

TITANIS API

TITANIS предоставляет функции вычисления текстовых параметров и обработки естественного языка для оценки психоэмоционального состояния автора текста или группы авторов. Помимо широко используемых подходов обработки естественного языка, таких как tf-idf и анализ тональности, TITANIS включает в себя психолингвистический, семантический, дискурсивный и другие виды анализа, которые позволяют обнаруживать различия в текстах пользователей с разными психоэмоциональными состояниями.

1 Внешний интерфейс

Входные данные

Параметр	Тип	Default	Описание
text	string		Непустой текст на русском языке в кодировке UTF-8. Для каждой функции установлен свой минимальный размер текста в символах.
Опциональные параметры, определяющие список вызываемых обработчиков			
psy_cues	bool	False	Рассчитывать психолингвистические/морфологические маркеры
psy_cues_normalization	string	'words'	Условия нормализации для психолингвистических/морфологических маркеров
psy_dict	bool	False	Расчет словарных маркеров
psy_dict_normalization	string	'words'	Условия нормализации для словарных маркеров
syntax	bool	False	Расчет синтаксических признаков
			Найти семантические роли при эмотивах
depression_clf	bool	False	Предсказать депрессивность автора текста по формату эссе Вне зависимости от заданных параметров запускает анализ признаков psy_cues и psy_dict
			Предсказать депрессивность автора текста по

			формату социальных сетей
frustration_clf	bool	False	Предсказать состояние фрустрации у автора текста
discourse	bool	False	Расчет дискурсивных признаков
discourse_long_texts_only	bool	True	Находить дискурсивные признаки только для текстов длинее одного предложения
rosenzweig	bool	False	Выявление типа реакции на фрустрацию по Розенцвейгу

Выходные данные

Результат имеет формат словаря с ключами для обработчиков:

- 'Psycholinguistic cues' - [Расчет психолингвистических/морфологических маркеров](#), [Предсказание депрессии по формату эссе](#)
- 'Psydict' - [Расчет словарных маркеров](#), [Предсказание депрессии по формату эссе](#)
- 'Syntax' - [Расчет синтаксических признаков](#)
- [2.4]
- 'Predicted depression' - [Предсказание депрессии по формату эссе](#)
- [2.6]
- 'Predicted frustration' - [Предсказание состояния фрустрации у автора текста](#)
- 'Discourse features' - [Расчет дискурсивных признаков](#)
- 'Rosenzweig reaction' - [Выявление типа реакции на фрустрацию по Розенцвейгу](#)

2 Внутренний интерфейс обработчиков

2.1 Расчет психолингвистических/морфологических маркеров

Входные данные

Параметр	Тип	Описание
text	string	Непустой текст на русском языке
normalization	string, {'words', 'sentences', 'abs'}, default: 'words'	Условия нормализации для маркеров по частям речи: 'words' - по кол. слов; 'sentences' - по кол. предложений; 'abs' - абсолютные значения.

Основа

Леммы слов, морфологическая аннотация mystem, словарь тональности linis-crowd.

Используемые анализаторы

Название	Ключи
MyStem	lemma, postag (не конвертированный)

