



RUATD



ИРА ЛОБАНОВА



НАСТЯ ШАБАЕВА



Мотивация



- Программа, распознающая сгенерированные тексты, необходима для защиты пользователей от получения фейкового и мошеннического контента.
- Такая программа также может быть полезна для разработчиков генераторов текста, поскольку она будет способствовать повышению их качества (имеет смысл делать генераторы такими, чтобы их текст программа посчитала естественным, подобрать оптимальную модель).

Данные



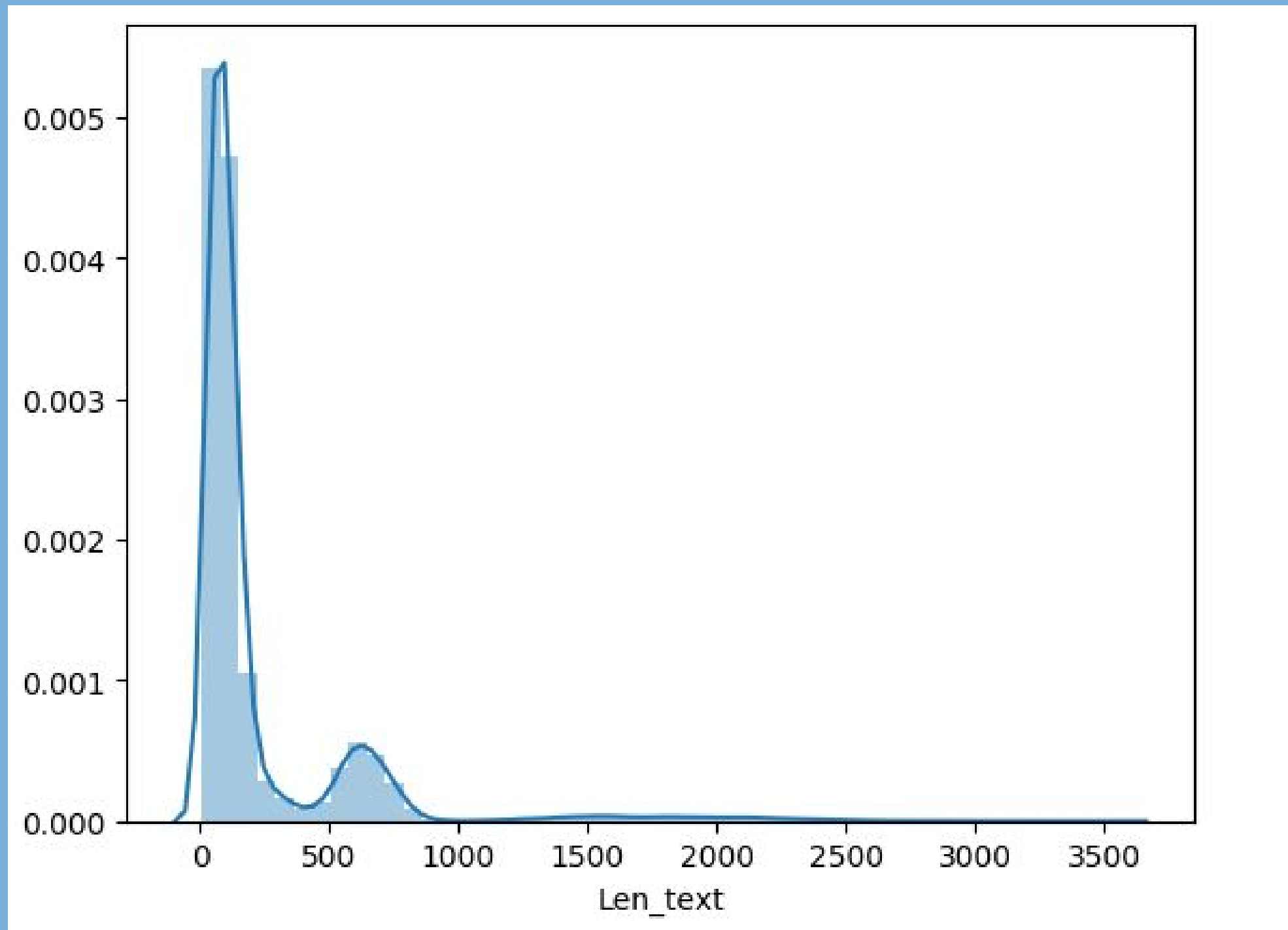
Бинарная:

- 3 датасета: для обучения, для валидации и для предсказаний
- Train: 129065 текстов, val: 21511 текстов, test: 64533 текстов.
- Классы: М (машина), Н (человек)
- Средняя длина текстов: 229 символов.
- Классы в обучающих данных распределены равномерно.

