Описание функциональных характеристик

В данном программном документе приведено описание применения программного обеспечения «Фреймворк корпусов языков России», содержащего набор функций, предназначенных для агрегации результатов, поступивших из разных источников.

В данном программном документе, в разделе «Назначение программы» приведено описание назначения программного модуля, возможности данного программного модуля, а также его основные характеристики и ограничения, накладываемые на область применения программного модуля.

В разделе «Условия применения» указаны условия, необходимые для работы программного модуля (требования к необходимым для данной программы техническим средствам, и другим программам, общие характеристики входной и выходной информации, а также требования и условия организационного, технического и технологического характера).

В данном программном документе, в разделе «Описание задачи» указаны определения задачи и методы ее решения.

В разделе «Входные и выходные данные» указаны сведения о входных и выходных данных.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящем описании программного продукта применяют следующие термины с соответствующими определениями:

Агрегация — процесс объединения в виде единой удобной для обработки величины нескольких единиц информации.

Агрегатор — программа или программный продукт, выполняющий задачу агрегации.

Глосса — специализированная помета, соответствующая одно из возможных грамматических явлений.

Грамматика — описание грамматического строя языка.

Контекст — текст правее и левее какой-то части текста.

Корпус — коллекция текстов, когерентная и снабженная метаинформацией.

Корпусный — предназначенный или имеющий отношение к корпусам.

Малые языки России — языки России, отличные от титульных для регионов, или же с малым числом говорящих на этих языках как на родных.

Типологический — относящийся к лингвистической типологии.

Токен сессии — специальный шифр, присваиваемый пользователю веб-сервиса, с целью его идентификации в процессе работы с сервисом.

Cookies — набор специальных шифров, получаемых пользователем веб-сервиса для реализации механизма запоминания состояния работы с сервисом между сессиями.

1. Назначение программы

1.1. Назначение программы

ПО будет использоваться исследователями и разработчиками многоязычных систем автоматической обработки языка, использующими в качестве источников данных грамматически аннотированные тексты на языках России. ПО предназначено для использования специалистами в области языкового искусственного интеллекта. ПО обеспечивает возможность подготовки данных для дообучения, оценки и интерпретации языковых моделей на языках России.

1.2. Возможности программы

«Фреймворк корпусов языков России» – это платформа, которая позволяет проводить типологические корпусные исследования: с помощью нее существующие корпуса на платформе TSAKorpus (https://github.com/timarkh/tsakorpus) были собраны в единую поисковую систему. Программа способна выполнять поиск по заданному пользователем списку языков. На данный момент поисковый запрос может состоять из следующих частей:

а) поиск по подстроке оригинала или перевода;

б) поиск по грамматике;

в) поиск по глоссам.

Это базовый набор функций, доступный практически в каждой сборке TSAKorpus.

Система выдаёт результаты по выбранным пользователем корпусам в интерфейсе со вкладками. Каждая вкладка отвечает за определенный корпус и не зависит от других вкладок.

Также разработан отдельный упрощенный интерфейс Tsakorpus, который позволяет:

а) выбирать набор корпусов, по которым будет осуществляться поиск;

б) использовать только базовый функционал поискового движка Tsakorpus – поиск по подстроке, глоссам и грамматике.

На данный момент агрегатор корпусов способен поддерживать работу с большим числом корпусов на платформе TSAKorpus, что позволяет делать типологические исследования, опираясь на данные корпусов малых языков.

1.3. Основные характеристики программы

Исходным языком программирования для программного модуля является Python версии 3.7. Языками верстки интерфейса являются HTML и CSS, для создания интерактивных элементов используется язык JavaScript.

1.4. Ограничения, накладываемые на область применения программы

Программный модуль предназначен для работы с набором данных, содержащихся в корпусах на платформе TSAKorpus.