Выходные данные

Параметр	Тип	Описание
char_count	Int	Число символов
word_count	Int	Число слов
sentence_count	Int	Число предложений
unique_words_count	Int	Число уникальных слов
punctuation_count	Int	Число знаков препинания
punctuation_per_word	float	Число знаков пунктуации / Число слов
uppercase_rate	float	Доля прописных символов
mean_word_len	float	Средняя длина слов (в количестве символов)
mean_sentence_len	float	Средняя длина предложения (в количестве слов)
unique_words_rate	float	Число уникальных слов / Число слов
verbs_1p_rate	float	Доля глаголов 1 лица
verbs_2p_rate	float	Доля глаголов 2 лица
verbs_3p_rate	float	Доля глаголов 3 лица
verbs_past_tense_rate	float	Доля глаголов прошедшего времени
infinitives_rate	float	Доля инфинитивов
pro_1p_rate	float	Доля местоимений 1 лица
pro_1p_sing_rate	float	Доля местоимений 1 лица ед. числа
pro_1p_plural_rate	float	Доля местоимений 1 лица мн. числа
pro_2p_rate	float	Доля местоимений 2 лица
pro_3p_rate	float	Доля местоимений 3 лица
trager_coef	float	Коэффициент Трейгера
logical_coh_coef	float	Коэффициент логической связности
verbs_per_nouns_coef	float	Коэффициент опредмеченности действия (кол. глаголов / кол. существительных)
participles_gerunds_coef	float	Причастия + деепричастия / Число предложений
negation_rate	float	Число отрицательных приставок и форм (не,ни,бес,без) / Число слов

postag_A	float	Часть речи: прилагательное
postag_ADV	float	Часть речи: наречие
postag_ADVPRO	float	Часть речи: местоименное наречие
postag_ANUM	float	Часть речи: числительное-прилагательное
postag_APRO	float	Часть речи: местоимение-прилагательное
postag_COM	float	Часть речи: часть композита - сложного слова
postag_CONJ	float	Часть речи: союз
postag_INTJ	float	Часть речи: междометие
postag_NUM	float	Часть речи: числительное
postag_PART	float	Часть речи: частица
postag_PR	float	Часть речи: предлог
postag_S	float	Часть речи: существительное
postag_SPRO	float	Часть речи: местоимение-существительное
postag_V	float	Часть речи: глагол

2.2 Расчет словарных маркеров

Входные данные

Параметр	Тип	Описание
text	string	Непустой текст на русском языке
normalization	string, {‘words’, ‘sentences’,‘abs’}, default: ‘words’	Условия нормализации встречаемости слов из словарей: ‘words’ - по кол. слов; ‘sentences’ - по кол. предложений; ‘abs’ - абсолютные значения.

Основа

Леммы слов и словари.

Используемые анализаторы

Название	Ключи
MyStem	lemma

Выходные данные

Параметр	Тип	Описание
Наименования словарей: 1. tgw_positive_assessment 2. tgw_positive_social 3. tgw_positive_emotions 4. tgw_negative_assessment 5. tgw_negative_social 6. tgw_negative_emotions 7. tgw_motivation_activity 8. tgw_cognitive_communication 9. tgw_destructive_activity 10. tgw_affect_lex 11. tgw_bodily_states_emotions 12. tgw_invectives 13. tgw_soft_invectives 14. tgw_obscene_lex 15. tgw_youth_jargon 16. tgw_hcs 17. tgw_economics 18. tgw_catastrophes 19. tgw_security_structures 20. tgw_healthcare_demography_ecology 21. tgw_authority	float	Частотность встречаемости слов из словарей: 1. Рациональная оценка положительная 2. Социальность положительная 3. Эмоции позитивные 4. Рациональная оценка отрицательная 5. Социальность отрицательная 6. Эмоции негативные, стеничные 7. Мотивация, деятельность и напряжение 8. Когнитивная деятельность и коммуникация 9. Деструктивная активность 10. Аффектогенная лексика 11. Эмоции и телесные состояния негативные пассивные 12. Инвективы 13. Мягкие инвективы 14. Обценная лексика 15. Молодежный жаргон 16. Тематическая ЖКХ 17. Тематическая Экономика 18. Тематическая Катастрофы 19. Тематическая Силовые структуры 20. Социальность предметная Здравоохранение, демография и экология 21. Тематическая Власти
positive	float	Частотность позитивных эмотивов в тексте
negative	float	Частотность негативных эмотивов в тексте
ambivalent	float	Частотность амбивалентов в тексте
de_emotives	float	Частотность деэмотивов в тексте
sentiment_rate	float	Тональность текста на основе словаря linis-crowd

2.3 Расчет синтаксических признаков

Входные данные

Параметр	Тип	Описание
text	string	Непустой текст на русском языке

normalization	string, {‘words’, ‘sentences’, ‘abs’}, default: ‘words’	Условия нормализации для маркеров по частям речи: ‘words’ - по кол. слов; ‘sentences’ - по кол. предложений; ‘abs’ - абсолютные значения.
---------------	---	--

Основа

Синтаксическая аннотация.

