить. Нам тоже можно в Экселе взять книжку, ввести по абзацам и работать с ней. Но здесь можно прямо целиком экселевский, то есть этот самый, текстовый файл ввести и работать с текстовыми файлами. Тоже можно лематизацию проводить, всё. То есть это файловый интерфейс. То есть не в строчках в Экселе, а прямо вот в файлах вводится информация. Я пробовал книжки анализировать, вполне всё нормально получается. В этих вот статьях, которые там приведены по ссылочке, описано, как это всё делается, как получается. Вот, ребята. На этом занятие заканчивается. До свидания. До свидания. До свидания. До свидания. Я вам эту работу по ссылочке, вот посмотрите, почитайте. У вас возникнет более развитое представление, чем то, которое я отдал. Ну я о нём сказал, что есть такая возможность, а вы там увидите. Хорошо. Хорошо. До свидания.