|  |  |
| --- | --- |
| Gerb-BMSTU_01 | Министерство науки и высшего образования Российской Федерации  Федеральное государственное бюджетное образовательное учреждение  высшего образования  «Московский государственный технический университет  имени Н.Э. Баумана  (национальный исследовательский университет)»  (МГТУ им. Н.Э. Баумана) |
| ФАКУЛЬТЕТ «Информатика и системы управления»  КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии» (ИУ7)  РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА  К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ  НА ТЕМУ:  «Метод выделения шаблонов мультимодальной коммуникации в корпусе REC»  Студент группы ИУ7-82Б Н. А. Жарова  (Подпись, дата)  Руководитель ВКР В. М. Градов  (Подпись, дата)  Консультант Л. Л. Волкова  (Подпись, дата)  Нормоконтролер Ю. В. Строганов  (Подпись, дата) | |
| 2021 г. | |

# РЕФЕРАТ

Расчетно-пояснительная записка с. 62, рис. 20, табл. 5, источников 10, прил.1.

DATA MINING, ПОИСК ШАБЛОНОВ, ВЫДЕЛЕНИЕ АССОЦИАТИВНЫХ ПРАВИЛ, МУЛЬТИМОДАЛЬНАЯ КОММУНИКАЦИЯ

Объектом исследования является Русскоязычный эмоциональный корпус – REC (Russian Emotional Corpus, НИЦ «Курчатовский институт»).

Целью данной работы является разработка метода анализа шаблонов мультимодальной коммуникации (ШМК) в корпусе REC.

Поставленная цель достигается за счёт анализа корпуса REC, разработки метода анализа шаблонов мультимодальной коммуникации и ПО реализующего данный метод, а также апробации разработанного метода на данных корпуса.

В первом разделе проводится анализ корпуса REC. Задача выделения шаблонов определяется как поиск ассоциативных правил. Для упрощения процесса поиска шаблонов при подготовке поведенческих сценариев предлагается использовать кластерный анализ.

Во втором разделе представлен разработанный метод анализа шаблонов мультимодальной коммуникации. Описываются алгоритмы выделения транзакций и особенности применения ConApriori и алгоритмов кластеризации на данных корпуса REC. Приводится разработанная модель визуализации ШМК.

В третьем разделе описываются ключевые моменты программной реализации разработанного метода, приводятся результаты тестирования разработанного ПО RECAnalyzer, а также руководство пользователя.

В четвёртом разделе представлены результаты апробации разработанного метода на данных корпуса. Приведены рекомендации по настройке каждого этапа метода и результаты их применения.

СОДЕРЖАНИЕ

[РЕФЕРАТ 5](#_Toc10734301)

[ВВЕДЕНИЕ 8](#_Toc10734302)

[1 Аналитический раздел 10](#_Toc10734303)

[1.1 Разметка корпуса REC в ELAN 10](#_Toc10734304)

[1.1.1 Описание формата ETF 12](#_Toc10734305)

[1.1.2 Описание формата EAF 14](#_Toc10734306)

[1.1.3 Шаблон разметки корпуса REC 17](#_Toc10734307)

[1.2 Интеллектуальное извлечение данных 23](#_Toc10734308)

[1.2.1 Основные задачи Data Mining 23](#_Toc10734309)

[1.2.2 Поиск ассоциативных правил 25](#_Toc10734310)

[1.3 Выводы 27](#_Toc10734311)

[2 Конструкторский раздел 29](#_Toc10734312)

[2.1 Метод выделения шаблонов в REC 29](#_Toc10734313)

[2.2 Разбор EAF-файлов 29](#_Toc10734314)

[2.3 Метод преобразования данных 33](#_Toc10734315)

[2.4 Выводы 35](#_Toc10734316)

[3 Технологический раздел 37](#_Toc10734317)

[3.1 Выбор технологии программирования, языка программирования и среды разработки 37](#_Toc10734318)

[3.2 Основные классы 37](#_Toc10734319)

[3.3 Руководство пользователя 40](#_Toc10734320)

[3.4 Выводы 47](#_Toc10734321)

[4 Исследовательский раздел 48](#_Toc10734322)

[4.1 Разбиение на одинаковые интервалы времени 48](#_Toc10734323)

[4.2 Разбиение в соответствии с аннотациями на главном слое 49](#_Toc10734324)

[4.3 Разбиение в соответствии с моментами времени 52](#_Toc10734325)

[4.4 Пример использования разработанного метода для более детально анализа 52](#_Toc10734326)

[4.5 Выводы 56](#_Toc10734327)

[ЗАКЛЮЧЕНИЕ 58](#_Toc10734328)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 59](#_Toc10734329)

[ПРИЛОЖЕНИЕ А 61](#_Toc10734330)

# ВВЕДЕНИЕ

Корпус как информационно-справочная система на данный момент является важным инструментом в лингвистике. Корпус обычно основывается на наборе текстов/видеозаписей/аудиозаписей, размеченных в соответствии с правилами, определяющимися, в частности, предметной областью, для которой данный корпус разрабатывался. Наиболее полными, как правило, являются т.н. национальные корпусы, которые создаются с целью максимально полно представить конкретный язык и используются в качестве базы для большинства лингвистических исследований. Примером такого корпуса могут служить Национальный корпус русского языка (НКРЯ), который активно развивается с 2003 года, и Генеральный интернет-корпус русского языка (ГИКРЯ). Существуют и более специализированные корпусы, такие как Русскоязычный эмоциональный корпус – REC (Russian Emotional Corpus, НИЦ «Курчатовский институт»). Корпусы имеют важное прикладное значение и могут быть использованы как базы знаний в различных областях: от машинного перевода до выявления потенциально противоправных текстов.

Корпус REC используется для создания поведенческих сценариев для аффективного робота-собеседника Ф-2. В работе [1] указывается, что важной задачей при моделировании эмоционального поведения является анализ корпуса. На основе этого анализа выделяются поведенческие шаблоны, которые должен воспроизвести робот, чтобы выразить некоторую коммуникативную функцию: продемонстрировать радость или грусть, обратиться к человеку, выразить задумчивость и др. Так, в корпусе размечено мультимодальное коммуникативное поведение людей, выражаемое речью, мимикой и жестами: для видеозаписи человека, коммуницирующего в эмоциональной ситуации (подкорпус экзаменов, подкорпус счастливых людей), приведена разметка элементов поведения. Единица разметки — событие — включает в себя метки времени начала и окончания, метку слоя (например, слой рук, слой движений головы, слой текста и др.) и название выполняемого действия (например, «облизывает губы», «манипулирует волосами», «улыбается»). Нахождение зависимостей в таких данных, в частности, повторяющихся цепочек мимических и жестовых элементов, которые могут быть привязанных к определенной коммуникативной функции (радость, страх, огорчение и др.), позволит выделять на материале корпуса паттерны поведения и в дальнейщем переносить подобные элементы коммуникативного поведения на робота-собеседника Ф-2.

Таким образом, целью данной работы является разработка метода анализа шаблонов мультимодальной коммуникации в корпусе REC. Стоит отметить, что разрабатываемый метод может стать полезным инструментом не только для моделирования эмоционального поведения роботов, но и стать основой для автоматизации процесса разметки новых видеозаписей, тогда как в настоящее время разметка производится вручную.

Для достижения поставленной цели необходимо решить следующие задачи:

— провести анализ корпуса REC и изучить существующие методы интеллектуального извлечения данных;

— разработать метод анализа шаблонов в корпусе REC;

— разработать ПО, реализующее разработанный метод и обеспечивающее возможность визуализации данных корпуса и просмотра результатов анализа;

— провести анализ результатов апробации разработанного программного продукта на тестовом наборе данных из корпуса REC и сделать выводы о применимости разработанного метода.

## Аналитический раздел

### Разметка корпуса REC в ELAN

Как уже говорилось выше, корпус REC – это Русскоязычный эмоциональный корпус [1]. Он разрабатывается с 2008 года и на данный момент содержит 815 размеченных вручную видеозаписей. Для выполнения данной работы был предоставлен доступ к 157 файлам с разметкой в формате EAF и файлу шаблона разметки в формате ETF. Разметка видеозаписей производилась в программе ELAN [2] Интерфейс программы представлен на рис. 1.1.

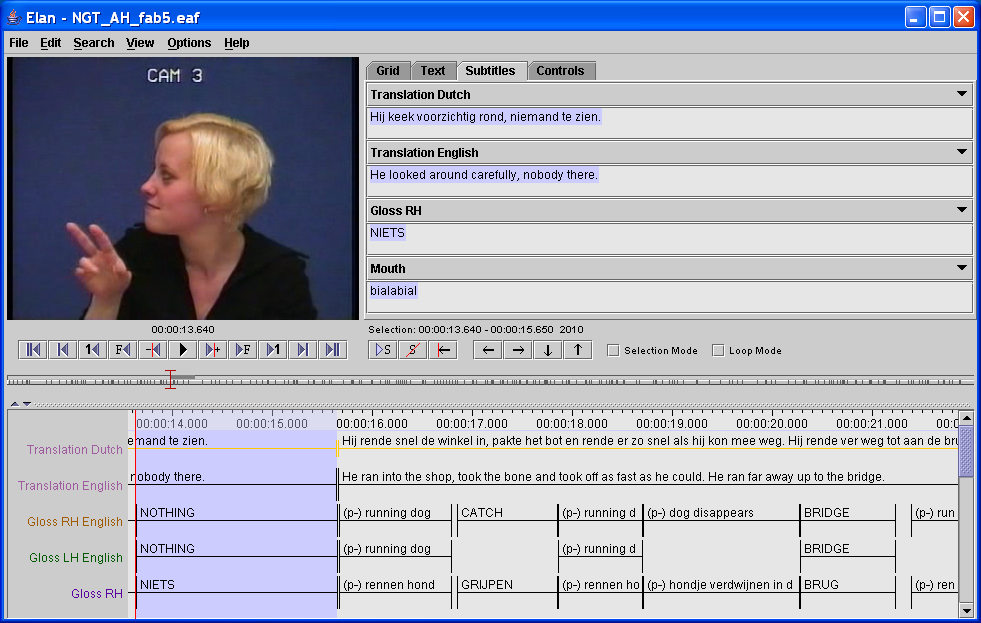


Рисунок 1.1 — Интерфейс программы ELAN.

В соответствии с документацией, представленной на сайте [2], в программе ELAN используются два основных понятия: слои и аннотации. Аннотация – это некоторое событие или его характеристика, привязанные ко времени размечаемого файла. Слой представляет собой контейнер для аннотаций. Аннотации одного слоя не пересекаются.

Слои могут создавать иерархию с помощью отношений «родитель» – «ребёнок», при этом аннотации на дочерних слоях обязаны быть внутри временного интервала родительской аннотации. На дочерние слои могут накладываться ограничения одного из четырёх типов:

— Symbolic Association – временной интервал дочерней ассоциации в точности совпадает с временным интервалом родительской аннотации;

— Symbolic Subdivision – временной интервал родительской аннотации равномерно разбивается на число дочерних аннотаций, и дочерние аннотации по порядку привязываются к полученным временным интервалам. Обычно такое ограничение используется, чтобы показать в целом порядок возникновения событий без точной привязки к моментам времени;

— Time Subdivision – дочерние аннотации разбивают временной интервал родительской аннотации на подынтервалы без промежутков. Данное ограничение похоже на Symbolic Subdivision, но в отличие от него при Time Subdivision дочерние аннотации имеют точную привязку к моментам времени;

— Included In – дочерние аннотации находятся внутри временного интервала родительской аннотации, при этом между ними могут быть промежутки.

Эти четыре типа соответствуют различным способам разграничения элементов разметки. На рис. 1.2 приведен пример для разграничения слов в звуковой дорожке: соотнесение 1 родительского элемента с единственным дочерним, равномерное распределение интервалов дочерних элементов, неравномерное (ручное) и неравномерное с промежутками между ними.

Кроме того, на дочерние слои накладываются следующее ограничение: у слоя с ограничением символьного типа (Symbolic Association, Symbolic Subdivision) дочерние слои должны также иметь ограничение символьного типа.

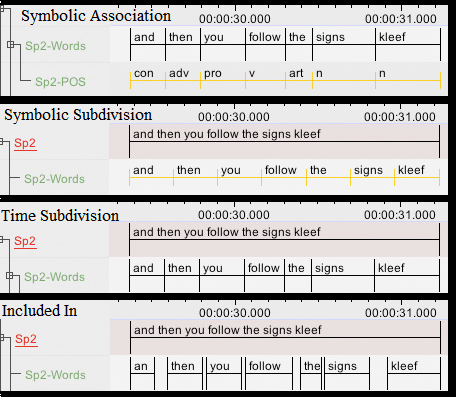


Рисунок 1.2 — Примеры использования ограничений

ELAN поддерживает шаблоны разметки, которые хранятся в файлах ETF.