2. Условия применения

2.1. Требования к техническим (аппаратным) средствам

Состав технических средств:

- Компьютер оператора, включающий в себя:
  - Процессор x86_64 с тактовой частотой не менее 2 ГГц;
  - Оперативную память объемом не менее 16 ГБ (для запуска ОС и одного программного модуля);
  - Мышь, клавиатура, монитор
  - доступ к сети Интернет
- Серверный компьютер (в случае размещения для публичного использования), включающий в себя:
  - Процессор x86_64 с тактовой частотой не менее 2 ГГц;
  - Оперативную память объемом не менее 16 ГБ (для запуска ОС и одного программного модуля);
  - доступ к сети Интернет.

2.2. Требования к программным средствам (другим программам)

Для функционирования в установленном режиме необходимо обеспечить установку следующих программных средств:

- операционная система Ubuntu не ниже 16.04;
- интерпретатор Python версии не ниже 3.7.
- настроенная система docker compose

Для реализации программного модуля использовались следующие сторонние открытые библиотеки:

- Pandas (библиотека Python для обработки и анализа структурированных данных);
- Flask (фреймворк Python для создания веб-приложений);
- Re (модуль Python для поддержки регулярных выражений);
- Requests (библиотека, позволяющая отправлять запросы HTTP в Python);
- Kioskboard (библиотека JavaScript для использования виртуальных клавиатур);
- DevBridge Autocomplete (библиотека для создания полей автоматического заполнения/подсказок для полей ввода текста);
- Bootstrap Icons (библиотека с набором иконок).

Данные библиотеки имеют свободные лицензии, никак не ограничивающие их использование в коммерческих и некоммерческих проектах.

2.3. Общие характеристики входной информации

Входными данными является запрос пользователя, полнотекстовый или дополненный указанием интересующих пользователя глосс, сформированный в веб-интерфейсе.

2.4. Общие характеристики выходной информации

Выходными данными является веб-страница с пагинацией, предоставляющая возможность просмотреть все данные из сконфигурированных источников, удовлетворяющие запросу пользователя.

2.5. Требования и условия организационного характера

Для корректной работы с программным модулем персонал должен убедиться в правильности формата входных данных при запуске программы, необходим доступ к сети Интернет для запуска программы.

2.6. Требования и условия технического характера

Для работы программного модуля не требуется обеспечения каких-либо особых требований и условий технического характера.

2.7. Требования и условия технологического характера

Для работы программного модуля не требуется обеспечения каких-либо особых требований и условий технологического характера.

3. Описание задачи

3.1. Определение задачи

Основная задача агрегатора ПО на данный момент – собрать существующие корпуса на платформе Tsakorpus (https://github.com/timarkh/tsakorpus) в единую поисковую систему; создать такой ресурс, который бы помог лингвистам-типологам в изучении корпусных данных. TSAKorpus выбран в качестве отправной точки по той причине, что выкладывание корпусов на этой платформе всё больше набирает популярность среди полевых лингвистов, занимающихся малыми языками России. На данный момент существует около сорока таких корпусов, что уже приближается к минимальной языковой выборке.

3.2. Методы решения задачи

Основной метод, используемый в данном программном модуле - агрегация множественных результатов из различных источников информации. В основном такая агрегация происходит автоматически, однако для ресурсов допустима также ручная доработка логики агрегации с целью поддержки специфичных для этих ресурсов функций.

Агрегация корпусов устроена следующим образом. Ввиду того, что обкачать все корпуса на платформе TSAKorpus невозможно (существует ограничение на 10000 предложений), каждый раз для выполнения запроса посылается запрос в соответствующий TSAKorpus. Программа имитирует работу пользователя: заходит на сайт с определенным токеном сессии и затем с помощью этого токена делает запросы. Для хранения этих токенов используется система браузерных cookies, в которых кроме токена сессии для каждого корпуса хранятся и номера страниц в выдаче. Это необходимо, чтобы сделать выдачи независимыми друг от друга – чтобы при переключении страницы в одной вкладке (в одном корпусе) не происходило переключения страницы во всех остальных. Таким образом, агрегатор не собирает данные по всем языкам в одно место, а обращается к веб-интерфейсу существующих хранилищ и получает данные в режиме онлайн. Такой способ получения информации лучше, чем просто обкачивание выложенных данных, потому что данные в корпусах могут обновляться и изменяться, но это не станет проблемой для системы в режиме онлайн.