Используемые анализаторы

Название	Ключи
UDPipe	syntax_dep_tree

Выходные данные

Параметр	Тип	Описание
max_synt_tree	int	Максимальная глубина синтаксического дерева
min_synt_tree	int	Минимальная глубина синтаксического дерева
mean_synt_tree	float	Средняя глубина синтаксического дерева
Наименование синтаксических связей: 1. synt_rel_flat:foreign: 2. synt_rel_csubj 3. synt_rel_acl 4. synt_rel_acl:relcl 5. synt_rel_advcl 6. synt_rel_advmod 7. synt_rel_amod 8. synt_rel_appos 9. synt_rel_aux 10. synt_rel_aux:pass 11. synt_rel_case 12. synt_rel_cc 13. synt_rel_cc:preconj 14. synt_rel_ccomp 15. synt_rel_conj 16. synt_rel_cop 17. synt_rel_det 18. synt_rel_discourse 19. synt_rel_fixed 20. synt_rel_flat 21. synt_rel_goeswith 22. synt_rel_iobj 23. synt_rel_list 24. synt_rel_mark 25. synt_rel_nmod 26. synt_rel_nsubj	float	Частота встречаемости синтаксических связей в тексте: 1. Синтакс. связь: flat:foreign: 2. Синтакс. связь: csubj 3. Синтакс. связь: acl 4. Синтакс. связь: acl:relcl 5. Синтакс. связь: advcl 6. Синтакс. связь: advmod 7. Синтакс. связь: amod 8. Синтакс. связь: appos 9. Синтакс. связь: aux 10. Синтакс. связь: aux:pass 11. Синтакс. связь: case 12. Синтакс. связь: cc 13. Синтакс. связь: cc:preconj 14. Синтакс. связь: ccomp 15. Синтакс. связь: conj 16. Синтакс. связь: cop 17. Синтакс. связь: det 18. Синтакс. связь: discourse 19. Синтакс. связь: fixed 20. Синтакс. связь: flat 21. Синтакс. связь: goeswith 22. Синтакс. связь: iobj 23. Синтакс. связь: list 24. Синтакс. связь: mark 25. Синтакс. связь: nmod 26. Синтакс. связь: nsubj

27. synt_rel_nsubj:pass 28. synt_rel_nummod 29. synt_rel_nummod:gov 30. synt_rel_obj 31. synt_rel_obl 32. synt_rel_orphan 33. synt_rel_parataxis 34. synt_rel_punct 35. synt_rel_root 36. synt_rel_xcomp 37. synt_rel_compound 38. synt_rel_flat:foreign		27. Синтакс. связь: nsubj:pass 28. Синтакс. связь: nummod 29. Синтакс. связь: nummod:gov 30. Синтакс. связь: obj 31. Синтакс. связь: obl 32. Синтакс. связь: orphan 33. Синтакс. связь: parataxis 34. Синтакс. связь: punct 35. Синтакс. связь: root 36. Синтакс. связь: xcomp 37. Синтакс. связь: compound 38. Синтакс. связь: flat:foreign
---	--	---

2.4 Семантические роли при эмотивах

Входные данные

Параметр	Тип	Описание
text	строка	Анализируемый текст, длина от 10 символов
dict	строка	Идентификатор словаря/набора правил

Используемые анализаторы

1. isanlp_rst (возможно использовать обрезанный вариант, только для извлечения элементарных дискурсивных единиц/клауз)
2. sanlp_srl_framebank (необходимо использовать измененную версию, выполняющую только извлечение предикатов-аргументов)
3. udpipe: postag, syntax_dep_tree