#### Описание формата ETF

Шаблон разметки (ETF-файл) – это XML-файл, содержащий теги, которые будут использоваться при создании файлов разметки. Основной задачей шаблона служит описание слоёв разметки, их взаимосвязей и ограничений.

Все теги находятся внутри тега ANNOTATION\_DOCUMENT. Атрибутами этого тега являются автор, дата, формат, версия и некоторая другая общая информация о файле.

Далее идёт тег HEADER, который описывает связанные с данной разметкой медиа-файлы (название, путь, временные единицы (обычно миллисекунды) и т.д.). В шаблоне атрибуты пустые, а в файлах разметки будут описывать размечаемые медиа-файлы.

Далее идёт *список временных слотов* (тег TIME\_ORDER). В шаблоне он пустой.

А затем идёт описание слоёв, лингвистических типов, ограничений и словарей.

Собственно, *слой* разметки (тег TIER) обладает следующими атрибутами:

— DEFAULT\_LOCALE – языковой стандарт («ru»);

— LINGUISTIC\_TYPE\_REF – ссылка на лингвистический тип;

— TIER\_ID – идентификатор слоя;

— PARENT\_REF (необязательный атрибут) – ссылка на родительский слой.

*Лингвистический тип* (тег LINGUISTIC\_TYPE) представляет собой набор ограничений и свойств, присущих слою, который будет ссылаться на данный лингвистический тип. Лингвистический тип имеет следующие атрибуты:

— GRAPHIC\_REFERENCES – флаг, показывающий, может ли аннотация, принадлежащая слою с данным лингвистическим типом ссылаться на графический 2D-объект. Обычно не используется и имеет по умолчанию значение false;

— LINGUISTIC\_TYPE\_ID – идентификатор лингвистического типа;

— TIME\_ALIGNABLE – флаг, показывающий, можно ли привязать ко времени аннотации, принадлежащие слою с данным лингвистическим типом. Если флаг принимает значение «true», значит, аннотации привязываются непосредственно к интервалу времени размечаемого видеофайла (в ELAN доступна возможность работы также с аудиофайлами, однако в корпусе REC хранятся только видеофайлы). По умолчанию принимает значение true. В случае, если атрибут CONSTRAINT является ограничением символьного типа, принимает значение false;

— CONTROLLED\_VOCABULARY\_REF (необязательный атрибут) – ссылка на словарь допустимых значений аннотаций слоя с данным лингвистическим типом;

— CONSTRAINTS (необязательный атрибут) – тип ограничения, принимает одно из 4 значений: Symbolic Association, Symbolic Subdivision, Time Subdivision, Included In, которые были подробно рассмотрены выше.

*Словарь допустимых значений* (тег CONTROLLED\_VOCABULARY) имеет один атрибут – CV\_ID – идентификатор словаря.

Внутри тега CONTROLLED\_VOCABULARY определяется тег DESCRIPTION, содержащий описание словаря, а также некоторое количество тегов CV\_ENTRY\_ML – сущностей словаря, внутри которых, в свою очередь определяется по одному тегу CVE\_VALUE – значению сущности словаря. Текст тега CVE\_VALUE представляет собой значение, которое могут принимать аннотации слоя, ссылающегося на данный словарь через свой лингвистический тип.

#### Описание формата EAF

Файл разметки (EAF-файл) также представляет собой XML-файл. Все теги, которые были описаны для ETF-файла, и их порядок используются и в EAF-файлах. Кроме того, в EAF-файлах используется теги аннотаций и временных слотов.

Как говорилось выше, тег TIME\_ORDER – список временных слотов, поэтому в EAF-файлах внутри этого тега находится некоторое количество временных слотов.

*Временной слот* (тег TIME\_SLOT) представляет момент времени (от начала записи в файле), к которому может быть привязана аннотация в пределах файла. Имеет два атрибута:

— TIME\_SLOT\_ID – идентификатор слота;

— TIME\_VALUE – момент времени, связанный с медиа-файлом.

Далее идёт собственно разметка: внутри слоёв (тегов TIER) находятся аннотаций (теги ANNOTATION).

Аннотации могут быть двух типов: аннотации с привязкой ко времени и аннотации, ссылающиеся на другие аннотации. В зависимости от типа, соответствующий тег помещается внутри тега ANNOTATION.

*Аннотации с привязкой ко времени* (тег ALIGNABLE\_ANNOTATION) содержат атрибуты:

— ANNOTATION\_ID – идентификатор аннотации;

— TIME\_SLOT\_REF1 – ссылка на временной слот начала временного интервала, к которому привязана данная аннотация;

— TIME\_SLOT\_REF2 – ссылка на временной слот конца временного интервала, к которому привязана данная аннотация.

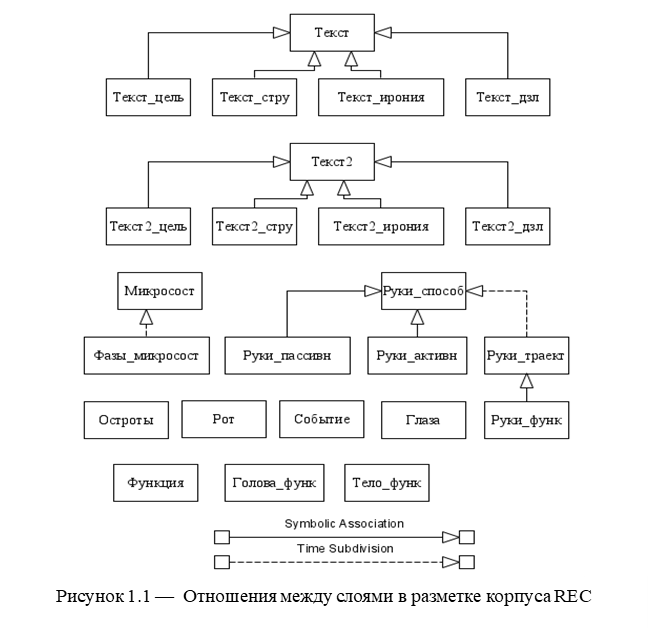
Аннотации, ссылающиеся на другие аннотации, (тег REF\_ANNOTATION) содержат атрибуты:

— ANNOTATION\_ID – идентификатор аннотации;

— ANNOTATION\_REF – ссылка на родительскую аннотацию.

Внутри тегов ALIGNABLE\_ANNOTATION и REF\_ANNOTATION помещается тег ANNOTATION\_VALUE, содержащий *значение аннотации*. Для аннотаций на слоях с лингвистическим типом, ссылающимся на словарь, значение аннотации должно соответствовать одному из значений сущности (CVE\_VALUE) этого словаря.

Отношения между слоями представлены на рисунке 1.3.



#### Шаблон разметки корпуса REC

В шаблоне разметки корпуса REC указано, что моменты времени измеряются в миллисекундах. Определено 24 слоя:

— «текст» – основной текст говорящего;

— «текст2» – текст основного оппонента в коммуникации;

— «текст\_стру», «текст2\_стру» – структура высказывания (текста и текста2 соответственно);

— «текст\_цель», «текст2\_цель» – цель высказывания (текста и текста2 соответственно);

— «текст\_дзл», «текст2\_дзл» – действия, затрагивающие лицо (ДЗЛ) адресата, аналог Face Threatening Act (FTA) в теории вежливости (текста и текста2 соответственно);

— «текст\_ирония», «текст2\_ирония» – наличие иронии, шуток, игры в тексте и тексте2 соответственно;

— «остроты»;

— «событие»;

— «глаза» – действия, выполняемые глазами;

— «рот» – движения рта;

— «руки\_способ» – движения рук;

— «руки\_активн» – активный орган;

— «руки\_пассивн» – пассивный орган;

— «руки\_траект» – траектория рук;

— «функция» – коммуникативная функция мимики и значимых движений корпусом;

— «голова\_функ» – коммуникативная функция движений головы;

— «тело\_функ» – коммуникативная функция движений тела;

— «руки\_функ» – коммуникативная функция движений рук;

— «микросост» – общая разметка изменений в коммуникативном

состоянии;

— «фазы\_микросост» – детальный состав последовательностей микросостояний.

Некоторые слои ссылаются через свой лингвистический тип на словари. Для них допустимые значения аннотаций приведены в приложении A.

Для дочерних слоёв разметки в REC используются только два типа ограничений: Symbolic Association и Time Subdivision.

Связи между слоями сведены в схему, представленную на рисунке 1.4. Для анализа данных из корпуса REC следует разработать специализированный метод. Рассмотрим подробно два ключевых подхода к поиску шаблонов: интеллектуальное извлечение данных и поиск подстрок.

### Интеллектуальное извлечение данных

Интеллектуальное извлечение данных, или Data Mining (далее Data Mining), является частью системы методов обнаружения знаний в базах данных, или KDD (knowledge discovery in databases), и используется для выявления шаблонов в статических (заранее полученных) данных. Другими словами, методы Data Mining используются для поиска скрытых, нетривиальных, практически полезных знаний в больших объёмах «сырых» данных. Особенностью Data Mining можно назвать то, что оно находит шаблоны, но не объясняет, почему они возникли. Кроме того, ищутся преобладающие значения, исключения не включаются в анализ и получаемые шаблоны, или *закономерности*.

Для решения конкретных задач Data Mining применяются конкретные *методы* Data Mining. В [3] выделяют три *стадии* Data Mining:

1. свободный поиск;
2. прогностическое моделирование;
3. анализ исключений.

#### Основные задачи Data Mining

Сведения об основных задачах Data Mining, приведённые в [3], [4], [5], сведены в таблицу 1.2.

В рамках данной работы, особый интерес представляют задачи поиска ассоциативных правил.

Таблица 1.2 — Основные задачи Data Mining

|  |  |  |
| --- | --- | --- |
| **Задача** | **Описание** | **Используемые модели и методы** |
| Классификация | Выполняет разбиение исходного множества объектов на заранее определённые классы. Одна из наиболее распространённых задач Data Mining. | Нейронные сети, деревья решений, метод k-ближайших соседей и др. |
| Кластеризация | Выполняет разбиение исходного множества объектов на группы, внутри которых объекты обладают наибольшей степенью схожести между собой и наименее схожи с элементами других групп. В отличии от классификации, классы заранее не известны. | Самоорганизующиеся карты Кохонена. |
| Поиск ассоциативных правил | Выполняет поиск закономерностей внутри транзакций – группы событий, связанных между собой каким-либо образом. Отличительной особенностью этой задачи является то, что анализируется не каждое отдельное событие, а группы событий. | Apriori |

Продолжение таблицы 1.2.

|  |  |  |
| --- | --- | --- |
| Поиск последовательностей (последовательных ассоциаций) | Является более общей задачей, чем поиск ассоциаций, так как учитывает не только разбиение на транзакции, но и их порядок. | Generalized Sequential Pattern (GSP, расширение алгоритма Apriori),  Sequential PAttern Discovery using Equivalence classes (SPADE) |
| Прогнозирование | На основе имеющихся данных оцениваются пропущенные или предсказываются новые данные. | Методы математической статистики, нейронные сети и др. |

Рассмотрим подробнее задачу поиска ассоциативных правил.

#### Поиск ассоциативных правил

Пусть – множество элементов, *N* – общее количество элементов, – транзакция, а – множество транзакций.

Пусть также *X* – произвольное подмножество *I*, т.е. . Тогда транзакция содержит *X,* если .

*Поддержкой* произвольного *набора* X называют число транзакций из , которые содержат этот набор, т.е.

|  |  |
| --- | --- |
| . | (1) |

*Ассоциативным правилом* называют импликацию вида

*Поддержкой* *правила*  называют число транзакций из , которые содержат

|  |  |
| --- | --- |
| . | (2) |

Стоит заметить, что .

*Достоверностью правила* называют число

|  |  |
| --- | --- |
|  | (3) |

Достоверность правила показывает вероятность того, что из *X* следует *Y.*

Целью поиска ассоциативных правил является нахождение зависимостей следующего вида: если в транзакции присутствует набор X, то вероятно в нём также присутствует набор Y.

Задачу нахождения ассоциаций можно разбить на два этапа.

1. Нахождение всех часто встречающихся наборов, то есть таких наборов, чья поддержка больше заданного минимального значения поддержки.
2. Генерация правил из наборов элементов, найденных на первом этапе, с достоверностью, превышающей заданное минимальное значение достоверности.

При этом, выделяют 3 типа получаемых правил:

— полезные – новые, нетривиальные правила, которые можно объяснить;

— необъяснимые – новые, нетривиальные правила, которые нельзя объяснить;

— бесполезные – очевидные правила.