Данные



График распределения длины текстов в датасетах

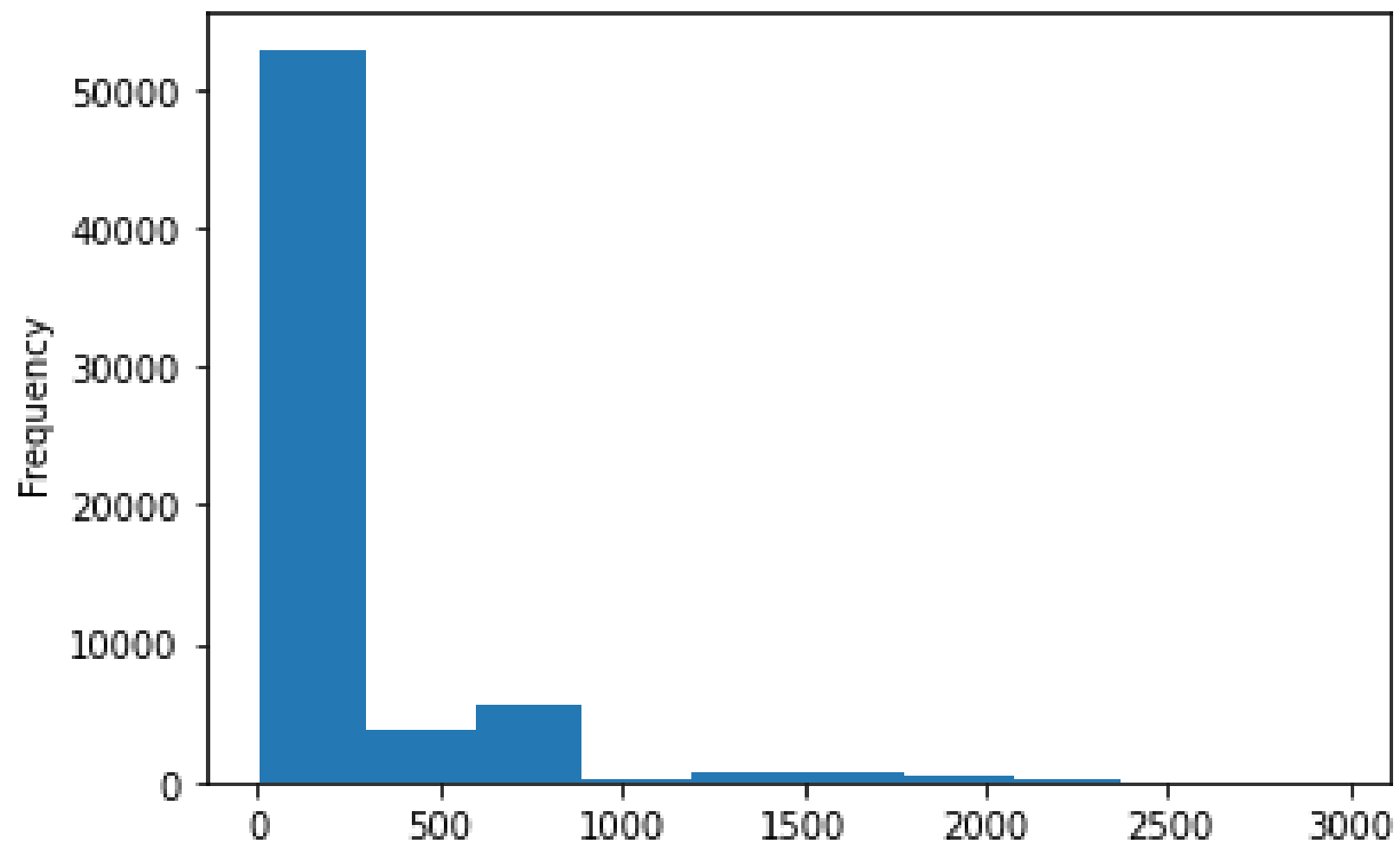


