Команда: “НеИИ”
Кейс: “Университет 20.35”

Необходимость прохождения множества тестов для определения характеристик учащихся

Пройти один тест ~10-20 минут

Анализ стиля и тона текстовых данных для автоматической прикидки характеристик

Оставить комментарий о пройденном обучении - 5 минут

Подготовка данных

- ★ Заполнение пропусков в данных диагностики
- ★ Агрегация результатов диагностики по каждому юзеру
- ★ Выделение признаков из текста
- ★ Агрегация текстовых признаков по юзеру
- ★ Объединение датасетов
- ★ Разбиение на трейн (db0-7) и тест (db8-10)

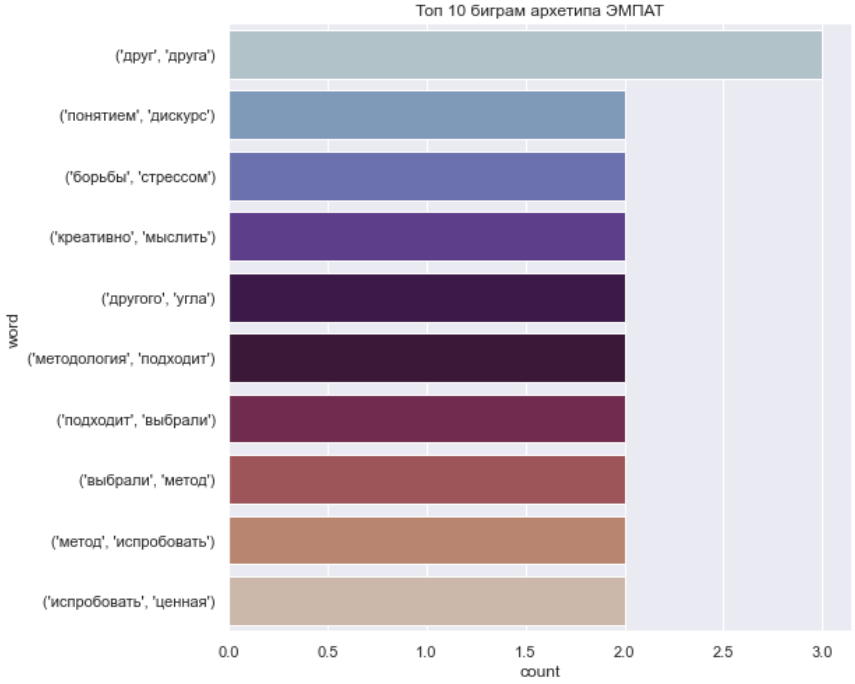
Выделение признаков из текста

- ★ Сентимент (deeppavlov, ELMO обученная на русском Twitter):
  - Классы: негатив, нейтралитет, позитив, вежливость, бессмыслица
- ★ Наличие NER (deeppavlov):
  - Присутствие в тексте именованных сущностей (названия организаций, фреймворков, людей).
  - Среднее кол-во используемых NER по юзеру.
- ★ Характеристики текста:
  - Среднее количество сообщений с позитивной тональностью
  - Среднее количество сообщений с негативной тональностью
  - Среднее количество сообщений с нейтральной тональностью
  - Среднее количество сообщений, содержащих вежливые обороты
  - Среднее количество используемых NER (именованных сущностей)
  - Средняя длина сообщений
  - Среднее количество слов в заглавной буквы
  - Среднее количество знаков препинания в сообщении

Дополнительные характеристики текстов

- ★ Тематическое моделирование (LDA в gensim). НЕ ЗАШЛО
  - Определение наиболее частой темы комментариев.
  - Использование её в качестве признака при прогнозе показателей диагностик
- ★ Векторизация текстов рефлексий для прогноза класса “Архетип”. ЗАШЛО
  - Гипотеза: частота использования определенных слов коррелирует с определенным типом личности
- ★ N grams (словосочетания). ЗАШЛО

Анализ N-grams (словосочетания из N эл-тов)



★ «Эмпат»: Стремится к гармонии, миру, отсутствию конфликта. Позитивно настроен, всегда сопереживает, готов выслушать и помочь. Любит все живое.