Алгоритмы, решающие задачу поиска ассоциативных правил, – это алгоритм Apriori [6] и его улучшения:

— AprioriTid – не использует базу данных для подсчёта поддержки после первого прохода, использует кодирование кандидатов;

— DHP [7] – использует хеширование хеш-таблицу для ускорения подсчёта поддержки;

— Partition [8] – разбивает большие базы данных на разделы, которые полностью помещаются в оперативной памяти, сначала рассчитывается поддержка в разделах, а потом относительно всей БД;

— ConApriori [9] – использует представление транзакций в виде бинарных векторов, таким образом, операция проверки включения произвольного множества в транзакцию сводится к операции побитового «И», а также сокращает количество операций при построении ассоциативных правил, за счёт итеративного построения правил, основывающегося на подсчёте достоверности, подобно построению кандидатов, основывающегося на подсчёте поддержки.

Так как корпус REC относительно небольшой, то не имеет смысла использовать алгоритмы, разбивающие базу данных на части. Кроме того, исходя из анализа разметки корпуса, можно предположить, что наибольший интерес будут представлять небольшие наборы элементов (до 24, если искать закономерности в одновременных событиях на всех слоях, и до 7, если рассматривать несколько близких по времени событий на одном слое). Таким образом, алгоритм AprioriTid будет уступать алгоритму Apriori, так как он даёт существенное увеличение скорости только на достаточно больших наборах данных [10], [11]. Использование бинарных векторов транзакций представляет достаточно удобный и универсальный способ представления данных и может быть применено в данной работе.

### Выводы

В данном разделе приведён разбор формата разметки корпуса REC, приведено описание используемых слоёв и их взаимодействий, рассмотрены основные задачи интеллектуального извлечения данных.

В результате проведённого анализа в рамках данной работы решено решать задачу поиска ассоциативных правил в корпусе REC и использовать алгоритма Apriori. Как следствие, возникает необходимость разработки метода преобразования данных корпуса в виде, пригодном для поиска ассоциативных правил.

## Конструкторский раздел

### Метод выделения шаблонов в REC

Разрабатываемый метод можно представить с помощью IDF0-диаграмм, представленных на рисунках 2.1-2.2.

Задачи, представленные блоками А3 и А4 (поиск частых наборов и ассоциативных правил соответственно), решаются посредством алгоритма Apriori. Значения минимальной поддержки и минимальной достоверности, максимально достоверности, а также минимального и максимального количества элементов зависят от целей анализа.

Задача разбора файла (парсинга) будет основываться на разборе формата файлов разметки, приведённом в аналитическом разделе.

Задача выделения транзакций в случае с корпусом REC раньше не рассматривалась, по крайней мере, результаты такого исследования не удалось найти в свободном доступе. Поэтому основное внимание в данном разделе будет уделено именно разработке метода выделения транзакций из данных корпуса REC.

### Разбор EAF-файлов

Схема разбора EAF-файлов, специфичных для корпуса REC, приведена на рисунке 2.3. В результате выполнения этого алгоритма:

— создаётся ассоциативный контейнер временных слотов, где ключом будет выступать идентификатор стола, а значением – момент времени;

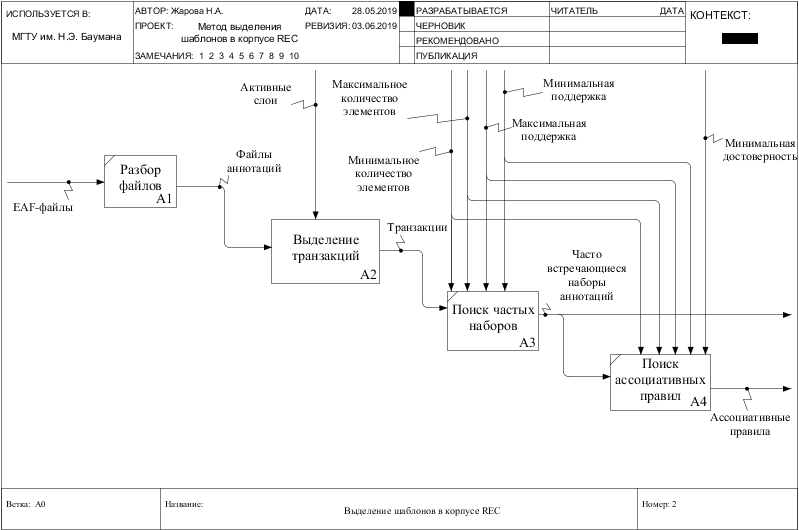


Рисунок 2.3 — IDEF0-диаграмма метода выделения шаблонов в REC.

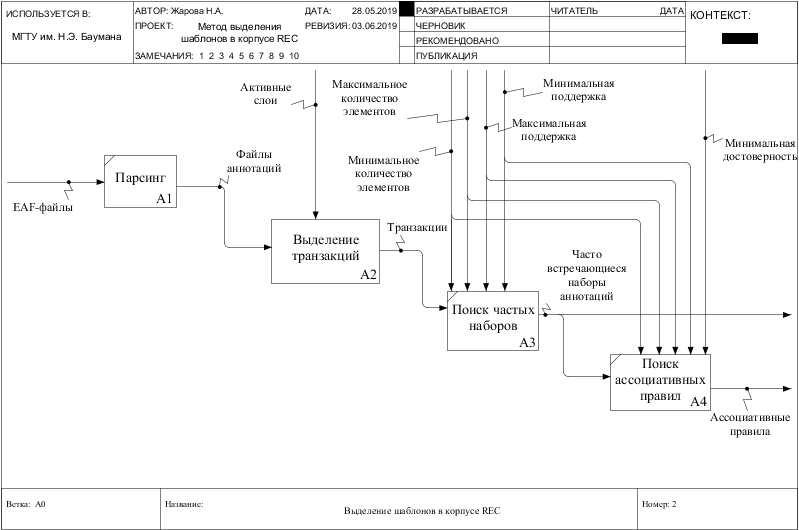


Рисунок 2.4 — IDEF0-диаграмма метода выделения шаблонов в REC. Детализация блока А0.

— создаётся массив слоёв, которые представлены массивом аннотаций: каждая аннотация содержит имя файла, из которой она была получена, идентификатор родительской аннотация (по умолчанию, нулевой) и идентификаторы двух временных слотов (возможно, нулевые, если указан идентификатор родительской аннотации);

— создаётся ассоциативный контейнер аннотаций, где ключом выступает идентификатор аннотации, а значением – указатель на соответствующую аннотацию.

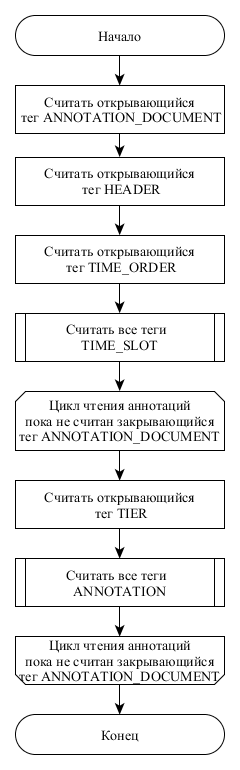


Рисунок 2.5 — Схема разбора EAF-файлов

Для получения готовых к выделению транзакций данных необходимо разрешить ссылки между перечисленными выше структурами данных. То есть необходимо для каждой аннотации получить по два значения временных слотов по их идентификаторам. В тех аннотациях, где используется ссылка на родителя, сначала необходимо приравнять идентификаторы временных слотов родительской и дочерней аннотации, а затем получить значения временных слотов.

### Метод преобразования данных

В транзакцию должны объединяться несколько элементов (событий, аннотаций), составляющих единую смысловую группу.

В одну транзакцию могут входить только аннотации из одного файла, таким образом. При этом учитываться те слои, которые будут помечены пользователем как активные. По умолчанию все слои помечены как активные.

После этого подсчитывается и запоминаются в порядке нахождения все уникальные значения аннотаций. При этом допустимые значения аннотаций будут учитываться как уникальные для каждого слоя.

Теперь каждая транзакция будет представлена как массив битов длиной, соответствующей количеству найденных уникальных значений. Если конкретный бит установлен в 1, то соответствующая аннотация присутствует в транзакции, иначе отсутствует.

Выделение транзакций происходит пофайлово. Всё время, для которого представлена разметка в текущем файле (от минимального значения временного слота (0) до максимального), разбивается на интервалы, в зависимости от используемого подхода. Затем для каждого интервала определяется, какие аннотации попали на данный интервал. Аннотация считается попавшей в интервал, даже если попадает в рассматриваемый интервал только частично. При этом пустые транзакции (в которые не вошла ни одна аннотация) удаляются.

IDEF0-диаграмма выделения транзакций представлена на рисунке 2.4.

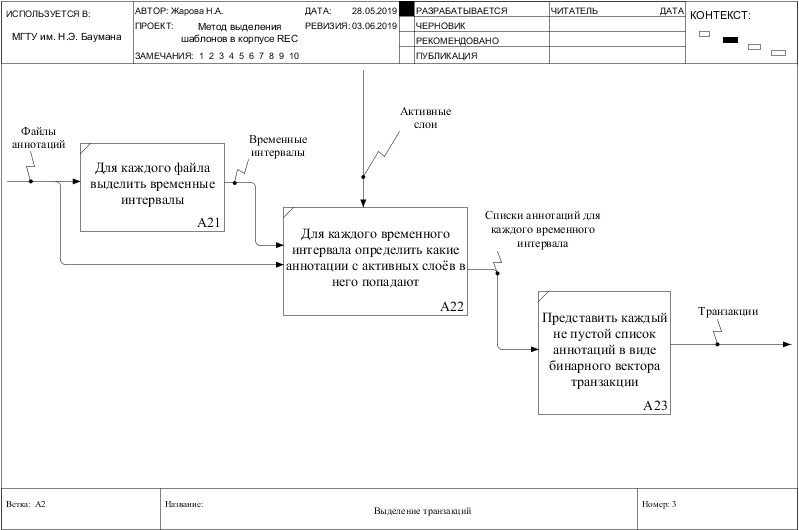


Рисунок 2.6 — IDEF0-диаграмма 1.1 метода преобразования данных

В данной работе предлагается рассмотреть три способа разбиения на временные интервалы.

*Разбиение на одинаковые временные интервалы*. При данном подходе временной интервал разметки в текущем файле разбивается на одинаковые интервалы с длительностью (шириной «окна») и множителем смещения. Множитель смещения представляет собой числа от 0,01 до 1. При этом начало следующего интервала времени смещается относительно начала текущего интервала на величину равную произведению длительности на множитель смещения, округлённому в большую сторону.

*Разбиение на временные интервалы разной длины в соответствии с аннотациями на главном слое*. Один из слоёв помечается пользователем как главный (ведущий), после чего происходит поиск аннотаций, принадлежащих данному слою. Так как аннотации одного слоя могут быть привязаны к временным интервалам разной длины, то временной интервал разметки в текущем файле также разбивается на интервалы разной длины в соответствии с временными интервалами найденных аннотаций на ведущем слое. При этом интервалы времени, которые не содержат аннотаций на ведущем слое, не учитываются в анализе.

*Разбиение на временные интервалы разной длины в соответствии с временными слотами*. При данном подходе временной интервал разметки в текущем файле разбивается на интервалы времени, которые определяются двумя соседними по значению временными слотами.

### Выводы

В данном разделе приведено общее описание разрабатываемого метода, кроме того подробно разобраны задачи парсинга EAF-файлов и выделения транзакций. Разработано три подхода к выделению транзакций: разбиение на одинаковые временные интервалы, разбиение на временные интервалы разной длины в соответствии с аннотациями на главном слоем, разбиение на временные интервалы разной длины в соответствии с временными слотами.

## Технологический раздел

### Выбор технологии программирования, языка программирования и среды разработки

Программа будет реализована с использованием объектно-ориентированного подхода, так как классы позволяют разрабатывать программы на основе информационной модели, абстрагируясь от реализации, инкапсуляция защищает данные от случайных изменений, наследование позволяет довольно легко добавлять функциональность, не изменяя ранее разработанные методы, а полиморфизм предоставляет возможность использовать объекты разных классов через один интерфейс. Всё это удобно использовать в текущем проекте.

Для реализации программы выбран язык С++, поскольку это универсальный язык, обладающий высокой вычислительной производительностью. Кроме того, он поддерживает парадигму ООП.

В качестве среды разработки выбрана среда Qt Creator, так как она предоставляет удобные инструменты для разработки программ на С++, особенно для разработки пользовательского интерфейса (Qt Designer). Кроме того, Qt Creator поддерживает активно развивающийся фреймворк Qt, который предоставляет различные классы, упрощающие разработку сложных систем на C++.