Основа

Клаузы, извлеченные через isanlp_rst

Выходные данные

Параметр	Тип	Описание
clauses	список	Список структур, по одной для каждой клаузы в тексте
clauses['clause_text']	строка	Текст клаузы
clauses['parse']	структура	Результат установления ролей внутри клаузы
clauses['parse']['predicate']	строка	слово-предикат + позиция слова в предложении, пример: “купил/4”

<code>clauses['parse']['arguments']</code>	список	Список аргументов для предиката
<code>argument['argument']</code>	строка	слово-аргумент + позиция в предложении
<code>argument['roles']</code>	список строк	Список ролей, может состоять из более чем одного элемента при неоднозначном разборе

2.5 Предсказание депрессивности автора текста по формату эссе

Входные данные

Параметр	Тип	Описание
text	string	Русскоязычный текст размером от 500 до 5000 символов

Основа

Психолингвистические и словарные признаки; классификационная модель, обученная на эссе.

Используемые анализаторы

Название	Ключи
MyStem	lemma, postag

Выходные данные

Параметр	Тип	Описание
prediction	Int	Предсказание модели (0 или 1)

2.6 Предсказание депрессивности автора текста по формату социальных сетей

Входные данные

Параметр	Тип	Описание
text	string	Русскоязычный текст размером от 2000 до 20000 символов

Основа

Психолингвистические и словарные признаки; классификационная модель, обученная на текстовых сообщениях пользователей социальных сетей.

Используемые анализаторы

Название	Ключи
MyStem	lemma, postag

Выходные данные

Параметр	Тип	Описание
prediction	Int	Предсказание модели (0 или 1)

2.7 Предсказание состояния фрустрации у автора текста

Входные данные

Параметр	Тип	Описание
text	string	Русскоязычный текст размером от 2000 до 20000 символов

Основа

Модель BERT; классификационная модель, обученная на текстовых сообщениях пользователей социальных сетей.

Выходные данные

Параметр	Тип	Описание
prediction	Int	Предсказание модели (0 или 1)

2.8 Расчет дискурсивных признаков

Входные данные

Параметр	Тип	Описание
text	string	Непустой текст на русском языке
long_texts_only	bool, default: True	Находить дискурсивные признаки, только если длина текста превышает одно предложение

Основа

Иерархическая дискурсивная аннотация в рамках Теории риторических структур.

Используемые анализаторы

Название	Ключи
MyStem	postag, morph
UDPipe	tokens, sentences, lemma, syntax_dep_tree
IsaNLP RST	rst

Выходные данные

Параметр	Тип	Описание
number_of_units	int	Количество дискурсивных единиц
number_of_trees	int	Количество дискурсивных деревьев
first_tree_edus_number	int	Количество элементарных дискурсивных единиц, входящих в первое дерево
last_tree_edus_number	int	Количество элементарных дискурсивных единиц, входящих в последнее дерево
suspended_edus_number	int	Количество элементарных дискурсивных единиц, не входящих ни в одно дерево
average_du_length	float	Средняя длина дискурсивной единицы в словах
average_edu_length	float	Средняя длина элементарной дискурсивной единицы в словах
average_tree_depth	float	Средняя глубина дискурсивного дерева
symmetric_relations	float	Доля мультаядерных (NN) риторических отношений
first_word_uniqueness	float	Доля уникальных первых слов элементарной ДЕ
du_len_imbalance	float	Среднее отношение длины правой части отношения к длине левой части
simple_sentence_ratio	float	Доля простых предложений среди ЭДЕ
Параметры - наименования риторических отношений: 1. attribution_NS 2. attribution_SN 3. background_NS 4. cause-effect_NS 5. cause-effect_SN 6. comparison_NN 7. concession_NS 8. condition_NS 9. condition_SN 10. contrast_NN 11. elaboration_NS 12. evidence_NS 13. interpretation-evaluation_NS 14. interpretation-evaluation_SN 15. joint_NN 16. preparation_SN 17. purpose_NS 18. purpose_SN	int	Количество вхождений определенных риторических отношений: 1. Источник информации с ядром слева 2. Источник информации с ядром справа 3. Фон 4. Причина с ядром слева 5. Причина с ядром справа 6. Сравнение 7. Уступка 8. Условие с ядром слева 9. Условие с ядром справа 10. Контраст 11. Детализация 12. Обоснование 13. Интерпретация/оценка с ядром слева 14. Интерпретация/оценка с ядром справа 15. Конъюнкция 16. Подготовка 17. Цель с ядром слева 18. Цель с ядром справа

