

Кейс № 4 “Викторина Клевер”



Команда: *BigChungus*
Лукьянов А.Е.
Криворучко Д.А.

	ID	Question
0	1	Как зовут лодочника на реке Стикс в древнегреч...
1	2	Как в химии обозначается свинец?
2	3	Какой химический элемент преобладает в составе...
3	4	Кто из перечисленных был пажом во времена Екат...
4	5	Когда началась 2 мировая война?
...
41082	41083	В каком году распался СССР
41083	41084	Сколько калорий в 100 гр арбуза?
41084	41085	Сколько хвостов у лиса, который является демон...
41085	41086	Сколько раз магнитогорский металлург становилс...
41086	41087	Какая численность людей в 2018 году?

41087 rows × 2 columns

	ID	Answer
0	1	0
1	2	1
2	3	0
3	4	0
4	5	0
...
29995	29996	0
29996	29997	0
29997	29998	0
29998	29999	0
29999	30000	0

30000 rows × 2 columns

Зависимости и закономерности в данных, возможные пути решения











- ❑ В вопросах экспертов часто встречается НЕ, НЕТ;
- ❑ Общее количество слов в вопросах экспертов - 26878;
- ❑ Общее количество слов в вопросах не экспертов - 226504;
- ❑ Средняя длина предложений у экспертов и не экспертов \approx одинакова;
- ❑ Значения TF и IDF у экспертов в среднем больше чем у не экспертов.

Что сделано?

- ❑ Очистка и предварительный анализ данных;
- ❑ MLP - 56%.
- ❑ SVM на сыром тексте - 58%
- ❑ LSTM - 63%

Эталонное решение

Ссылка: <https://github.com/BraginIvan/vkcup2019>

	BraginIvan Create tfliite_inference.py	c3b9acf on 9 Mar 2020	 11 commits
	analysis	final 0.8875	16 months ago
	.gitignore	final 0.8875	16 months ago
	EDA.ipynb	first commit	16 months ago
	README.md	Update README.md	16 months ago
	stage2.ipynb	stage 2 https://mlbootcamp.ru/round/25/rating/	13 months ago
	stage2_boosting.ipynb	stage 2 boosting https://mlbootcamp.ru/round/25/rating/	13 months ago
	tfliite_inference.py	Create tfliite_inference.py	13 months ago
	training2.ipynb	show feature importance	16 months ago

Что дальше?