### Основные классы

Диаграмма основных классов представлена на рисунке 3.1. На диаграмме не представлены классы, связанные с визуализацией.

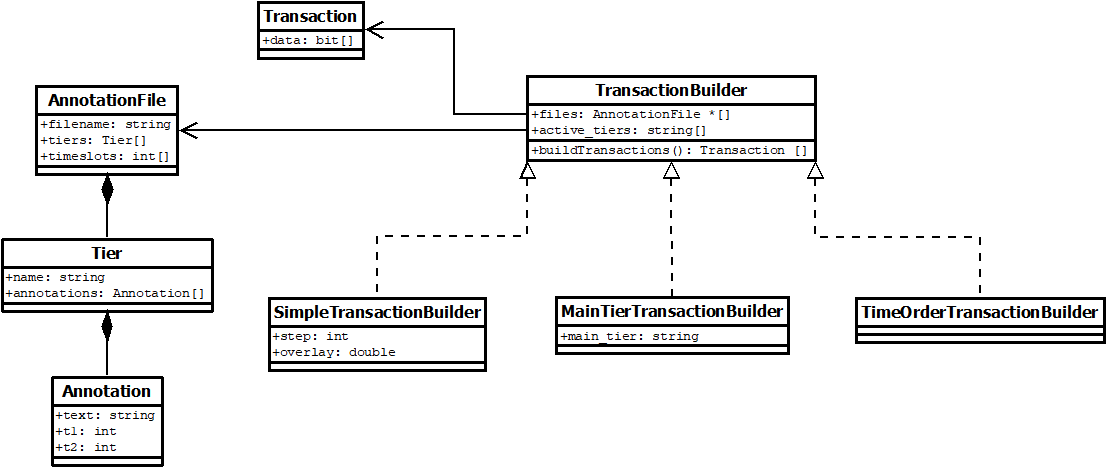


Рисунок 3.7 — Диаграмма основных классов.

*Класс Annotation.* Представляет тег Annotation. Хранит информацию о тексте аннотации и временных метках, к которым она привязана.

*Класс Tier.* Представляет тег Tier. Хранит информацию о названии слоя и содержит аннотация этого слоя.

*Класс AnnotationFile.* Хранит имя файла разметки и содержит данные разметки из файла.

*Класс Transaction.* Объект класса представляет данные одной транзакции в виде бинарного вектора. В статическом массиве класса хранятся все встречающиеся уникальные тексты аннотаций. Таким образом i-ый бит в конкретном объекте класса является флагом наличия в отображаемой транзакции i-ого элемента массива уникальных текстов аннотаций. Для работы с массивом битов используется класс QBitArray фреймворка Qt.

*Класс TransactionBuilder.* Является базовым чисто виртуальным классом для всех классов, реализующих алгоритмы выделения транзакций. Класс-интерфейс. Поле *files* – массив выбранных файлов аннотаций, в которых будет произведён поиск транзакций. Поле *active\_tiers –* массив с названиями активных слоёв. Основной метод – buildTransactions() – возвращает массив выделенных транзакций.

*Класс SimpleTransactionBuilder.* Является реализацией метода выделения транзакций на одинаковых интервалах времени. Поле *step* определяет длительность временных интервалов. Поле *overlay* – множитель, определяющий, смещение интервалов и их пересечение, может принимать значения от 0,01 до 1. Интервалы будут смещены друг относительно друга на значение, полученное путём округления overlay\*step в большую сторону. При максимальном значении *overlay* – 1 – интервалы будут следовать друг за другом, без промежутков и не пересекаясь.

*Класс MainTierTransactionBuilder.* Является реализацией метода выделения транзакций на интервалах разной длины по главному слою. Поле *main\_tier* содержит название главного слоя.

*Класс TimeOrderTransactionBuilder.* Является реализацией метода выделения транзакций на интервалах разной длины по временным слотам.

### Руководство пользователя

На рисунке 3.2 представлена use-case диаграмма разработанного ПО.

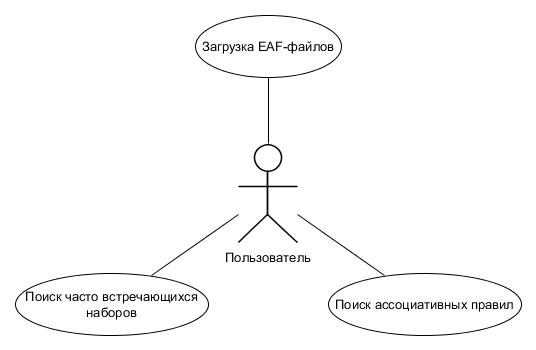


Рисунок 3.8 — Use-case диаграмма.

На рисунках 3.3-3.6 представлен пользовательский интерфейс разработанной программы RECAnalyzer.

В качестве входных данных программа принимает EAF-файлы корпуса REC. Выходные данные представлены несколькими текстовыми файлами.

*Файлы in\*.txt.* Вместо \* в названии файла указан номер анализа, в результате которого были выделены транзакции, представленные в файле. Транзакции разделяются символом переноса строки. Элементы транзакций строятся по следующему принципу: <называние слоя>-<текст\_аннотации>, при этом все пробелы в тексте аннотации заменяются на символ нижнего подчёркивания («\_»). Элементы разделяются пробелами.

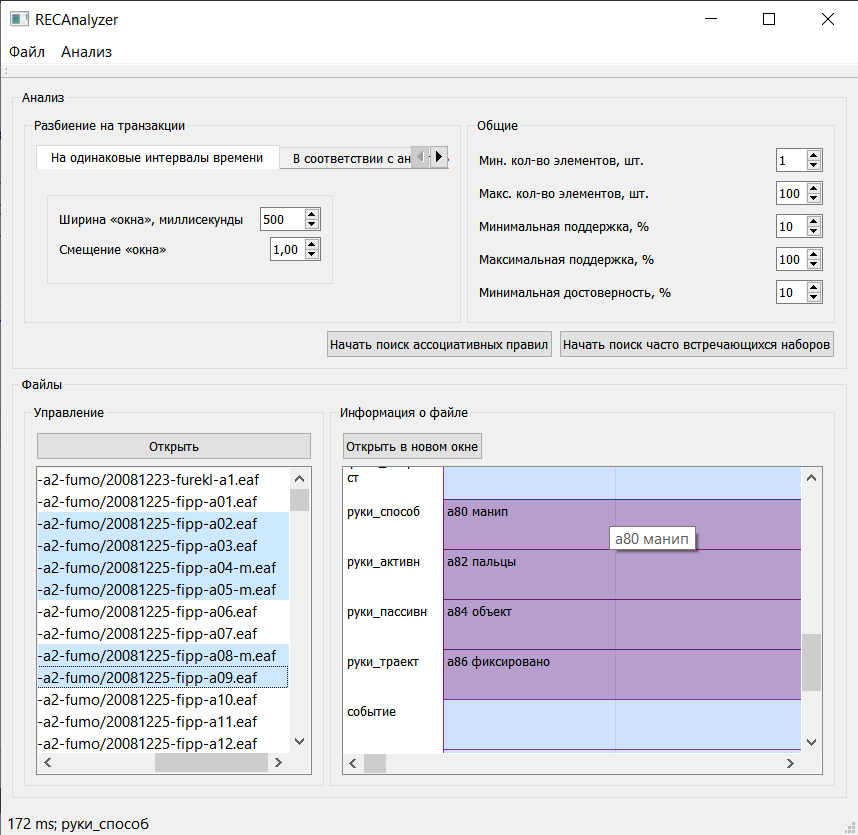


Рисунок 3.9 — Главное окно RECAnalyzer.

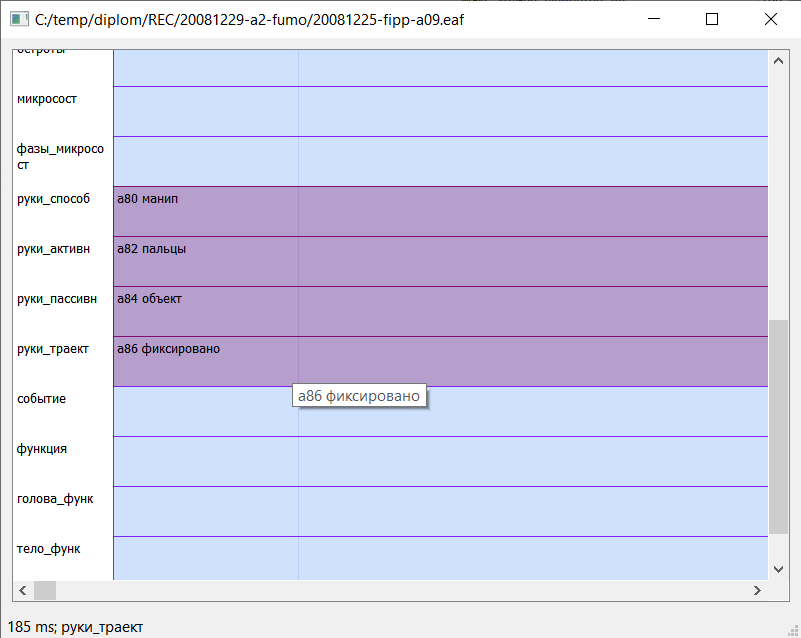


Рисунок 3.10 — Визуализация EAF-файла.

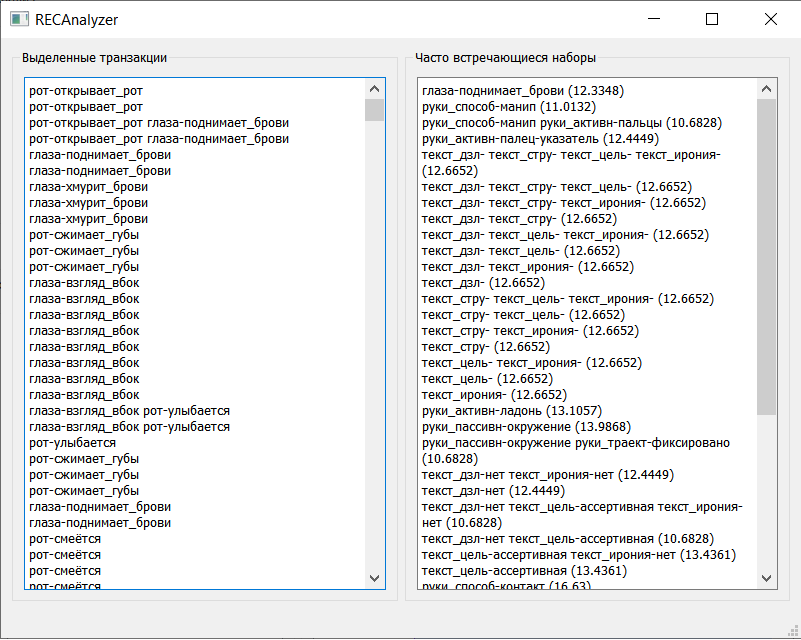


Рисунок 3.11 — Результаты поиска часто встречающихся наборов.

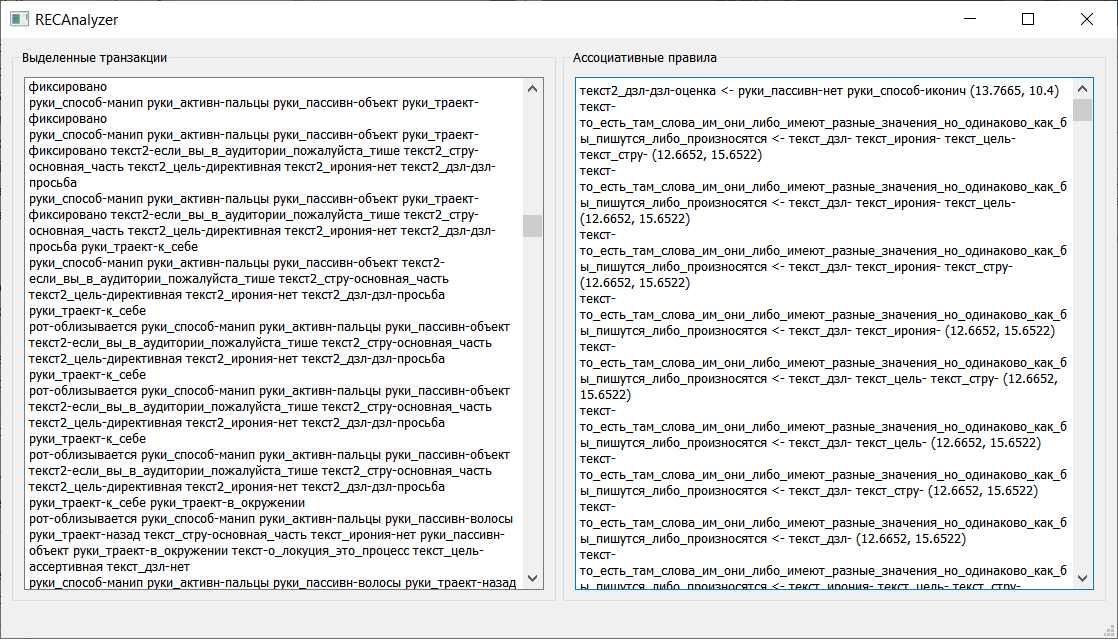


Рисунок 3.12 — Результаты поиска ассоциативных правил.