Данные



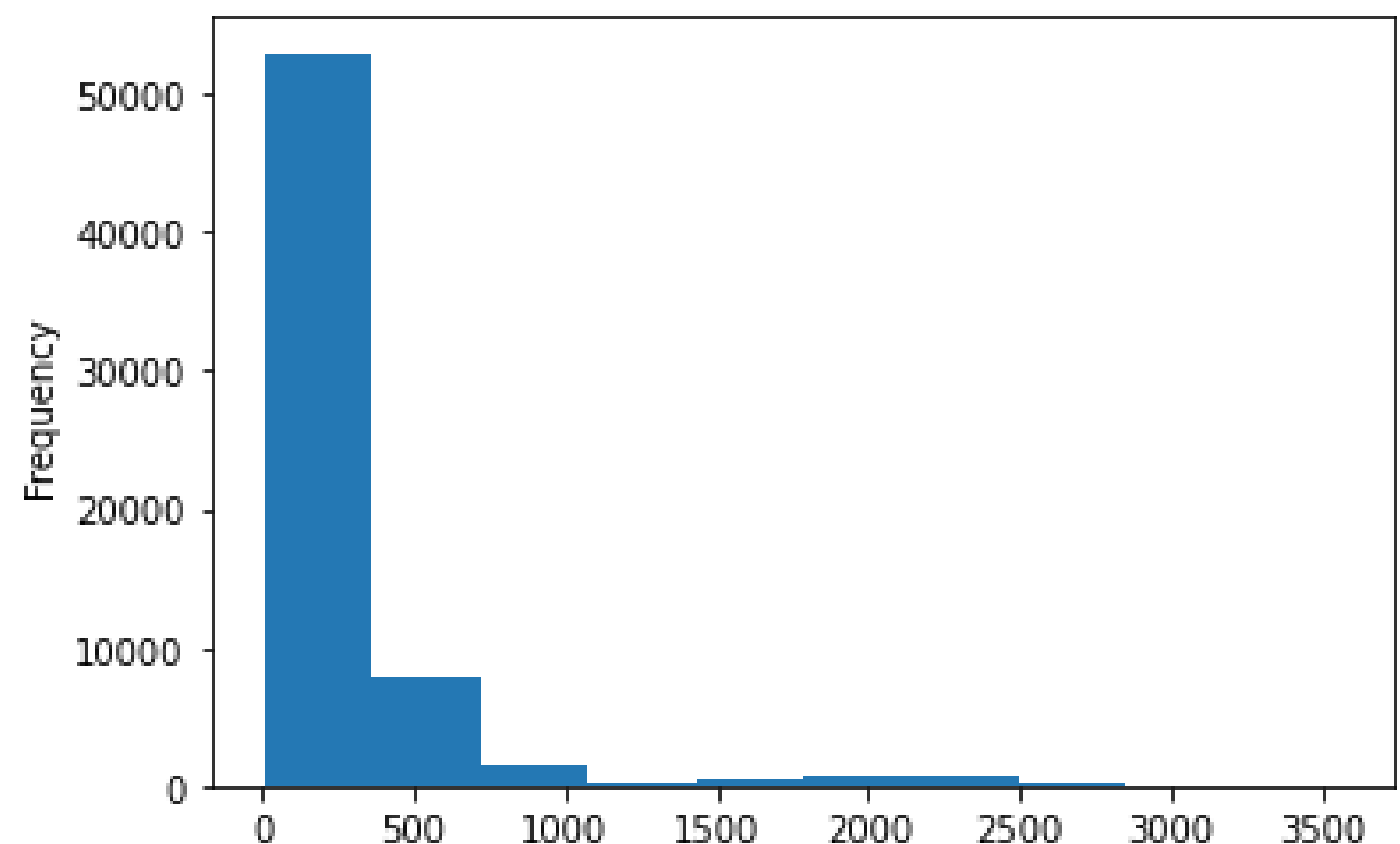
Естественные тексты:

Средняя длина: 221 символ



Сгенерированные тексты:

Средняя длина: 237 символов



Данные



Использованные в текстах языки (по версии langdetect)

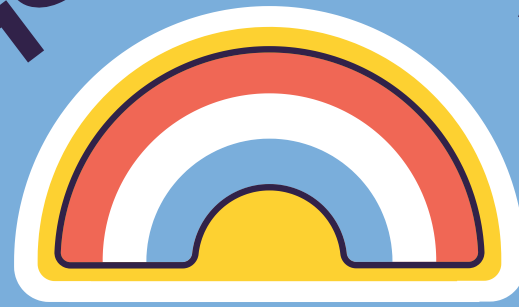
Естественные тексты (28 штук):

'ru', 'ro', 'bg', 'uk', 'mk', 'de',
'Nan', 'et', 'fr', 'da', 'en', 'es', 'pl',
'af', 'ca', 'nl', 'so', 'it', 'pt', 'cy', 'sv',
'vi', 'fi', 'sk', 'el', 'no', 'tl', 'sw'

Сгенерированные тексты (18 штук):

'ru', 'uk', 'bg', 'en', 'mk', 'fr',
'ca', 'et', 'it', 'Nan', 'de', 'so',
'pt', 'lt', 'sl', 'nl', 'sv', 'cy'

Данные



Использованные в текстах символы

Естественные тексты:

- Всего 818 символов
- Среди них 393 символа, которые есть в естественных текстах и которых нет в сгенерированных
- Например: 'ה', 'ב', 'צ', 'א', 'ת', 'ק', '鬼', '灯', '\u2004', '兴', '县', '专', '榆', '次', '离', '山', '晋', '地', '璧', '瑗'

Сгенерированные тексты:

- Всего 802 символа
- Среди них 377 - символы, которые есть в сгенерированных текстах и которых нет в естественных
- Например: '❖', '✂', '▶', '☎', '◆', '°C', '␣', '回', '義', '▼', 'ï', '\u200a', ' ', '●', '\x97', '★', ' ', '\u200d', '___', ' ', '✓'

Данные



20 самых частотных слов

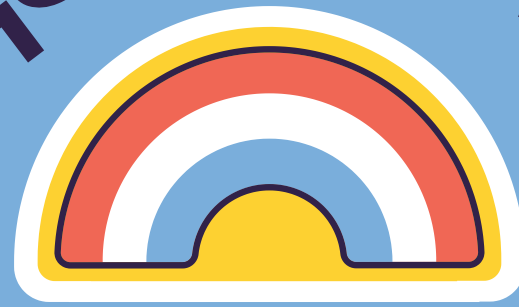
Естественные тексты:

('год', 11109), ('это', 4389), ('который', 3701),
('муниципальный', 3198), ('весь', 3165),
('район', 2494), ('свой', 2480), ('область', 2352),
('человек', 2162), ('программа', 2130),
('российский', 1915), ('также', 1906),
('развитие', 1761), ('время', 1672), ('мочь',
1662), ('город', 1526), ('государственный',
1429), ('реализация', 1421), ('№', 1404),
('образование', 1372)

Сгенерированные тексты:

('год', 8161), ('это', 7342), ('весь', 4726),
('который', 4224), ('свой', 3587),
('человек', 3335), ('мочь', 2670),
('россия', 2607), ('время', 2574), ('-', 2548),
('также', 2406), ('российский', 2167),
('город', 2000), ('1', 1953),
('программа', 1823), ('страна', 1692),
('работа', 1678), ('развитие', 1676),
('день', 1519), ('стать', 1493)

Данные



Частотные биграммы

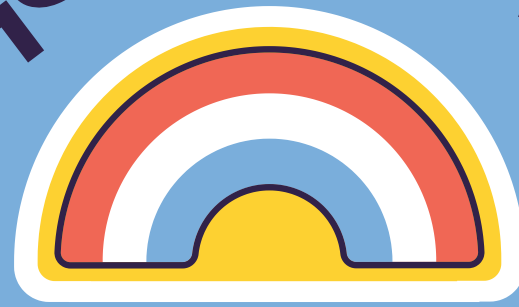
Естественные тексты:

[('российский', 'федерация'), 1255],
[('муниципальный', 'образование'), 762],
[('муниципальный', 'программа'), 734],
[('муниципальный', 'район'), 729],
[('домашний', 'хозяйство'), 578],
[('городской', 'округ'), 390],
[('федеральный', 'закон'), 384],
[('местный', 'самоуправление'), 380],
[('2020', 'год'), 359],
[('социальноэкономический', 'развитие'), 318]]

Сгенерированные тексты:

[('российский', 'федерация'), 1020],
[('настоящий', 'время'), 455],
[('муниципальный', 'образование'), 382],
[('2015', 'год'), 326],
[('муниципальный', 'программа'), 302],
[('2014', 'год'), 270],
[('оригинал', 'запись'), 265],
[('федеральный', 'закон'), 265],
[('-', 'это'), 256],
[('это', 'год'), 247]]

Данные



Частотные триграммы

Естественные тексты:

('орган', 'местный', 'самоуправление'), 228),
(('реализация', 'муниципальный', 'программа'),
188), (('малое', 'среднее', 'предпринимательство'),
185), (('правительство', 'российский', 'федерация'),
171), (('год', 'плановый', 'период'), 139),
(('домашний', 'хозяйство', 'представлять'), 136),
(('домашний', 'хозяйство', 'состоять'), 136),
(('кодекс', 'российский', 'федерация'), 127),
(('физический', 'культура', 'спорт'), 123),
(('президент', 'российский', 'федерация'), 122),
(('общий', 'принцип', 'организация'), 122)

Сгенерированные тексты:

('орган', 'местный', 'самоуправление'), 159),
(('субъект', 'российский', 'федерация'), 131),
(('орган', 'исполнительный', 'власть'), 121),
(('великий', 'отечественный', 'война'), 112),
(('орган', 'государственный', 'власть'), 96),
(('правительство', 'российский', 'федерация'), 85),
(('второй', 'мировой', 'война'), 79),
(('муниципальный', 'образование', 'город'), 75),
(('президент', 'российский', 'федерация'), 72),
(('реализация', 'муниципальный', 'программа'), 71)










Что мы планировали сделать изначально


- Использовать бейзлайн tf-idf с логистической регрессией и написать свою нейросеть, попробовав разные архитектуры;
- Это оказалось нерелевантно:(


Что мы сделали к предзащите



- Мы решили воспользоваться ruBert (это был наш запасной вариант), попробовали бейзлайны;
- Нашли новую литературу;
- Решили использовать ruBert-tiny;
- ruBert-tiny работает значительно быстрее, но качество на модели, аналогичной бейзлайну с ruBert хуже (однако не сильно: accuracy бейзлайна - 0.79622; ruBert-tiny - 0.75574)

	Baseline Bert		0.79622		
10	Tumanov Alexander		0.79343	4	10d
11	Elizaveta Nosova		0.79278	4	10d
12	Nikita Selin		0.78689	22	20d
13	David Avagyan		0.78026	1	5d
14	Mikhail Yumanov		0.77294	11	3d
15	mipatov		0.76579	3	3d

16	Анастасия Шабеева		0.75574	2	2h
----	--------------------------	---	---------	---	----



Your Best Entry!
 Your most recent submission scored 0.75574, which is an improvement of your previous score of 0.74747. Great job!

Tweet this

Что мы делали дальше



- Сделали более подробный анализ данных (посмотрели на символы, языки, самые частотные слова, биграммы и триграммы)
- Попробовали добавить в ruBert от DeepPavlov EarlyStoppingCallback - раннюю остановку, если спустя заданное количество шагов обучения метрика (loss или accuracy) не улучшается

Что мы делали дальше



- Запустили tf-idf+логистическую регрессию, результат получился даже значительно хуже, чем у Bert-tiny (0,65).
- Попробовали сделать стекинг (метод создания ансамбля) для tf-idf+log.reg и Bert.

Литература

1. Learning without Forgetting

<https://arxiv.org/pdf/1606.09282.pdf>

2. Маленький и быстрый BERT для русского языка

<https://habr.com/ru/post/562064/>

3. How to Fine-Tune BERT for Text Classification?

<https://arxiv.org/pdf/1905.05583.pdf>