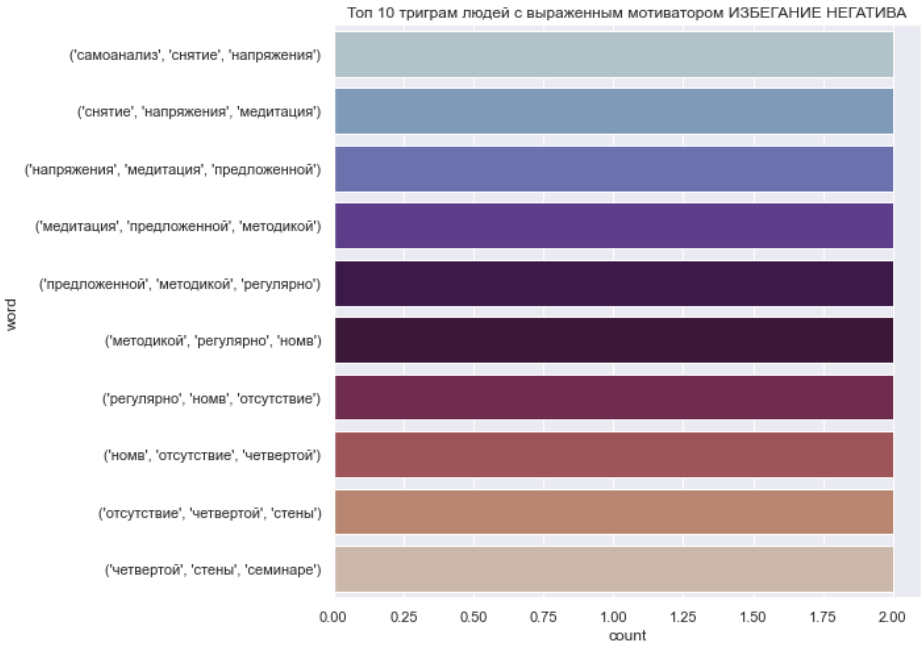
★ В топ биграмм по текстам людей с выраженным архетипом «Эмпат» попали:

- друг друга
- борьба (со) стрессом
- креативно мыслить

Перекликаются по смыслу с описанием архетипа

Подробный разбор N grams по остальным диагностикам: https://git.2035.university/maia-06/project2035/-/blob/master/4_Analysing_N_Grams.ipynb

Анализ N-grams (словосочетания из N эл-тов)



★ **Мотиватор «Избегание негатива»:** люди с выраженным мотиватором избегания негатива будут всеми силами стараться уйти от конфликтных и противоречивых ситуаций.

★ Как видим, среди слов, используемых людьми с выраженным показателем данного мотиватора выделяются:

- снятие напряжения
- медитация
- самоанализ

Подробный разбор N grams по остальным диагностикам: https://git.2035.university/maia-06/project2035/-/blob/master/4_Analysing_N_Grams.ipynb

Корреляции

- ★ Явные корреляции наблюдаются между некоторыми признаками из диагностики. Например, показатели **"Мотивация: достижение, стремление к лидерству"** имеют положительную корреляцию (0.7) с **"Архетипом аналитик"**.

- ★ Также признаки, извлеченные из текстов, коррелируют между собой. Например, есть положительная корреляция между **"Длиной сообщения"** и **"Количеством знаков препинания"** (0.77).

- ★ Корреляций между выделенными текстовыми признаками и значениями диагностик меньше.

Подробнее по корреляциям: https://git.2035.university/maia-06/project2035/-/blob/master/5_Correlations.ipynb

Поиск корреляций между показателями тестов “Архетипы” и “Модель культуры организационной деятельности”

хакер_sum	исследование неопределённости_sum	0.777857
боец_sum	гибкий график, удалённая работа_sum	0.751433
гибкий график, удалённая работа_sum	творец_sum	0.739696
творец_sum	открытая система управления знаниями_sum	0.705023

★ Возможность прогнозирования Архетипа без прохождения теста

ML модель #1: Прогноз значения мотиватора

Мотиватор	**MAE алгоритма**	**MAE константа=1**
Избегание негатива, размеренность, безопасность	**1.6**	2.3
Преодоление сложностей, препятствий	**2.2**	4.4
Осмысленность деятельности, собственная значимость	**1**	1.9
Социальный элемент, дружба, влияние, конкуренция	**1.3**	2.6
Самосовершенствование, работа над собой, труд	**2.2**	2.8
Достижение, стремление к лидерству	**1.7**	2.8

Градиентный бустинг

+ Находит нелинейные зависимости
+ Не требует вычислительных мощностей в отличии от нейронных сетей