*Файлы out\*.txt.* Вместо \* в названии файла указан номер анализа, в результате которого были выделены ассоциативные правила или часто встречающиеся шаблоны, представленные в файле. Правила или наборы разделяются символом переноса строки. Элементы строятся по тому же принципу, что и элементы транзакций. Элементы наборов разделяются пробелами, в конце строки в скобках указана поддержка набора в процентах (рисунок 3.5). Правила представлены следующим образом: в начале строки указан один элемент – следствие правила, – затем через пробелы идут символ «<-» и тело правила (некоторое количество элементов, разделённых пробелом), в конце строки в скобках указывается поддержка тела правила и достоверность правила через запятую (рисунок 3.6).

Для проведения анализа необходимо открыть интересующие EAF-файлы корпуса REC. Это можно сделать с помощью сочетания клавиш CTRL+O, через меню Файл -> Открыть или нажатием на кнопку «Открыть» в разделе Файлы-Управление. В появившемся диалоговом окне (рисунок 3.7) можно выбрать сразу несколько файлов. Список открытых файлов отображается под кнопкой «Открыть».

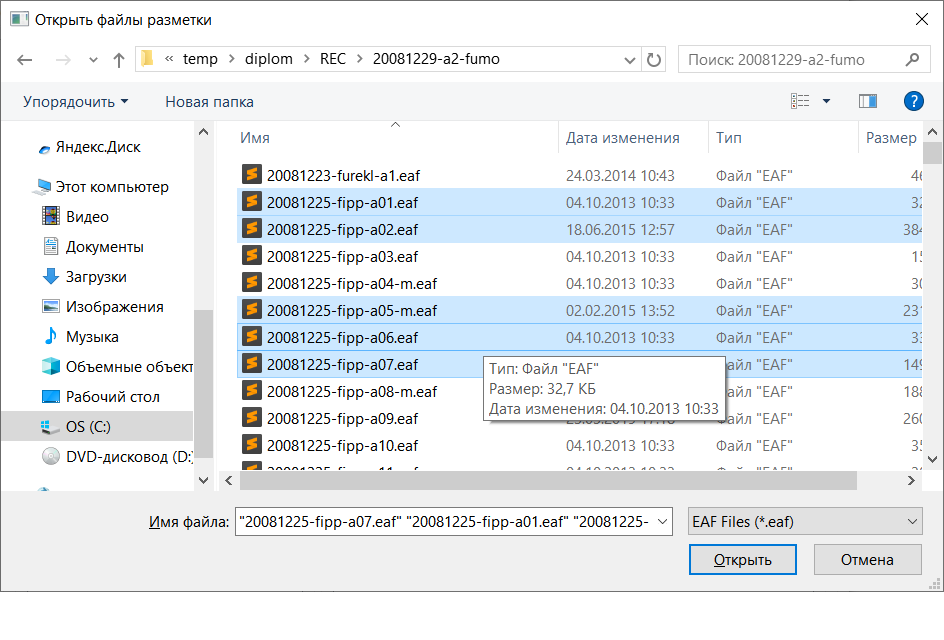


Рисунок 3.13 — Диалоговое окно «Открыть файлы разметки»

После того как файл открыт, можно посмотреть визуализацию разметки, приведённой в данном файле. Для этого необходимо выбрать файл из списка открытых файлов (щелкнуть левой кнопкой мыши по нужной строке) и визуализация появится в области визуализации в разделе Файлы-Информация о файле. Визуализацию файла можно открыть в новом окне, нажав кнопку «Открыть в новом окне» в том же разделе. Слои разметки представлены горизонтальными полосами голубого цвета (условная временная ось направлена вправо), аннотации – прямоугольниками сиреневого цвета, размещёнными на соответствующих слоях. Кроме того, при наведении курсора мыши в область визуализации в строке состояния появится название слоя и момент времени, на который указывает курсор. Для удобства сравнения времени начала и конца аннотаций также отображается «Курсор» – тонкая вертикальная синяя линия, которая проходит через один и тот же момент времени на всех слоях разметки. Если задержать курсор мыши на какой-то аннотации, появляется всплывающая подсказка с текстом аннотации.

По умолчанию, все слои активного файла (отображающегося в зоне визуализации главного окна) помечаются как активные, для того чтобы пометить слой как неактивный необходимо дважды щелкнуть по нему левой кнопкой мыши, после чего он перекрасится в более светлый цвет.

Далее необходимо выбрать метод разбиения на транзакции, для этого необходимо выбрать нужную вкладку в разделе Анализ-Разбиение на транзакции.

Вкладка «На одинаковые интервалы времени» содержит два настраиваемых поля: «Ширина окна, миллисекунды» и «Смещение окна», которые соответствуют по смыслу полям step и overlay класса SimpleTransactionBuilder, описанным выше. Выбор данной вкладки обеспечивает разбиение на одинаковые временные интервалы.

Вкладка «В соответствии с аннотациями на главном слое» содержит одно настраиваемое поле «Главный слой». В выпадающем списке необходимо выбрать слой, который будет назначен как главный. Если главный слой не входит в список активных слоёв, транзакции всё равно могут успешно быть выделены, но аннотации главного слоя не будут входить в транзакции. Если в файле главный слой отсутствует, в нём не будет найдено ни одной транзакции. Выбор данной вкладки обеспечивает разбиение на временные интервалы разной длины в соответствии с аннотациями на главном слоем.

Вкладка «В соответствии с TIME\_ORDER» не содержит настраиваемых полей, выбор данной вкладки обеспечивает разбиение на временные интервалы разной длины в соответствии с временными слотами.

В разделе Анализ-Общие приведены общие настройки анализа.

Поле «Мин. кол-во элементов, шт.» отвечает за минимальное количество элементов в результирующих часто встречающихся наборов или в телах результирующих правил.

Поле «Макс. кол-во элементов, шт.» аналогично предыдущему, но отвечает за максимальное количество элементов.

Поле «Минимальная поддержка, %» для поиска часто встречающихся наборов отвечает за минимальную поддержку результирующих наборов, а для поиска ассоциативных правил – за минимальную поддержку тел результирующих правил.

Поле «Максимальная поддержка, %» аналогично предыдущему, но отвечает за максимальную поддержку количество элементов.

Поле «Минимальная достоверность, %» имеет значение только при поиске правил и отвечает за минимальную достоверность результирующих правил.

Для начала анализа следует нажать на кнопку «Начать поиск часто встречающихся наборов» или «Начать поиск ассоциативных правил» после чего появится новое окно (рисунок 3.5 или 3.6 соответственно). При этом в левой части окна будут отображены найденные транзакции, а в правой наборы или правила. Формат отображения совпадает с форматов выходных файлов, описанных выше.

Стоит заметить, что анализ проводится только по активным слоям выбранных файлов (выделенных синим цветом в списке открытых файлов) и, возможно, по главному слою при выборе соответствующего режима.

Для сборки разработанного ПО используются Qt Creator 4.9.0, основанный на Qt 5.12.2 (MSVC 2017, 32 бита) и компилятор MSVC 2017, 32 бита.

### Выводы

В рамках данного раздела осуществлён выбор языка С++ и среды разработки Qt Creator. Приведено описание основных классов, используемых при анализе разметки. Приведено руководство пользователя разработанного ПО RECAnalyzer, в котором описаны форматы входных и выходных файлов, пользовательский интерфейс, порядок действий при работе с программой и необходимые для сборки инструменты.

## Исследовательский раздел

### Разбиение на одинаковые интервалы времени

Для того чтобы получить общее представление о возможностях разработанного метода на данных корпуса был проведён анализ по всем имеющимся 157 файлам, список которых приведён в приложении Б, и всем слоям со следующими значениями параметров: минимальное количество элементов – 1, максимальное – 100, минимальная поддержка – 10%, максимальная – 100%, минимальная достоверность – 10%. В рамках исследовательского раздела будем считать эти значения параметров значениями по умолчанию.

Вычисленная средняя длительность аннотации: 1743.57 миллисекунд.

Результаты такого анализа представлены в табл. 4.1.

Таблица 4.1 — Результаты анализа с использованием разбиения на одинаковые интервалы времени.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ширина «окна», миллисекунды | Множитель смещения | Количество наборов, шт. | Максимальный размер набора, шт. | Максимальное значение  поддержки, % |
| 100 | 0,25 | 23 | 4 | 25,18 |
| 100 | 1 | 23 | 4 | 25, 48 |
| 1000 | 0,25 | 40 | 4 | 30,82 |
| 1000 | 1 | 40 | 4 | 30,72 |
| 2000 | 0,25 | 88 | 6 | 36,00 |
| 2000 | 1 | 88 | 6 | 35,88 |
| 3000 | 0,25 | 335 | 8 | 40,48 |
| 3000 | 1 | 335 | 8 | 40, 47 |
| 4000 | 0,25 | 337 | 8 | 44, 28 |
| 4000 | 1 | 337 | 8 | 44,36 |

При дальнейшем увеличении ширины «окна» результаты перестают значительно отличаться от результатов анализа при шаге 4000 миллисекунд. Поэтому для анализа всех файлов по всем слоям следует выбирать ширину окна в пределах от 2000 до 4000. Изменение значения множителя смещения, позволяет производить более точную настройку, при которой выделяемые наборы почти не меняются, но меняется их значение поддержки.

Стоит отметить, что чаще всего выделенные наборы содержат однотипные аннотации со слоёв текста (например, текст2\_ирония\_, текст2\_дзл\_, текст\_стру\_ и др.). Для поиска более интересных наборов следует исключить слои текста, если не все, то частично (например, оставить слои текст\_цель и текст2\_цель).

Кроме того, при таком разбиении на временные интервалы анализ всех файлов приводит к выделению большого количества наборов с малым значением поддержки (около 10%). Для улучшения результатов, следует объединять анализируемые файлы в связанные группы, например, анализировать только файлы разметки видеозаписей с экзаменов.

### Разбиение в соответствии с аннотациями на главном слое

Для анализа данного метода на данных корпуса также был проведён анализ по всем имеющимся 157 файлам и всем слоям со значениями параметров по умолчанию.

Проводить исследования для каждого из существующих слоёв не целесообразно, так как дочерние слои, связанные с родительскими посредством символьной ассоциации, будут полностью учтены при разбиении по родительскому слою. Таким образом, можно провести анализ только по следующим слоям: текст, текст2, микросост, руки\_способ, остроты, событие, рот, глаза, функция, функция\_тело, функция\_голова).

Результаты такого анализа представлены в табл. 4.2.

Таблица 4.2 — Результаты анализа с использованием разбиения на интервалы разной длины в соответствии с аннотациями на главном слое.

|  |  |  |  |
| --- | --- | --- | --- |
| Главный слой | Количество наборов, шт. | Максимальный размер набора, шт. | Максимальное значение  поддержки, % |
| текст | 1165 | 9 | 72,88 |
| текст2 | 16293 | 11 | 74,20 |
| микросост | 572732/11039 | 17/11 | 29,1925/73,60 |
| руки\_способ | 989 | 9 | 56,88 |
| рот | 319 | 8 | 40,41 |
| глаза | 371 | 8 | 51,14 |
| функция | 8545 | 11 | 54,50 |
| тело\_функ | 79812 | 13 | 57,55 |
| голова\_функ | 9209 | 11 | 46,26 |
| остроты | 131421 | 16 | 83,33 |
| событие | 36281 | 12 | 84,29 |

При разбиении по аннотациям на слоях остроты и события при заданных параметрах выделяется слишком большое количество наборов (при записи в файл занимает более 27 Гб), поэтому для них данные приводятся при минимальной поддержке 30%.

При разбиении по аннотациям на слое микросост приведено по два значения (через «/»), так как при минимальной поддержке 10% выделяется большое количество аннотаций и их сложно анализировать, а при поддержке 20% анализ данных производится значительно быстрее (сортировка, поиск наборов содержащих конкретные аннотации). Тем не менее на рисунке 4.1 приведены 22 набора с наиболее высоким значением поддержки при минимальной поддержке 10%.

