Описание программы

«Вэб-сервис психотипирования человека по тексту: оценка темпераментов и радикалов.»

АННОТАЦИЯ

В данном программном документе приведено описание Системы рекомендаций «Вэб-сервис психотипирования человека по тексту: оценка темпераментов и радикалов.» (далее Clientype), представляющей собой сервис рекомендации оптимальной стратегии коммуникации с клиентами на основе психометрических моделей анализа их поведения.

В данном программном документе, в разделе «Общие сведения» приведено описание обозначения программы и её возможных наименований, программное обеспечение необходимое для функционирования программы, и языки программирования, на которых написана данная программа.

В разделе «Функциональное назначение» описано для исполнения каких функций была написана программа.

В данном программном документе, в разделе «Описание логической структуры» представлена логическая структура. В частности, рассказан алгоритм программы, какие методы для осуществления алгоритмов используются, структура программы с описанием функций составных частей и связей между ними.

В разделе «Используемые технические средства» указаны средства, необходимые для работы программы.

Раздел «Вызов и загрузка» поясняет как вызывать функции библиотеки с носителя данных и рассказывает о входных точках в программу.

В разделе «Входные данные» указаны сведения о характере и формате выходных данных.

В разделе «Выходные данные» указаны сведения о характере и формате выходных данных.

Оформление программного документа «Описание программы» произведено по требованиям ЕСПД ГОСТ

Т 19.402-78.[1]

[1] ГОСТ 19.402-78 ЕСПД. Описание программы. Требования к содержанию и оформлению

СОДЕРЖАНИЕ

1. Общие сведения

1.1. *Обозначение и наименование программы*

Способами обращения к Системе рекомендаций Clientype является либо обращение к серверу по REST API, либо через телеграм-бот, либо через jupyter notebook в развернутом python-окружении. В данном документе подробно рассмотрим именно последний способ.

1.2. *Программное обеспечение, необходимое для функционирования программы*

Для функционирования ЭО ПК СРОСКК входящей в комплекс необходимы следующие сторонние библиотеки:

- dostoevsky==0.6.0
- morpholog==1.6
- numpy==1.19.5
- pandas==1.3.5
- pingouin==0.3.11
- pymorphy2==0.9.1
- pymystem3==0.2.0
- seaborn==0.11.2
- simple-elmo==0.8.0
- snowballstemmer==2.1.0
- spacy==3.0.6
- stanza==1.2
- tensorflow==2.5.0
- termcolor==1.1.0
- wordfreq==2.5.0

1.3. *Языки программирования, на которых написана программа*

Библиотека написана на языке Python.

2. Функциональное назначение

Система рекомендаций Clientype предназначена для рекомендации оптимальной стратегии коммуникации с клиентами на основе психометрических моделей анализа их поведения. На вход система принимает русскоязычные текстовые данные.

3. Описание логической структуры

3.1. *Описание логической структуры*

Процесс отыскания распределения психотипов и темпераментов по тексту начинается с выбора квантилей текстовых признаков, относительно которых будут производиться дальнейшие расчеты. В общем случае, квантили обозначают выборку, относительно которой будет психотипироваться рассматриваемый человек. Сам процесс психотипирования построен на основе алгоритма, построенного на наборе индикаторов психотипов для каждого текстового признака нашими психологами. Другими словами, значение признака по тексту сравнивается с квантилями по выборке и в зависимости от интервала между квантилями, в который попал признак, срабатывает/не срабатывает соответствующий индикатор. Далее, в разрезе каждого психотипа индикаторы суммируются по всем признакам. Таким образом, на выходе получается распределение психотипов. Аналогично устроено получение распределения темпераментов

Как уже было описано в отчете по НИР, нами рассматривались следующие группы текстовых признаков: синтаксис, морфология, семантика. Для их детектирования использовались такие nlp-модели, как mystem, stanza, morpholog, pymorphy, spacy и dostoevsky.

Для того, чтобы выбрать кастомный тип квантилей (например, по группе клиентов вашей компании), необходимо обработать корпус текстов ваших клиентов моделями детектирования текстовых признаков. Полученную таблицу pandas.dataframe с результатами необходимо сохранить в соответствующую вашей компании папку внутри папки data.

Для распределения психотипов и темпераментов доступны 2 вида нормировок: по l1-норме (на сумму весов) и по Чебышевской норме (на максимальный вес).

3.2. *Алгоритм программы*

3.2.1. Фильтрация

Фильтрация решает проблему “плохих” токенов в тексте разработанными нашей командой правилами. В текстах часто встречаются неинформативные токены, являющиеся несвязанным набором букв или цифр. Это может происходить, например, из-за того, что человек таким образом дополнял текст до 300 символов. Также при детектировании большинства признаков мы подвергаем фильтрации союзы, предлоги и частицы (не

учитываем их при рассчёте общего количества слов в тексте). На основе EDA корпуса рассматриваемых нами текстов было принято решение об их фильтрации по причине наличия представителей данных частей речи практически в каждом тексте выборки Assessty, на которой разрабатывались семантических, синтаксические и морфологические модели. Напомним, что выборка текстов Assessty представляет собой корпус текстов людей одной профессии, прошедших психологическое тестирование.

3.2.2. Подготовка данных к расчету квантилей

Как можно понять выше, получение квантилей для моделей психометрии основывается на рассчёте текстовых характеристик для кастомного корпуса текстов. Для использования достаточного наличия корпуса текстов, принадлежащих клиентам одного сегмента (людям одной профессии и т.д.). Также необходимо, чтобы тексты были содержательные и имели длину не менее 300 символов.