Поиск по подстроке устроен следующим образом. Пользователь вводит любую подстроку, а затем выбирает, хочет ли он её найти в языке оригинала или в языке перевода (обычно это русский или английский). После этого агрегатор обращается к выбранным корпусам и выдаёт предложения в том виде, в котором они были представлены в исходной сборке TSAKorpus для этого языка. После этого выдача нормализуется (приводится к одному виду). Выдача может быть как в глоссированном виде, так и в виде простого текста с грамматическими тегами, появляющимися во всплывающих окнах при наведении курсора на словоформу (как и в оригинальном TSAKorpus).

Поиск по глоссам и грамматике устроен более сложным образом. Разные сборки TSAKorpus используют разный набор условных обозначений для описания грамматических категорий и морфемного членения словоформ. На данный момент в интерфейсе представлен свободный ввод любой глоссы/грамматической категории,

разные варианты одного и того же не объединяются. Например, для обозначения творительного падежа в одном корпусе используется обозначение INS, а в другом – INST.

Выдача также приведена к единому виду, чтобы система выглядела более цельно. Несмотря на изменения в интерфейсе, функционал TSAKorpus был практически полностью сохранен: пользователь получает выдачу с грамматическими характеристиками и глоссами и может скачать полученные результаты.

4. Входные и выходные данные

4.1. Сведения о входных данных

Входными данными программы являются: перечень подлежащих агрегации корпусных менеджеров. Во время работы программы поисковый интерфейс предполагает указание следующих входных полей: Word, Lemma, Grammar, Language.

Входные данные указываются в переменной скрипта на языке программирования Python версии 3.8 CORPORA, которая требует указания списка (list) триплетов (тоже типа list) следующей структуры:

<код языка в формате iso 639 или подобном, ссылка на сайт с веб-корпусом языка, названия языка на английском языке>

Настройка хоста и порта работы приложения предполагаются через прокси сервер, подобный NGinx.

Входные поля Word, Lemma, Grammar, Language указываются в формате, предложенном веб-интерфейсом (см. Рисунок 1).

