

Дмитрий Кузюрин

Решение задачи

Принятие законопроектов

RuCode 4.0, ноябрь 2021

<https://www.kaggle.com/c/regulations-outcome>



dkuzyurin@gmail.com



@dkuzyurin



Название законопроекта

Токенизация – Стоп-слова

Лемматизация и стемминг

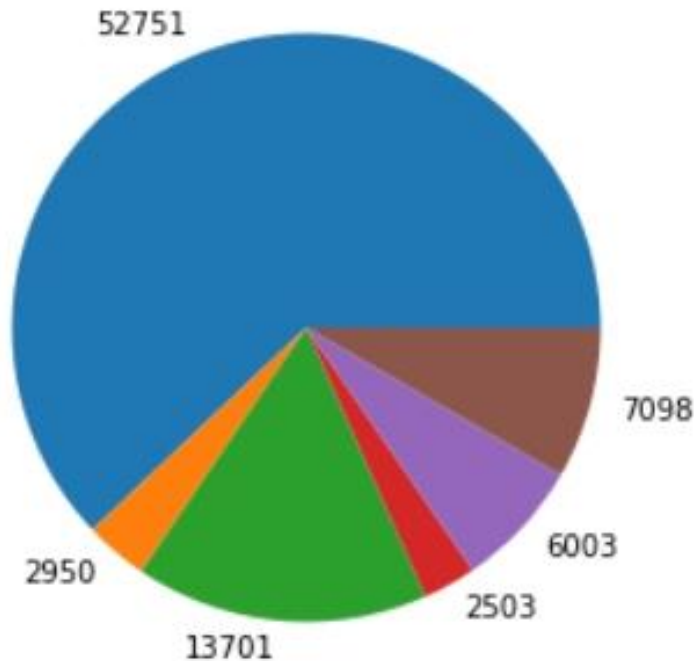
`nlk.stem.snowball.SnowballStemmer`

	act_title	act_title_clear	act_title_lem
0	Об утверждении тарифов на услуги по транспорти...	утверждении тарифов услуги транспортировке газ...	утвержден тариф услуг транспортировк газ газор...
1	О внесении изменений в отдельные законодательн...	внесении изменений отдельные законодательные а...	внесен изменен отдельн законодательн акт росси...
2	Об утверждении Положения об уведомлении лиц об...	утверждении положения уведомлении лиц включени...	утвержден положен уведомлен лиц включен список...
3	О внесении изменений в Положение о Министерств...	внесении изменений положение министерстве обра...	внесен изменен положен министерств образован н...
4	О внесении изменений в Правила подготовки и пр...	внесении изменений правила подготовки принятия...	внесен изменен прав подготовк принят решен пре...
5	О внесении изменений в федеральную целевую про...	внесении изменений федеральную целевую програм...	внесен изменен федеральн целев программ жилищ ...
6	О внесении изменений в Реестр должностей федер...	внесении изменений реестр должностей федеральн...	внесен изменен реестр должност федеральн госуд...
7	О создании на территории Наримановского муници...	создании территории наримановского муниципальн...	создан территор наримановск муниципальн район ...
8	О порядке представления информации об исполнит...	порядке представления информации исполнителях ...	порядк представлен информац исполнитель созда э...
9	Об утверждении федерального государственного о...	утверждении федерального государственного обра...	утвержден федеральн государственн образовательн...

Название законопроекта

Кластеризация на 6 кластеров

ngram_range=(1, 2)



организац, росс, прав, порядк, российск федерац, федерац, российск, государствен, федеральн, утвержден

федеральн государствен, подготовк, стандарт, образовательн, стандарт высш, государствен образовательн, образовательн стандарт, образован, высш образован, высш

изменен приказ, внесен, внесен изменен, изменен, приказ министерств, приказ, министерств, российск, российск федерац, федерац

стандарт работник, специалист област, производств, утвержден, специалист, стандарт специалист, стандарт, профессиональн, утвержден профессиональн, профессиональн стандарт

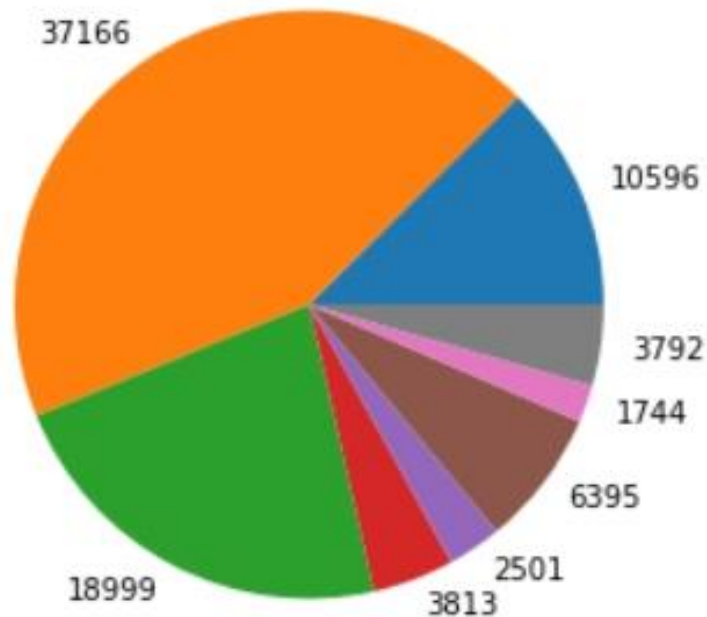
внесен, внесен изменен, российск, российск федерац, федерац, изменен постановлен, постановлен, постановлен правительств, правительств российск, правительств