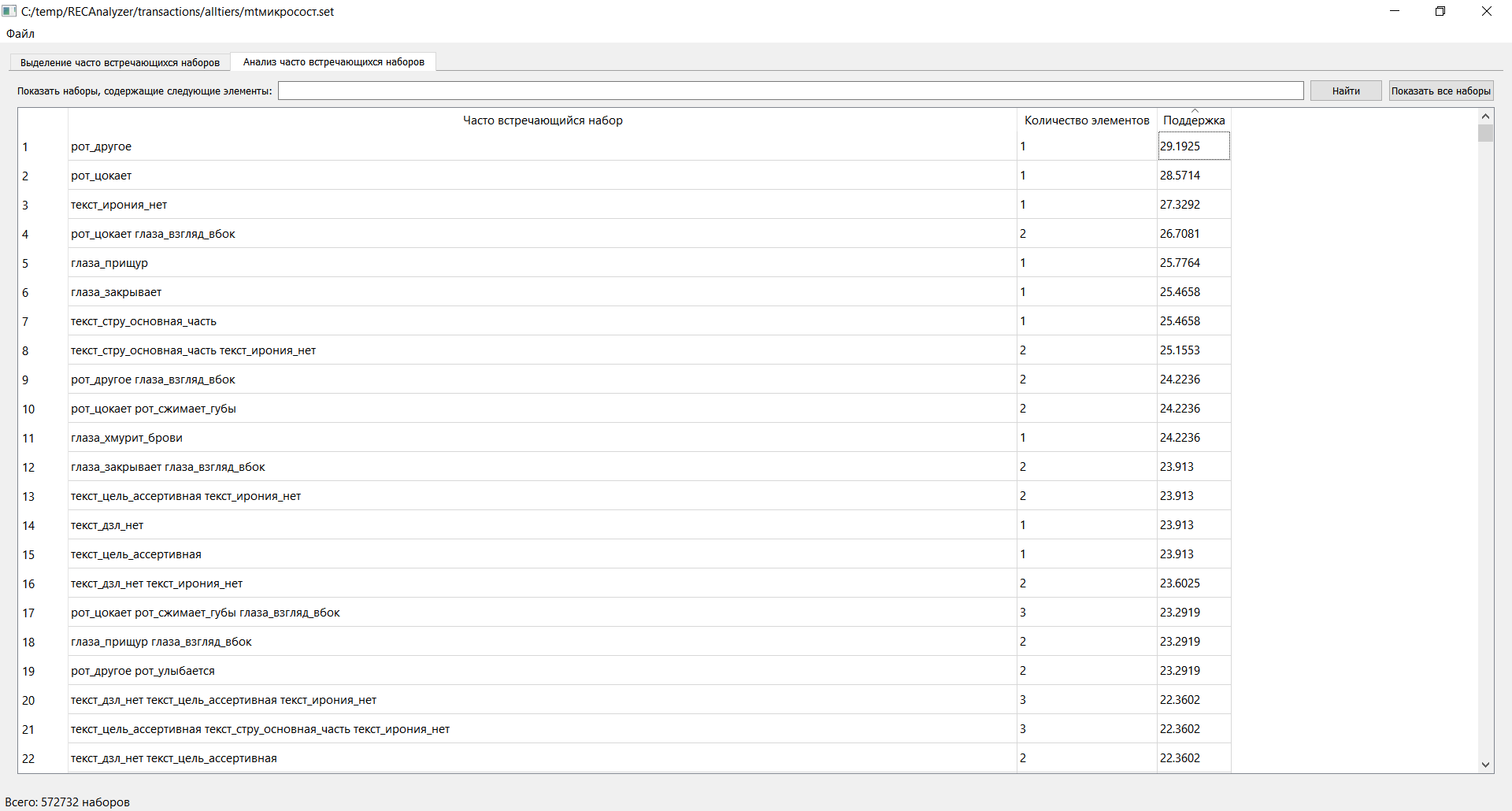


Рисунок 4.14 — Результаты анализа по слою микросост с минимальной поддержкой 10%.

При изучении полученных данных было выявлено, что аннотации текстовых слоёв при таком разбиении также создают эффект шума, наиболее интересными являются наборы с поддержкой порядка 50%. Примеры таких наборов для анализа по слою событие приведены на рисунке 4.2.

### Разбиение в соответствии с моментами времени

Как и для других разбиений, был проведён анализ со значениями параметров по умолчанию.

Получилось выделить 40 наборов, с максимальным количеством элементов 2 и максимальной поддержкой 15,26.

Данные результаты объясняются тем, что при этом типе разбиения в транзакциях будут представлены все, существующие в корпусе уникальные наборы элементов. При этом поддержка каждого такого набора будет низкой. Зато такой тип разбора позволяет определить каждое совместное появление аннотаций.

### Пример использования разработанного метода для более детально анализа

Как уже говорилось выше, слои текста могут рассматриваться как шум. Поэтому был проведён анализ при следующем списке активных слоёв: текст\_цель, глаза, рот, руки\_способ, руки\_активн, руки\_пассивн, руки\_траект, событие, функция, голова\_функ, тело\_функ, руки\_функ.

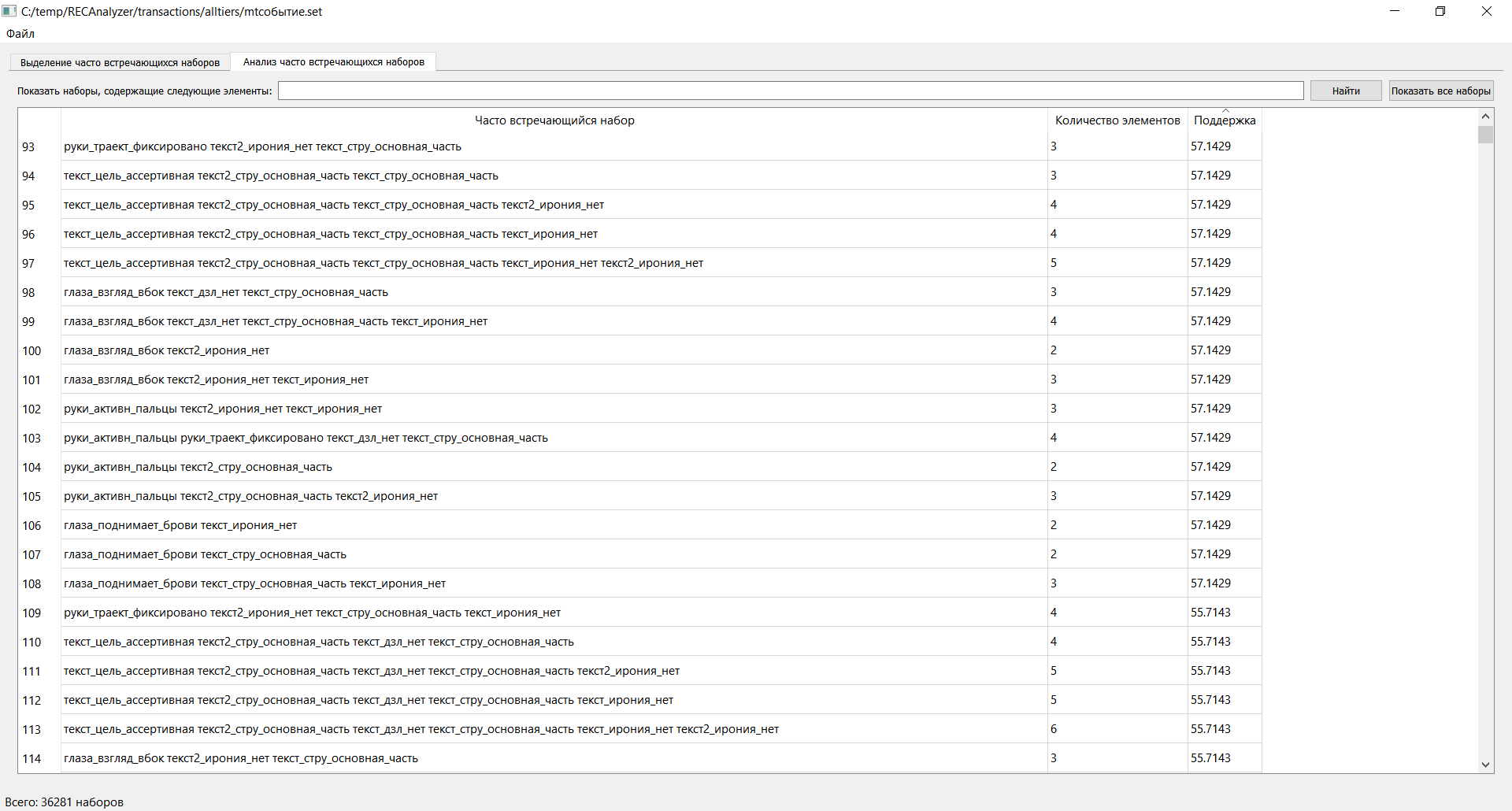


Рисунок 4.2 — Примеры выделенных наборов при разбиении по слою событие с поддержкой около 57%.

При разбиении на одинаковые интервалы времени с шириной окна 2000 и множителем смещения 0,25 со значениями параметров по умолчанию удалось выявить 19 наборов, которые представлены в отсортированном по значению поддержки порядке на рисунке 4.3. Как видно из рисунка, наибольшее значение поддержки – 37% для набора состоящего из 1 элемента – «глаза\_взгляд\_в\_бок». Наибольшее количество элементов в наборе – 3.

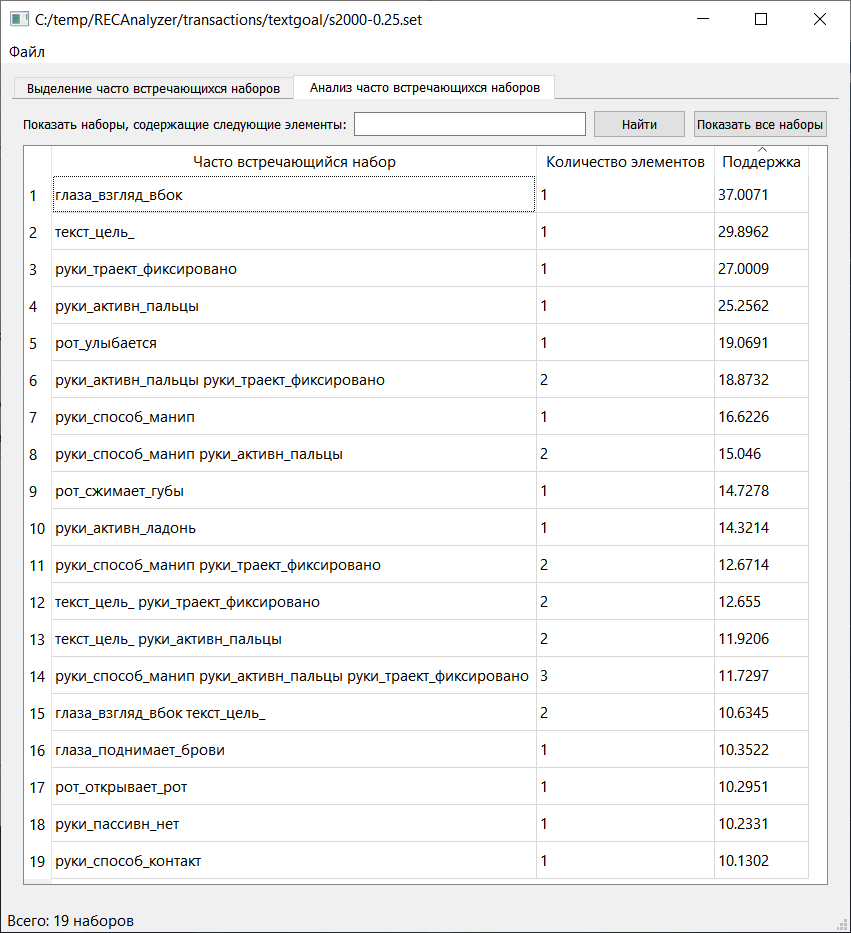


Рисунок 4.3 — Результат разбиения на одинаковые интервалы времени с шириной окна 2000 и множителем смещения 0,25

Для слоя событие при анализе выделяется большое количество наборов, в которых присутствуют аннотации со слоёв текста. При отсутствиии этих слоёв (кроме, текст\_цель, текст2\_цель) удаётся выявить более разнообразные наборы, примеры наибольшим количеством элементов представлены на рисунке 4.4.