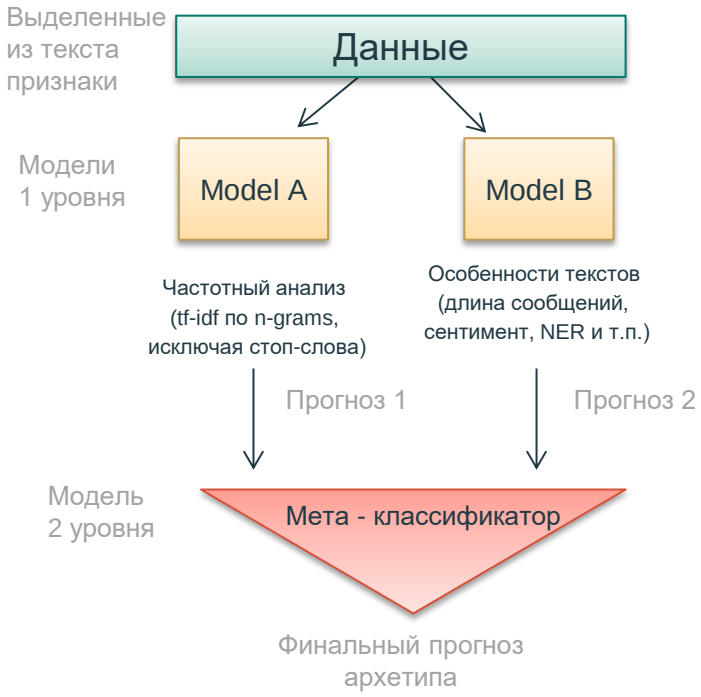
dmlc XGBoost

Подробнее (графики + feature importance) - https://git.2035.university/maia-06/project2035/-/blob/master/7_Model_Regressor.ipynb

ML модель #2: Прогноз архетипа

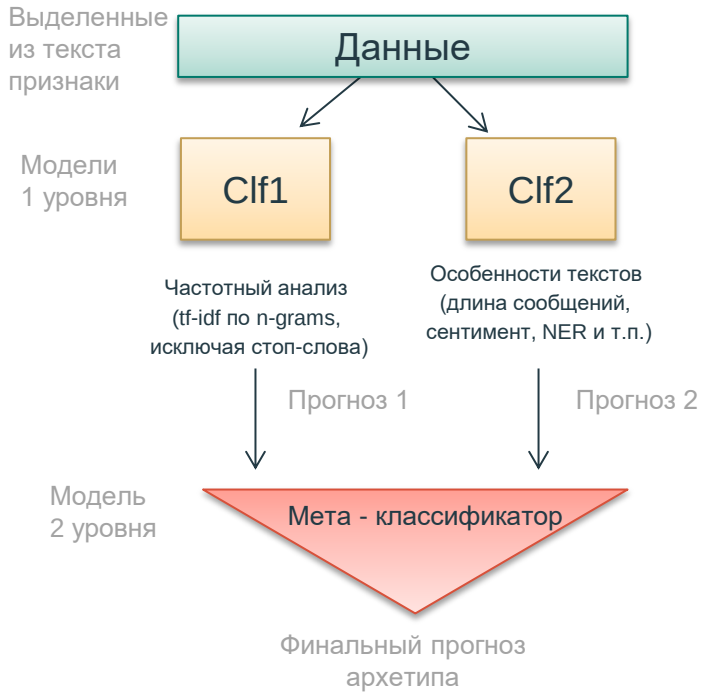
Двухуровневая модель (stacking)

- **Модель А уровня 1** прогнозирует архетип, исходя из семантики (tf idf по N граммам, исключая часто встречающиеся слова, т.н. Стоп-слова). Алгоритм – kNearestNeighbors.
- **Модель В уровня 1** прогнозирует архетип, исходя из тональности/сентимента текста (положительный, нейтральный, негативный, вежливый), количества именованных сущностей и статистических признаков: длина сообщения, кол-во знаков препинания и т.п. Алгоритм – kNearestNeighbors.
- **Мета-классификатор** - прогнозирует архетип на основе прогнозов Модели А и Модели В. Алгоритм – градиентный бустинг (LightGBM).



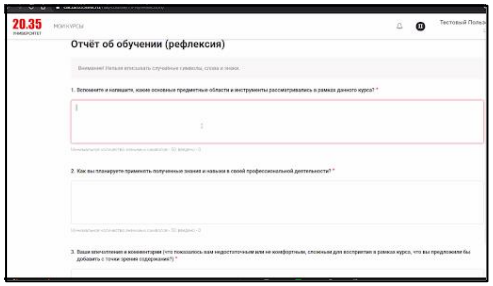
ML модель #2: Прогноз архетипа

Архетип	Точность (precision)
Боец	90 %
Аналитик	74.7 %
Исполнитель	72.7%
Эксперт	72.4%
Эмпат	71.4%
Визионер	71%



- ★ Средняя точность алгоритма ~ 75%.
- ★ Средняя точность константного прогноза модой - 0%

MLaaS: Персональные рекомендации на основе автоматического определения архетипа



Текст 1
Текст 2
Текст ...