российск федерац, изменен стат, кодекс российск, кодекс, стат, изменен, внесен, внесен изменен, федеральн закон, закон

Название законопроекта

Кластеризация на 8 кластеров

ngram_range=(1, 2)



надзор, гражданск, агентств, утвержден, государствен гражданск, федеральн государствен, федеральн служб, государствен, служб, федеральн

приказ, внесен, проект, изменен, федеральн, организац, прав, порядок, росс, утвержден

утвержден, приказ министерств, приказ, внесен изменен, внесен, изменен, министерств, российск федерац, федерац, российск

внесен, внесен изменен, российск, российск федерац, федерац, правительств российск, правительств, изменен постановлен, постановлен, постановлен правительств

стандарт работник, специалист област, производств, утвержден, специалист, стандарт специалист, стандарт, профессиональн, утвержден профессиональн, профессиональн стандарт

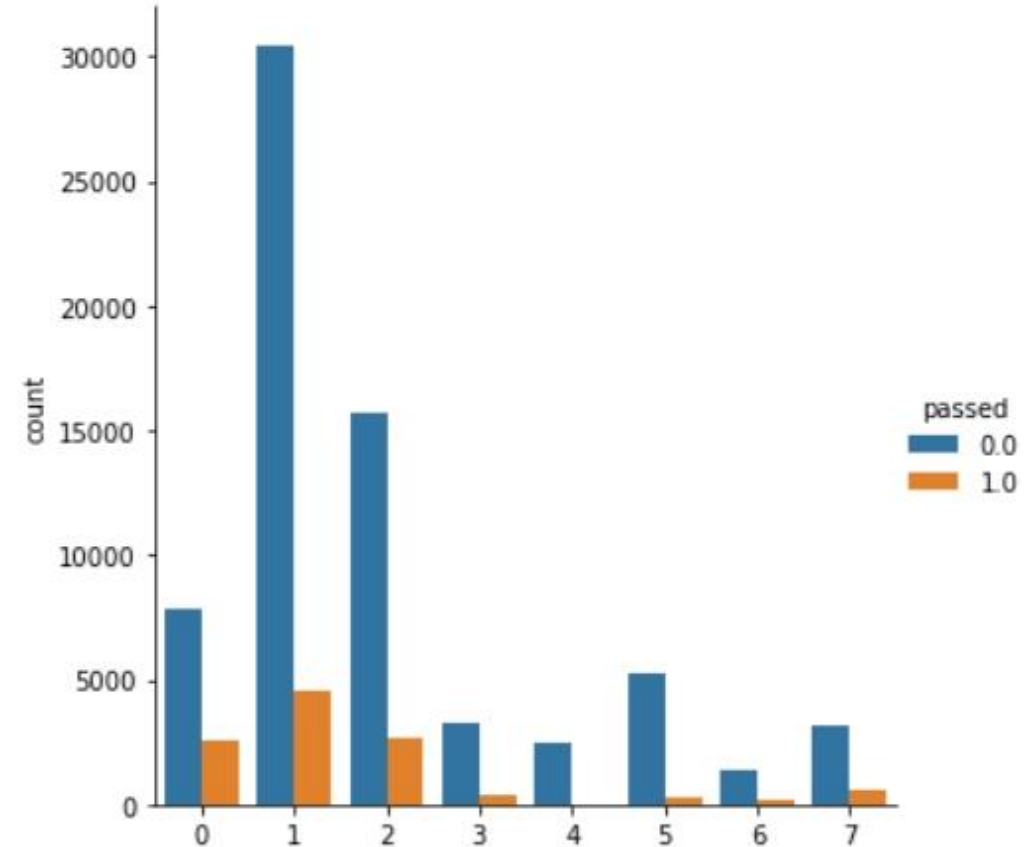
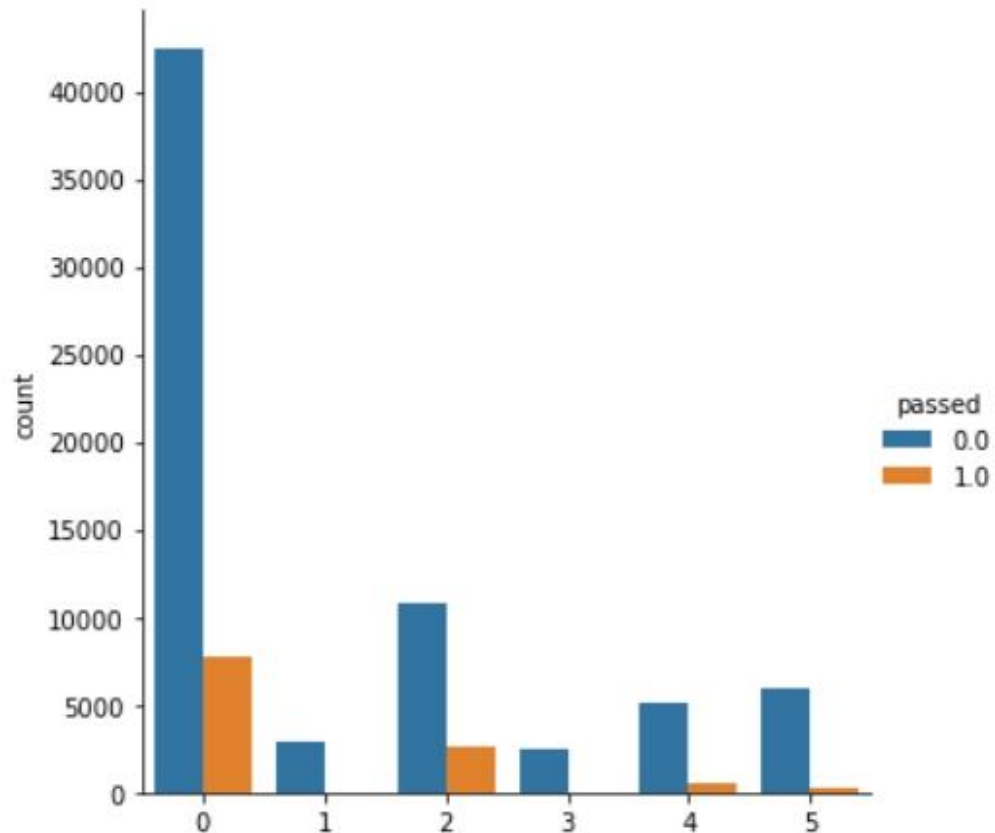
российск федерац, изменен стат, кодекс российск, кодекс, стат, изменен, внесен, внесен изменен, федеральн закон, закон