3.3 Используемые методы

Для реализации своих функций алгоритмы программы используют скрипты, позволяющие производить обработку нового корпуса текстов (рассчёт семантических, синтаксических и морфологических признаков), строить распределение психотипов и темпераментов по данному на вход тексту.

3.4 Структура программы с описанием функций составных частей и связи между ними

Основной модуль со скриптами и моделями:

– text.py: методы анализа одного конкретного текста (в разрезе текстовых характеристик)

– psychotype.py: методы психотипирования на основании рассчитанных текстовых характеристик и квантилей

– for_dict_processing.py: методы предобработки словарей с чувственными прилагательными, методы пополнения словаря глаголов, связанных с ощущениями

Модуль примеров работы и визуализации результатов:

– psychotype_by_text.ipynb: психотипирование клиента по тексту с визуализацией всех промежуточных шагов работы моделей

– df_for_average_text_features.ipynb: скрипт для обработки нового корпуса текстов

Модуль с данными:

– public_df.pkl: таблица pandas.DataFrame для рассчёта квантилей с синтаксическими признаками по корпусу текстов Assessty

– public_semantic_role_df.pkl: таблица pandas.DataFrame для рассчёта квантилей с семантическими признаками по корпусу текстов Assessty

– morph_df.pkl: таблица pandas.DataFrame для рассчёта квантилей с морфологическими признаками по корпусу текстов Assessty

4. Используемые технические средства

Для запуска Система рекомендаций Joys требуется компьютер с одной из оперативных систем: Ubuntu версии 14.04.5 или выше, Linux Mint версии 17.3 или выше, Arch Linux, Manjaro, Fedora 24 или выше, openSUSE Leap версии 42.1 или выше, MacOS-X, Windows версии 8 или выше.

5. Вызов и загрузка

Для начала работы необходимо убедиться, что имеются необходимые для работы stanza и elmo модели ресурсы для русского языка (директории analytics_lib.notebooks.stanza_resources и analytics_lib.notebooks.elmo_resources)

5.1 Способ вызова программы с соответствующего носителя данных

Необходимо создать Python-окружение с версией Python не ниже 3.8

Установить в него все требуемые для успешной работы программы библиотеки. Сделать это можно командой pip install requirements.txt

Скачать по ссылками из readme нашего github репозитория ресурсы для русского языка stanza и elmo моделей. Поместить их в директории analytics_lib.notebooks.stanza_resources и analytics_lib.notebooks.elmo_resources

Выгрузить ресурсы для русского языка для nlp-библиотек dostoevsky и spacy путём выполнения следующих команд:

python3 -m dostoevsky download fasttext-social-network-model

python3 -m spacy download ru_core_news_sm

5.2 Входные точки в программу

После выполнения всех предыдущих шагов необходимо запустить jupyter-notebook, перейти в директорию analytics_lib.notebooks и открыть файл psychotype_by_text.ipynb

6. Входные данные

6.1 Характер, организация и предварительная подготовка входных данных

Тетрадка **analytics_lib.notebooks.psychotype_by_text.ipynb** использует следующие файлы, которые должны лежать в соответствующих директориях

Директория analytics_lib.data:

– **public_df.pkl**: таблица с посчитанными синтаксическими признаками по всем текстам из БД для усреднения

– **public_semantic_role_df.pkl**: таблица с посчитанными семантическими признаками по всем текстам из БД для усреднения

– **morph_df.pkl**: таблица с посчитанными морфологическими признаками по всем текстам из БД для усреднения

– **df_sense.pkl**: таблица, содержащая словарь с чувственными прилагательными

– **public_modality_df.pkl**: таблица с посчитанными модальностями по всем текстам из БД для усреднения

– **verbs_df.pkl**: таблица, содержащий словарь с глаголами для внутренних и внешних предикатов, разбитых по категориям

– **df_sense.pkl**: таблица, содержащий словарь с чувственными прилагательными

– **df_for_new_labeling.pkl**: таблица с посчитанными значениями индикаторов шкал "Хаотичная/Системная" и "Локальная/Глобальная" по всем склеенным текстам из БД

– **df_for_rand_syst_indicators.pkl**: таблица с посчитанными значениями индикаторов шкалы "Хаотичная/Системная" по всем токенам из БД

Директория analytics_lib.nlp_texts:

– **text.py**: библиотека методов обработки конкретного текста в разрезе признаков (морфология, синтаксис, семантика)

– **psychotype.py**: библиотека моделей психометрии конкретного текста, предварительно обрабортанного в разрезе признаков

– **for_dict_processing.py**: скрипт для обновления словаря с глаголами, связанными с различными категориями ощущений

Директория analytics_lib.notebooks:

– **stanza_resources**: файлы с ресурсами для русского языка stanza-модели

– **elmo_resources**: файлы с ресурсами для русского языка elmo-модели

7. Выходные данные

7.1 Характер и организация выходных данных

Выходными данными скрипта psychotype.py являются вектора психотипов и темпераментов, а также визуализация процесса работы моделей в интерактивном формате jupyter-notebook

7.2 Формат, описание и способ кодирования выходных данных

Выходными данные можно сохранить в виде JSON-совместимых строки в кодировке utf-8.

Перечень принятых сокращений

Лист регистрации изменений									
Номера листов (страниц)					Всего листов (страниц) в докум.	№ докумен-та	Входящий № сопроводит. докум. и дата	Подп.	Дата
Изм.	измен-ных	замен-ных	новых	аннули-рованных					