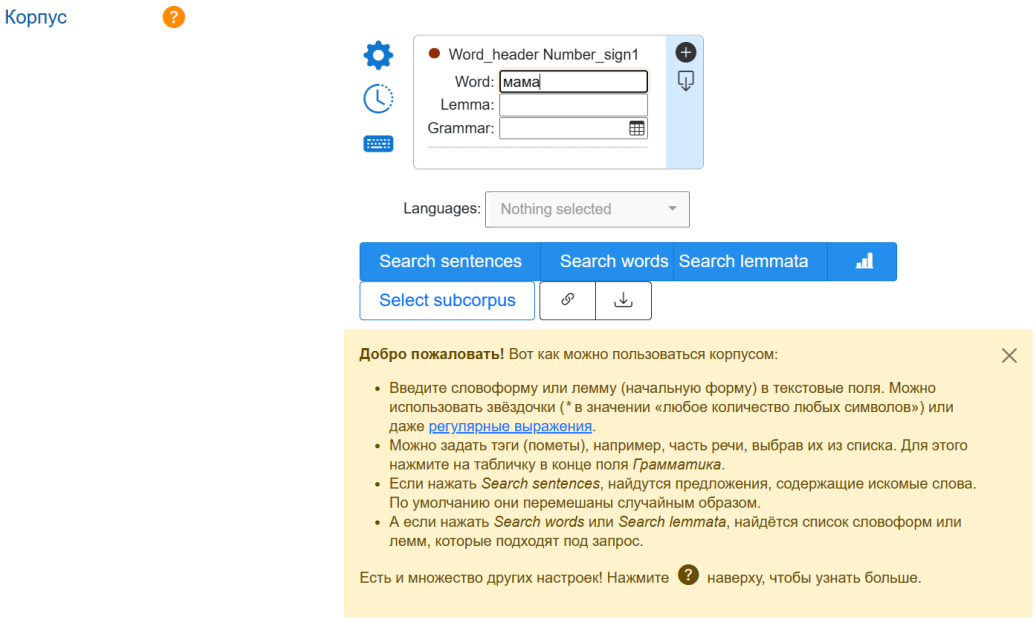

Рисунок 1 – Поля для заполнения

4.2. Сведения о выходных данных

Выходными данными программы являются веб-интерфейс поиска, в котором в ответ на настроенные входные данные запроса отображается результат поиска по соответствующим корпусам.

Пример выходных данных программы приведён на Рисунке 2.

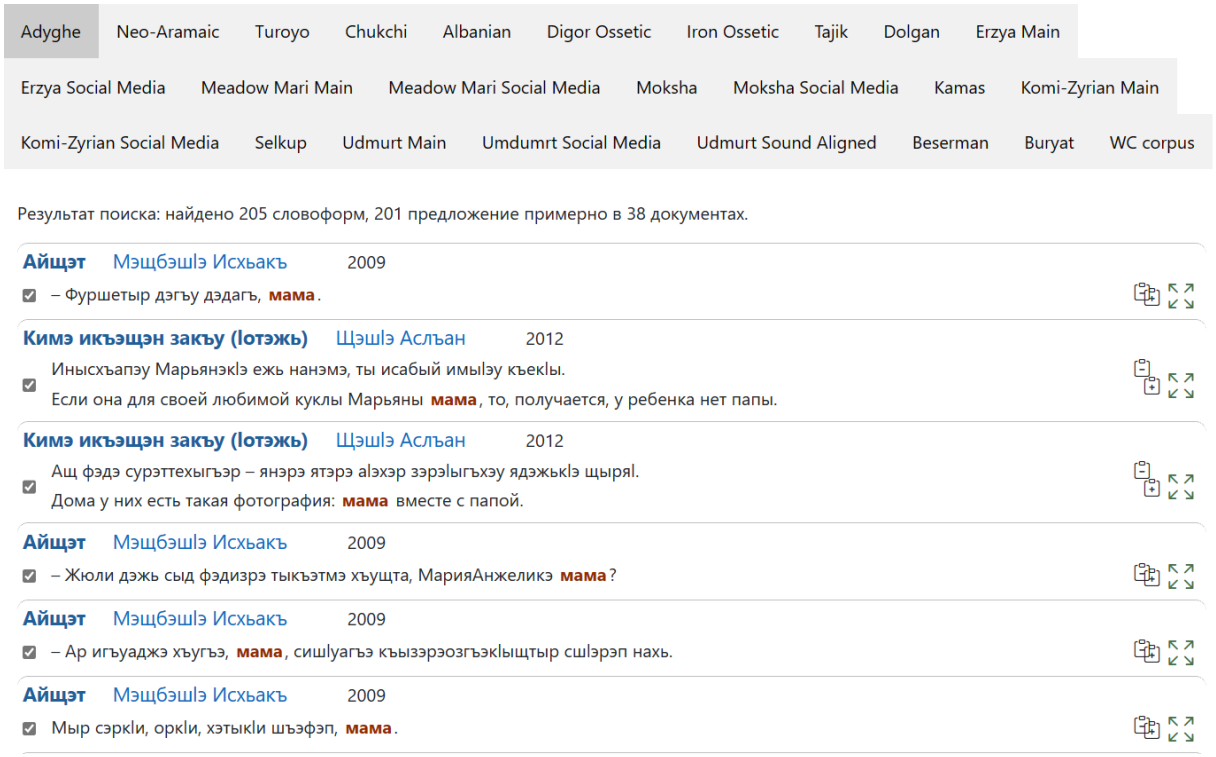

Рисунок 2 – Пример выходных данных программы