российск, российск федерац, федерац, правительств российск, правительств, акт, изменен некотор, некотор, акт правительств, некотор акт

утвержден федеральн, федеральн государствен, стандарт высш, стандарт, образовательн, высш образован, государствен образовательн, образовательн стандарт, высш, образован

Название законопроекта

Принадлежность к кластеру и целевой признак



Список ОКВЭД


Максимальное количество указанных ОКВЭД у одного законопроекта: 53

Всего различных упоминаемых ОКВЭД: 57

	0	1	2	3	4	5	6	7	8	9	...	47	48	49	50	51	52	53	54	55	56
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...
85001	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
85002	0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
85003	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
85004	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
85005	0	0	0	0	0	0	0	0	0	0	1	...	0	1	0	0	0	0	0	0	0

Признаки из regulations.csv

+ publication_date  publication_year, publication_month, publication_day

+ added_by, responsible  added_resp
Ответственный за законопроект – тот, кто его внес

+ na_counts
Количество пропусков в данных по законопроекту

— Удалены: problem_addressed, act_significance

= Остальные признаки: очистка и Label Encoding

imp_dict = {'Низкая': 1, 'Средняя': 2, 'Высокая': 3, 'Не определена': 0}

Признаки из ria_reports_main.csv

Cramér's V в качестве меры ассоциации между категориальными переменными и целевым признаком

degree: Степень регулирующего воздействия проекта акта **0.398**

other_notification_info: Источники данных **0.338**

public_discussion: Иные сведения о проведении публичного обсуждения проекта акта **0.331**