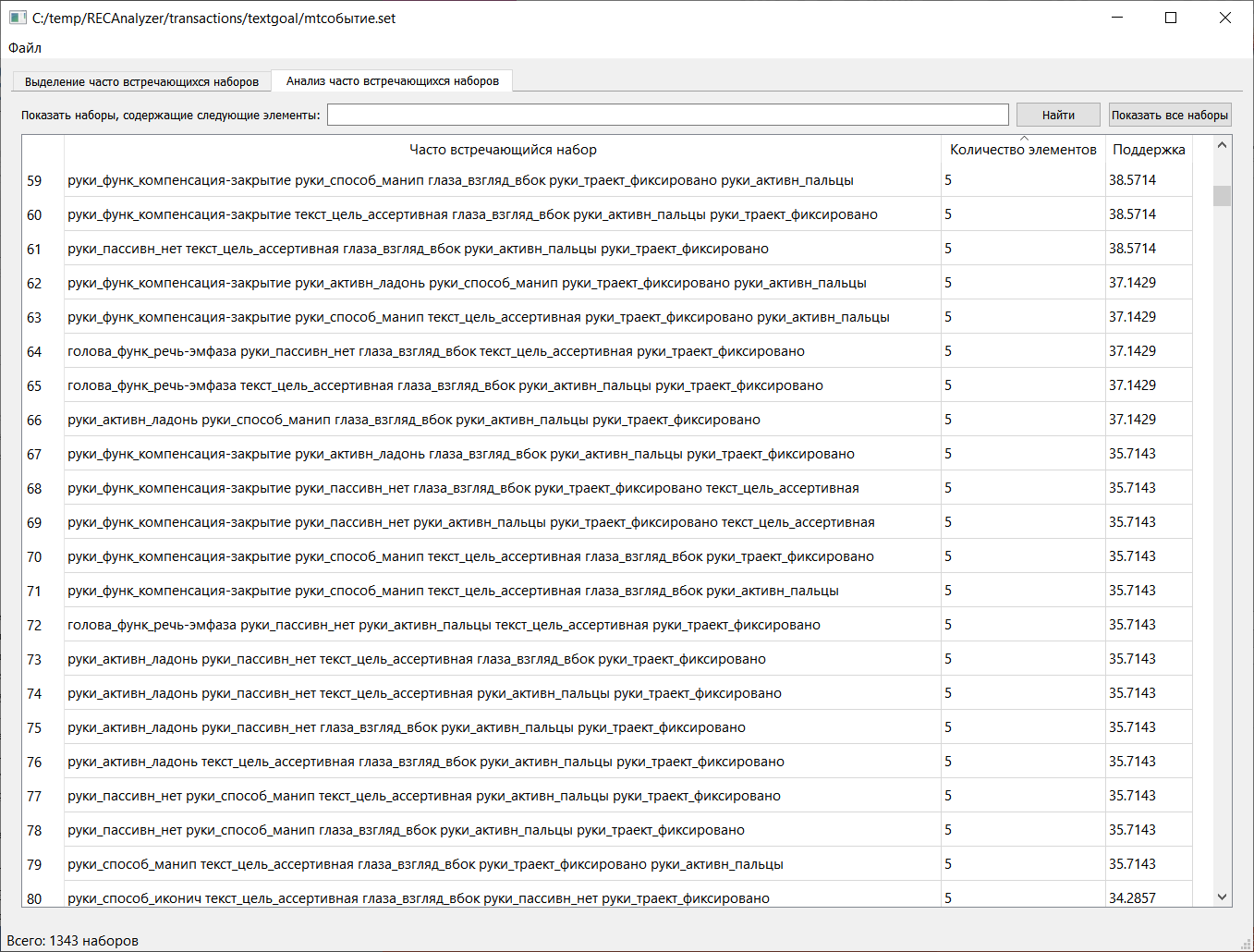


Рисунок 4.4 — Наборы с наибольшим количеством элементов

Рассмотрим более подробно набор: руки\_функ\_компенсация-закрытие, руки\_способ\_манип, текст\_цель\_ассертивная, руки\_траект\_фиксировано, руки\_активн\_пальцы. Можно заметить, что 4 из 5 аннотаций принадлежат слою руки и 1 – слою текст\_цель. Возникает вопрос, насколько применимо правило, у которого в заголовке будет аннотация текст\_цель\_ассертивная, а тело будет состоять из 4 оставшихся аннотаций.

В результате анализа было выявлено 140540 правил. 2617 правил содержат в своём теле тело рассматриваемого правила и только у 15 правил из них, заголовок совпадает с заголовком искомого правила (рисунок 4.5). У искомого правила достоверность 86%, то есть набор {руки\_функ\_компенсация-закрытие, руки\_способ\_манип, руки\_траект\_фиксировано, руки\_активн\_пальцы} не гарантирует наличие аннотации текст\_цель\_ассертивная. Однако если к набору добавить функция\_ожидание-обратной-связи, или руки\_пассивн\_нет, или набор {руки\_активн\_ладонь, руки\_пассивн\_нет}, или набор {руки\_пассивн\_нет, глаза\_взгляд\_вбок}, то при наличии такого набора транзакции в нём наверняка будет аннотация текст\_цель\_ассертивная.

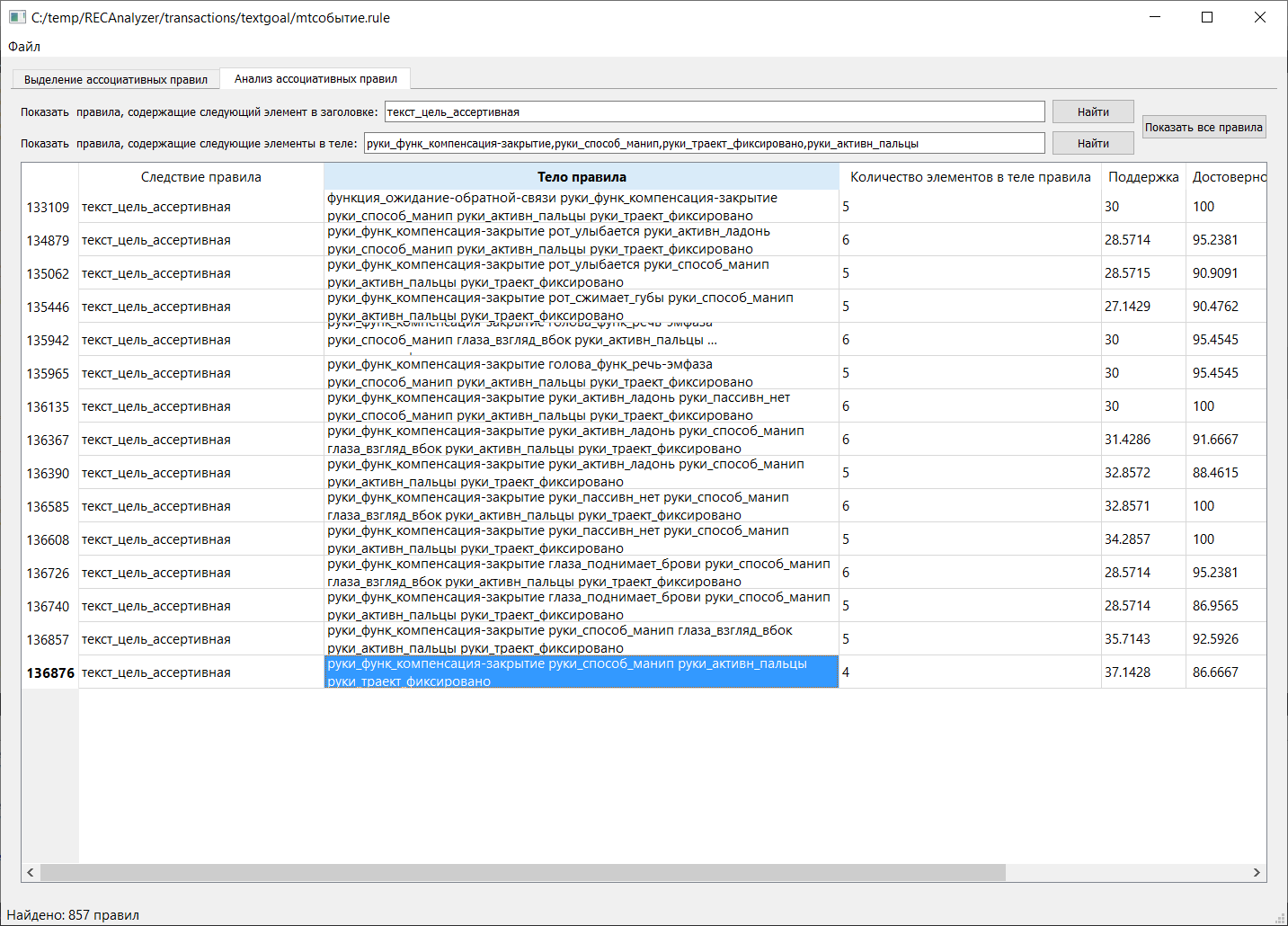


Рисунок 4.5

### Выводы

В рамках данного раздела было проведено первичное исследование корпуса REC с помощью разработанного метода. Для более детального исследования необходимы дополнительные данные о содержании файлов разметки (например, в каких файлах находится разметка видеозаписей с экзаменов, а на каких – с творческими людьми), которые позволят исследовать группы, связанных по смыслу файлов. Другим важным аспектом анализа корпуса является выбор активных слоёв. Игнорирование слоёв, создающих зашумление данных, может привести к выявлению более практически полезных шаблонов.

По результатам проведённого исследования, можно выделить следующие достоинства и недостатки разработанных методов разбиения на временные интервалы, представленные в таблице 4.3.

Таблица 4.3 — достоинства и недостатки методов разбиения на временные интервалы.

|  |  |  |
| --- | --- | --- |
| **Метод разбиения** | **Достоинства** | **Недостатки** |
| На одинаковые интервалы | Универсальный (не зависит от выбранных слоёв) | Больше подходит для получения общего представления о данных корпуса, чем для поиска полезных шаблонов |
| На интервалы разной длительности в соответствии с аннотациями на выбранном(главном) слое | Можно выбрать слой, представляющий наибольший интерес для анализа, так как в каждом из выделенных интервалов будет аннотация с этого слоя | Для каждого слоя необходимо проводить отдельный анализ |
| На интервалы разной длительности в соответствии с отметками времени, имеющимися в файле | Позволяет производить наиболее детальное разбиение, не требует настройки | Если для каждой аннотации отметки времени уникальны, разбиение будет слишком подробным и шаблоны перестанут выделяться |

# ЗАКЛЮЧЕНИЕ

В рамках данной работы был проведён анализ предметной области и разбор формата разметки корпуса REC. Кроме того были рассмотрены существующие алгоритмы поиска шаблонов. На основе анализа, было решено для поиска шаблонов использовать модификацию алгоритма Apriori, использующую для представления транзакций бинарные вектора.

Также был разработан и описан метод выделения шаблонов в корпусе REC. Предложен подход для выделения транзакций из данных REC и три способа разбиения на временные интервалы: на одинаковые временные интервалы, на интервалы разной длины по главному слою и на интервалы разной длины по временным слотам.

Разработано ПО RECAnalyzer, использующая в качестве входных данных EAF-файлы корпуса REC и возвращающая пользователю текстовую информацию о найденных транзакциях, часто встречающихся наборах и ассоциативных правилах.