19. restatement_NN 20. same-unit_NN 21. sequence_NN 22. solutionhood_SN		19. Переформулировка 20. Прерывающаяся единица 21. Последовательность 22. Решение
--	--	--

2.9 Выявление типа реакции на фрустрацию по Розенцвейгу

Входные данные

Параметр	Тип	Описание
text	string	Непустой текст на русском языке
postag	Список строк	Список частей речи
srl	Список объектов	Результаты семантического анализа в виде списка объектов класса Event ()
morph	Список слов	Морфологические признаки слов
lemma	Список строк	Нормальные формы слов анализируемого текста
syntax_dep_tree	Список объектов	Синтаксическое дерево в виде списка объектов класса WordSynt.

Используемые анализаторы

Название	Ключи
MyStem (isa-nlp)	postag, morph
Udpipe (isanlp)	tokens, sentences, lemma, syntax_dep_tree
IsaNLP SRL FrameBank	srl

Основа

Типология реакций на фрустрацию по-Розенцвейгу.

Выходные данные

Параметр	Тип	Описание
type	string	Тип реакции на фрустрацию. Строка из множества: {E, E', e, M, M', m, I, I', i} E - агрессия, E' - растерянность и неприятие, e - требование к другим исправить ситуацию, M - отрицание чьей-либо вины, M' - игнорирование, отрицание значимости препятствия m - ожидание, что само всё как-то образуется, I - самобвинение I' - признание ситуации в чём-то полезной, i - "я сам исправлю ситуацию"

3. Сторонние компоненты

Компонент	Лицензия и ссылка
MyStem	https://yandex.ru/legal/mystem/ “3.2. Программа может использоваться в коммерческих целях для разработки/создания каких-либо сервисов или программ, включаться и использоваться по прямому функциональному назначению в составе таких сервисов или программ, а также использоваться иным образом в процессе оказания услуг/выполнения работ, за исключением: <ul style="list-style-type: none">- использование Программы и/или сервисов или программ, созданных на её основе или с её использованием, для создания или распространения массовых рассылок и спама;- использование Программы и/или сервисов или программ, созданных на её основе или с её использованием, для поисковой оптимизации сайтов в сети Интернет;- использование Программы для создания сервисов или программ или в составе сервисов или программ, предлагающих услуги или функциональность, аналогичную программам и сервисам Правообладателя.”
Udpipe	Mozilla Public License Version 2.0 https://raw.githubusercontent.com/ufal/udpipe/master/LICENSE
isanlp	MIT License https://raw.githubusercontent.com/IINemo/isanlp/master/LICENSE
isanlp_srl_framebank	MIT License https://raw.githubusercontent.com/IINemo/isanlp_srl_framebank/master/LICENSE

isanlp_rst	MIT License https://raw.githubusercontent.com/tchewik/isanlp_rst/master/LICENSE
linis-crowd sentiment	https://creativecommons.org/licenses/by-nc-sa/4.0/ Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

4. Минимальные требования к оборудованию

- 16 ГБ оперативной памяти
- 8 ядер ЦП

5. Минимальные системные требования

- python 3.7
- docker 19.03.0
- docker-compose 1.25