NLP сервис

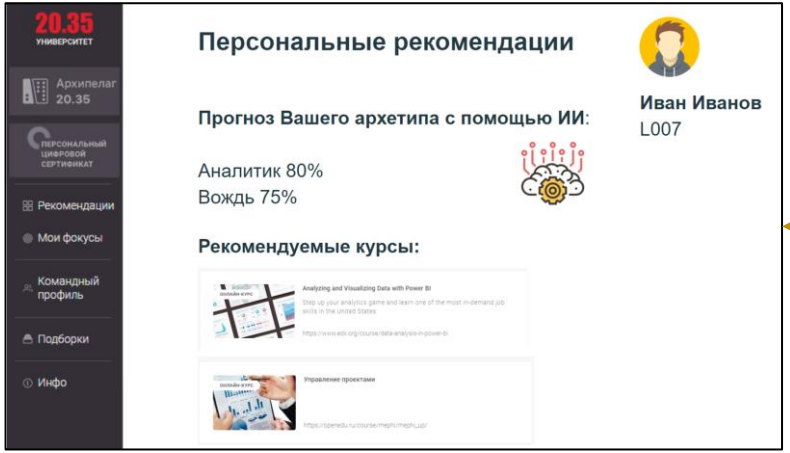
Feature Extraction

ner	answer_len	upper_case_word_count	punctuation_count
0.583333	1.75	0.027778	0.055556

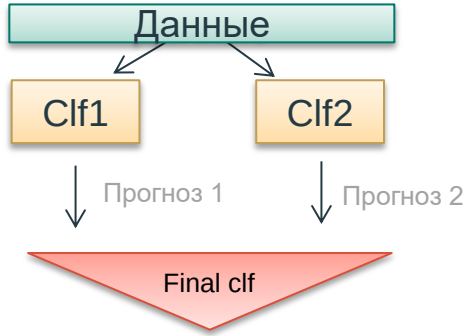
noun_count	verb_count	adj_count	adv_count	pron_count
81	58	25	16	36

sentiment_negative	sentiment_neutral	sentiment_positive
0.0	0.972222	0.027778

char_count	word_count	word_density
1785	314	5.666667



Machine Learning



Финальный прогноз архетипа

20.35
УНИВЕРСИТЕТ

Архипелаг 20.35

ПЕРСОНАЛЬНЫЙ ЦИФРОВОЙ СЕРТИФИКАТ

- Рекомендации
- Мои фокусы
- Командный профиль
- Подборки
- Инфо

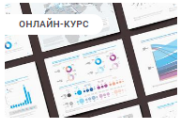
Персональные рекомендации

Иван Иванов
L007

Прогноз Вашего архетипа с помощью ИИ:

Аналитик 80%
Вождь 75%

Рекомендуемые курсы:



Analyzing and Visualizing Data with Power BI

Step up your analytics game and learn one of the most in-demand job skills in the United States.

https://www.edx.org/course/data-analysis-in-power-bi

Демо работы NLP сервиса - https://git.2035.university/maia-06/project2035/-/blob/master/nlp_service/demo.mp4
Анализ реализации проекта по критериям оценки - https://git.2035.university/maia-06/project2035/-/blob/master/CRITERIA.md

Используемые технологии

MLaaS

Feature Extraction

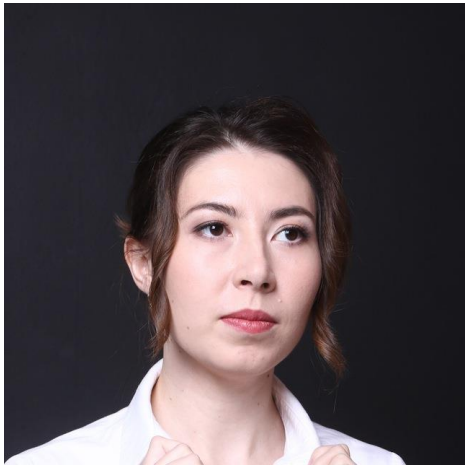
NLTK

Морфологический анализатор pymorphy2

DeepPavlov

Machine Learning

Наша команда



Марина Семенова

Data Engineer

semenovamarina1992@gmail.com

Майя Бикметова

Data Scientist

maya.bikmetova@gmail.com