В ходе апробации разработанного метода на данных корпуса REC наборы с наибольшим количеством аннотаций (и при этом поддержкой более 30%) удалось выявить при использовании разбиения на интервалы в соответствии с аннотациями на главном слое – остроты и событие. При разбиении на одинаковые интервалы замечена интересная тенденция, что с увеличением длительности интервала, несмотря на смещение частот основной массы событий в меньшую сторону, начинаю проявляться наборы с поддержкой более 30%. Кроме того, в большинстве случаев, текстовые слои создают шум, поэтому бывает полезно не рассматривать из в ходе анализа, или рассматривать избирательно, например, учитывая только слой структуры.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. А. А. Зинина, Н. А. Аринкин, Л. Я. Зайдельман, А. А. Котов. Разработка модели коммуникативного поведения робота Ф-2 на основе мультимодального корпуса REC. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». – М.:РГГУ. - 2018. – С. 831-843.
2. ELAN (Version 5.0.0). Nijmegen: Max Planck Institute for Psycholinguistics [Электронный ресурс]. – Режим доступа: https://tla.mpi.nl/tools/tla-tools/elan/ (дата обращения 27.05.2019).
3. Е. Б. Солонин. Интеллектуальные технологии поиска и анализа данных [Электронное текстовое издание] – Екатеринбург. – 2015. – С. 43. – Режим доступа: https://study.urfu.ru/Aid/Publication/13334/1/Solonin.pdf (дата обращения 27.05.2019).
4. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.: ил. + CD-ROM.
5. В. Дюк, А. Самойленко. Data mining: учебный курс – СПб.:Питер, 2001. – 368 с:ил. + CD.
6. R. Agrawal, R. Srikant. "Fast Discovery of Association Rules // In Proc. of the 20th International Conference on VLDB. - Santiago, Chile. - 1994.
7. Savasere, E. Omiecinski An Efficient Algorithm for Mining Association Rules in Large Databases // In Proc. 21st Int’l Conf. Very Large Data Bases. - San Francisco – 1995.
8. J.S. Park, M.-S. Chen. An Effective HashBased Algorithm for Mining Association Rules // In Proc. ACM SIGMOD Int’l Conf. Management of Data. - ACM Press - New York. - 1995.
9. V.A. Billig. Effective algorithm for constructing associative rules // Программные продукты и системы. - 2017. - Т. 30, № 2. - С.196–206. - DOI: 10.15827/0236-235X.030.2.196-206.
10. Zhi-Chao Li, Pi-Lian He. A high efficient AprioriTid algorithm for mining association rule. //2005 International Conference on Machine Learning and Cybernetics – China. - 2005. - DOI: 10.1109/ICMLC.2005.1527239.
11. Курс: «Data Mining» // Национальный открытый университет «ИНТУИТ» [Электронный ресурс]. – Режим доступа (свободный): https://www.intuit.ru/studies/professional\_skill\_improvements/2056/courses/504/lecture/11468?page=2 (дата обращения 27.05.2019).

# ПРИЛОЖЕНИЕ А

Таблица A.1 — Слои и словари

|  |  |  |
| --- | --- | --- |
| **Слой/слои** | **Идентификатор словаря** | **Допустимые значения** |
| текст\_дзл, текст2\_дзл | text\_fta | «нет», «возможно дзл», «дзл-просьба», «дзл-оценка», «дзл-отказ», «дзл-другое», «потеря лица», «неуверенность», «комплимент», «п2-оговорка», «п3-вовлечь-ата», «п9-озабоч-потребн-ата», «п11-оптимист», «п14-взаимно», «н3-пессимист», «н4-снизить-ущерб», «н5-уважение», «н6А-извиниться», «н6Б-веские-причины», «н7-деперсонализ», «н8-обычное-событие», «н9-компенсация» (22 сущности словаря) |
| текст\_стру, текст2\_стру | text\_struct | «основная часть», «междометие», «зачин», «неоконч», «дополн к своим», «дополн к другому» (6 сущностей словаря) |
| текст\_ирония,  текст2\_ирония | text\_irnoy | «нет», «ирония», «признаки иронии», «ролевая игра», «острота» (5 сущностей словаря) |
| руки\_способ | hand\_mode | «манип», «указывает», «перечисляет», «иконич», «контакт», «стучит», «трет», «чешет», «машет», «демонстр», «закрывает», «подпирает», «кусает», «скрещивает», «другое» (15 сущностей словаря) |

Продолжение таблицы А.1.

|  |  |  |
| --- | --- | --- |
| рот | Mouth | «облизывается», «улыбается», «смеётся», «трубочка», «жуёт», «сжимает губы», «втягивает губы», «кусает губу», «оскаливается», «двигает губами», «кашляет», «цокает», «плюёт», «втягивает воздух», «выдувает воздух», «открывает рот», «другое» (17 сущностей словаря) |
| событие | event\_type | «задача», «поручение», «неудача-внешн», «неудача-внутр», «успех», «подготовка», «выполнение действия» (7 сущностей словаря) |
| текст\_цель, текст2\_цель | text\_goal | «ассертивная», «директивная», «экспрессивная», «декларативная», «вопрос-разрешения», «вопрос-уточнение», «вопрос-другой», «понимание-согласие-одобрение», «отрицание-несогласие-возражение», «этикетное», «метавысказывание», «сопровождает действие», «коррекция оппонента», «коррекция себя», «комиссивная», «хезитация» (16 сущностей словаря) |

Продолжение таблицы А.1.

|  |  |  |
| --- | --- | --- |
| функция, голова\_функ, тело\_функ, руки\_функ | Func | «понимание-согласие-одобрение», «отрицание-несогласие-возражение», «апелляция», «побуждение», «ожидание-обратной-связи», «остановка-адресата», «отсутствие-невозможность», «обратная связь», «привлечение внимания», «демонстрация непонимания», «демонстрация неадекватности а-та», «я-воодушевление», «я-расслабление-удовлетворение», «я-смущение-фрустрация», «я-хезитация», «я-радость», «я-умиление-сочувствие», «я-размышление», «ты-позитивн-к-объекту», «ты-позитивн-к-адресанту», «ты-сочувствие-к-адресанту», «ты-пренебрежение», «ты-попытка-успокоить», «ты-негативн», «дистанцирование», «об-указатель», «речь-эмфаза», «речь-референт», «речь-операция-с-референтом», «речь-замена-референта», «выбор-варианта», «стимулирование», «компенсация-закрытие», «другое» (34 сущности словаря) |
| руки\_траек | hand\_traj | «фиксировано», «к объекту», «к оппоненту», «от себя», «к себе», «вверх», «вниз», «в дискурс простр», «в окружении», «круговое», «поперек», «назад», «экскурсия», «рекурсия», «реверс», «ожидание», «фрагмент», «другое» (18 сущностей словаря) |

Окончание таблицы А.1.

|  |  |  |
| --- | --- | --- |
| руки\_активн | hand\_act | «палец-указатель», «пальцы», «ладонь», «кулак», «рука-целиком», «адаптер», «другое» (7 сущностей словаря) |
| руки\_пассивн | hand\_pass | «нет», «волосы», «лоб», «глаза-брови», «нос», «щека», «рот-губы», «ухо», «подбородок», «пальцы», «ладонь», «кулак», «рука-целиком», «собств-одежда-тело», «объект», «человек», «окружение», «другое» (18 сущностей словаря) |
| глаза | eyes | «взгляд вбок», «взгляд вверх», «расширяет», «прищур», «закрывает», «подмиг», «поднимает брови», «хмурит брови», «морщит нос», «часто моргает», «другое» (11 сущностей словаря) |
| Итого | | 176 сущностей словаря |

# ПРИЛОЖЕНИЕ Б

"20081223-furekl-a1.eaf","20081225-fipp-a01.eaf", "20081225-fipp-a02.eaf", "20081225-fipp-a03.eaf", "20081225-fipp-a04-m.eaf", "20081225-fipp-a05-m.eaf", "20081225-fipp-a06.eaf", "20081225-fipp-a07.eaf", "20081225-fipp-a08-m.eaf", "20081225-fipp-a09.eaf", "20081225-fipp-a10.eaf", "20081225-fipp-a11.eaf", "20081225-fipp-a12.eaf", "20081225-fipp-a13.eaf", "20081225-fipp-a14-m.eaf", "20081225-fipp-a15.eaf", "20081225-fipp-a16-m.eaf", "20081225-fipp-a17-m.eaf", "20081225-fipp-a18.eaf", "20081225-fipp-a19.eaf", "20081225-fipp-a20.eaf", "20081225-fipp-a21.eaf", "20081225-fipp-a22.eaf", "20081225-fipp-a23.eaf", "20081225-fipp-a24-m.eaf", "20081225-fipp-a25-m.eaf", "20081225-fipp-a26.eaf", "20081225-fipp-a26-m.eaf", "20081225-fipp-a27.eaf", "20081225-fipp-a28-m.eaf", "20081225-fipp-a29-m.eaf", "20081225-fipp-a30.eaf", "20081225-fipp-a31.eaf", "20081225-fipp-b1-m.eaf", "20081225-fipp-b2.eaf", "20081225-fipp-b3-m.eaf", "20081225-fipp-b4.eaf", "20081225-fipp-b5.eaf", "20081225-fipp-b6.eaf", "20081225-fipp-b7.eaf", "20081225-fipp-b8.eaf", "20081225-fipp-b9.eaf", "20081225-fipp-b10-m.eaf", "20081225-furekl-a01.eaf", "20081225-furekl-a01-m.eaf", "20081225-furekl-a02.eaf", "20081225-furekl-a02-m.eaf", "20081225-furekl-a03.eaf", "20081225-furekl-a05.eaf", "20081225-furekl-a06-m.eaf", "20081225-furekl-a07.eaf", "20081225-furekl-a08-m.eaf", "20081225-furekl-a09-m.eaf", "20081225-furekl-a10.eaf", "20081225-furekl-a11.eaf", "20081225-furekl-a12.eaf", "20081225-furekl-a13.eaf", "20081225-furekl-a14.eaf", "20081225-furekl-a15.eaf", "20081225-furekl-a16.eaf", "20081225-zhurn-b1.eaf", "20081225-zhurn-b2-m.eaf", "20081225-zhurn-b3-m.eaf", "20081225-zhurn-b4-m.eaf", "20081225-zhurn-b5-m.eaf", "20081225-zhurn-b6-m.eaf", "20081225-zhurn-b8-m.eaf", "20081226-omoiai-a1.eaf", "20081226-omoiai-a2.eaf", "20081226-omoiai-a3.eaf", "20081226-omoiai-a4.eaf", "20081226-omoiai-a5.eaf", "20081226-zhurn-a1.eaf", "20081227-fumo-a01.eaf", "20081227-fumo-a02.eaf", "20081227-fumo-a03.eaf", "20081227-fumo-a04.eaf", "20081227-fumo-a05.eaf", "20081227-fumo-a06.eaf", "20081227-fumo-a07.eaf", "20081227-fumo-a08.eaf", "20081227-fumo-a09.eaf", "20081227-fumo-a10.eaf", "20081227-fumo-a11.eaf", "20081227-fumo-a12.eaf", "20081227-fumo-a13.eaf", "20081227-fumo-a14.eaf", "20081227-fumo-a15.eaf", "20081227-fumo-a16.eaf", "20081227-fumo-a17.eaf", "20081227-fumo-a18.eaf", "20081227-fumo-a19.eaf", "20081227-fumo-a20.eaf", "20081227-fumo-a21.eaf", "20081227-fumo-a22.eaf", "20081227-fumo-a23.eaf", "20081227-fumo-a24.eaf", "20081227-fumo-a25.eaf", "20081227-fumo-a26.eaf", "20081227-fumo-a27.eaf", "20081227-fumo-a28.eaf", "20081229-a1-fumo.eaf", "20081229-a2-fumo.eaf", "20081229-a3-fumo.eaf", "20081229-a5-fumo.eaf", "20081229-a6-fumo.eaf", "20081229-a7-fipppr.eaf", "20081229-a8-fipppr.eaf", "20081229-a9-fipppr.eaf", "20081230-a01-furekl.eaf", "20081230-a02-furekl.eaf", "20081230-a03-furekl.eaf", "20081230-a04-furekl.eaf", "20081230-a05-furekl.eaf", "20081230-a06-furekl.eaf", "20081230-a07-furekl.eaf", "20081230-a08-furekl-m.eaf", "20081230-a09-furekl.eaf", "20081230-a10-furekl.eaf", "20081230-a10-furekl-m.eaf", "20081230-a11-furekl.eaf", "20081230-a12-furekl.eaf", "20081230-a13-furekl-m.eaf", "20081230-a14-furekl.eaf", "20081230-a15-furekl.eaf", "20081230-a16-furekl.eaf", "20081230-a17-furekl.eaf", "20081230-a18-furekl.eaf", "20081230-a19-furekl.eaf", "20081230-a20-furekl.eaf", "20081230-a21-furekl-m.eaf", "20081230-a22-furekl.eaf", "20081230-a23-furekl.eaf", "20081230-a24-furekl-m.eaf", "20081230-a25-furekl-m.eaf", "20081230-a26-furekl.eaf", "20081230-a28-furekl.eaf", "20081230-a29-furekl.eaf", "20081230-a30-furekl.eaf", "20081230-b1-furekl.eaf", "20081230-b2-furekl.eaf", "20081230-b3-furekl-m.eaf", "20081230-b4-furekl-m.eaf", "20081230-b5-furekl.eaf", "20081230-b6-furekl.eaf", "20081230-b7-furekl.eaf", "20081230-b8-furekl.eaf", "20081230-b8-furekl-m.eaf", "20081231-a1-furekl.eaf", "20081231-a2-furekl-m.eaf", "20081231-a3-furekl.eaf", "20081231-a4-furekl-m.eaf", "20081231-a5-furekl-m.eaf", "20081231-a6-furekl.eaf", "20081231-a7-zhurn.eaf", "20081231-a8-furekl-m.eaf", "20081231-a9-furekl-m.eaf"