public_discussion: Сведения о лицах, представивших предложения **0.323**

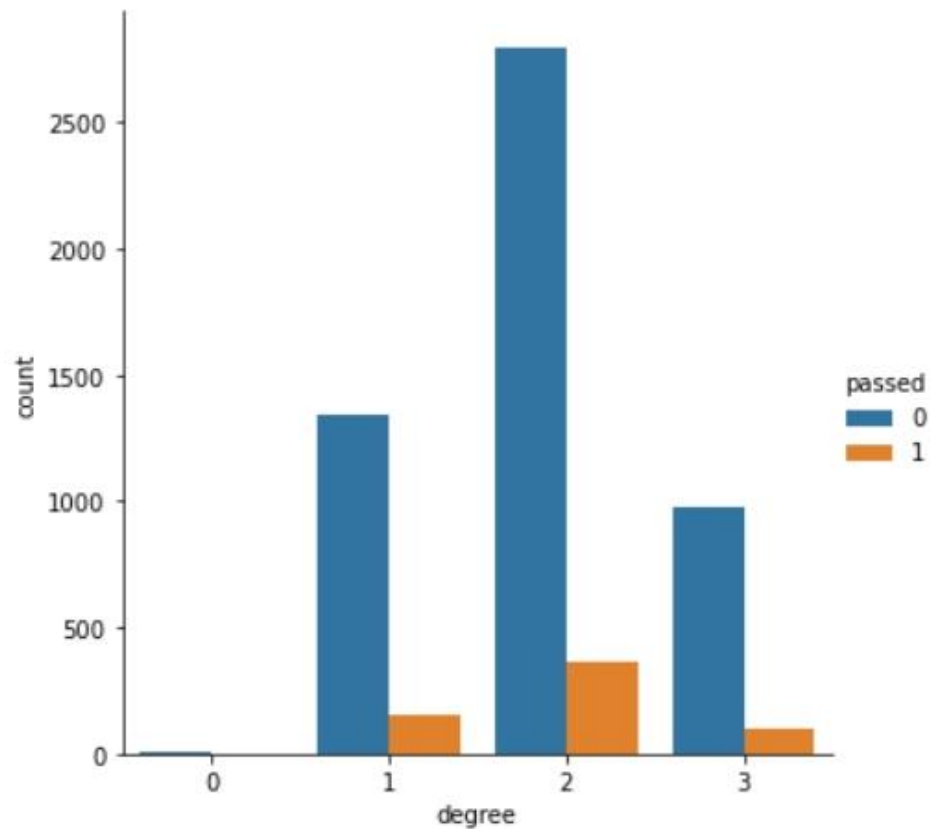
anticorr_expertise: Выявленные коррупциогенные факторы и их способы устранения **0.318**

public_discussion: Сведения о структурных подразделениях разработчика, рассмотревших предоставленные предложения **0.317**

other_notification_info: Иные необходимые, по мнению разработчика, сведения **0.315**

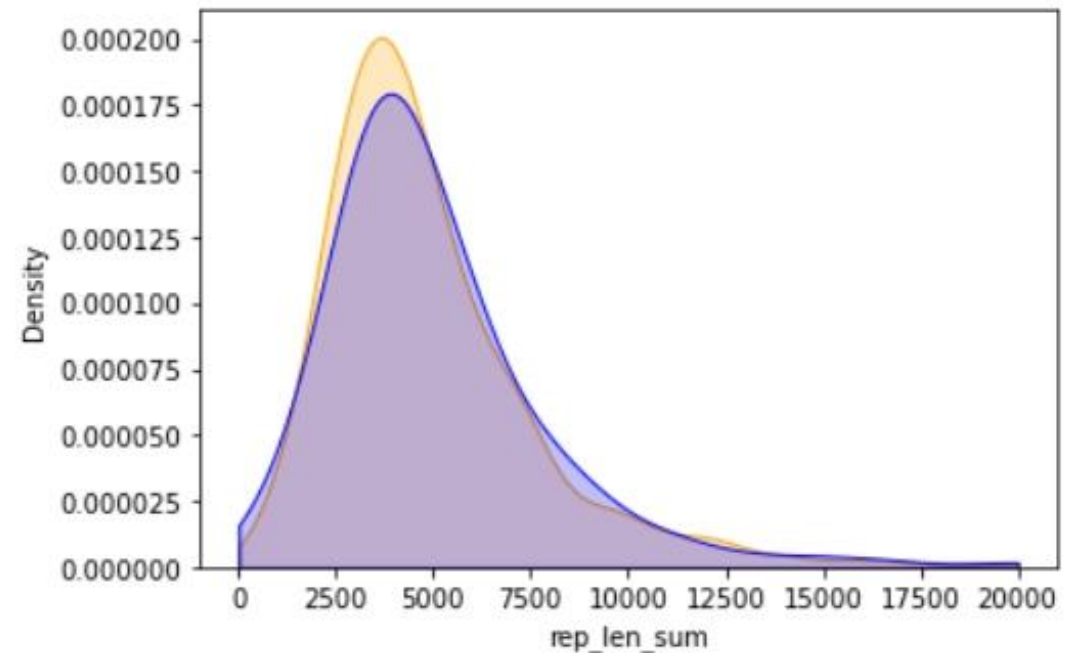
Признаки из ria_reports_main.csv

degree: Степень регулирующего
воздействия проекта акта
Очистка, Label Encoding



rep_len_sum

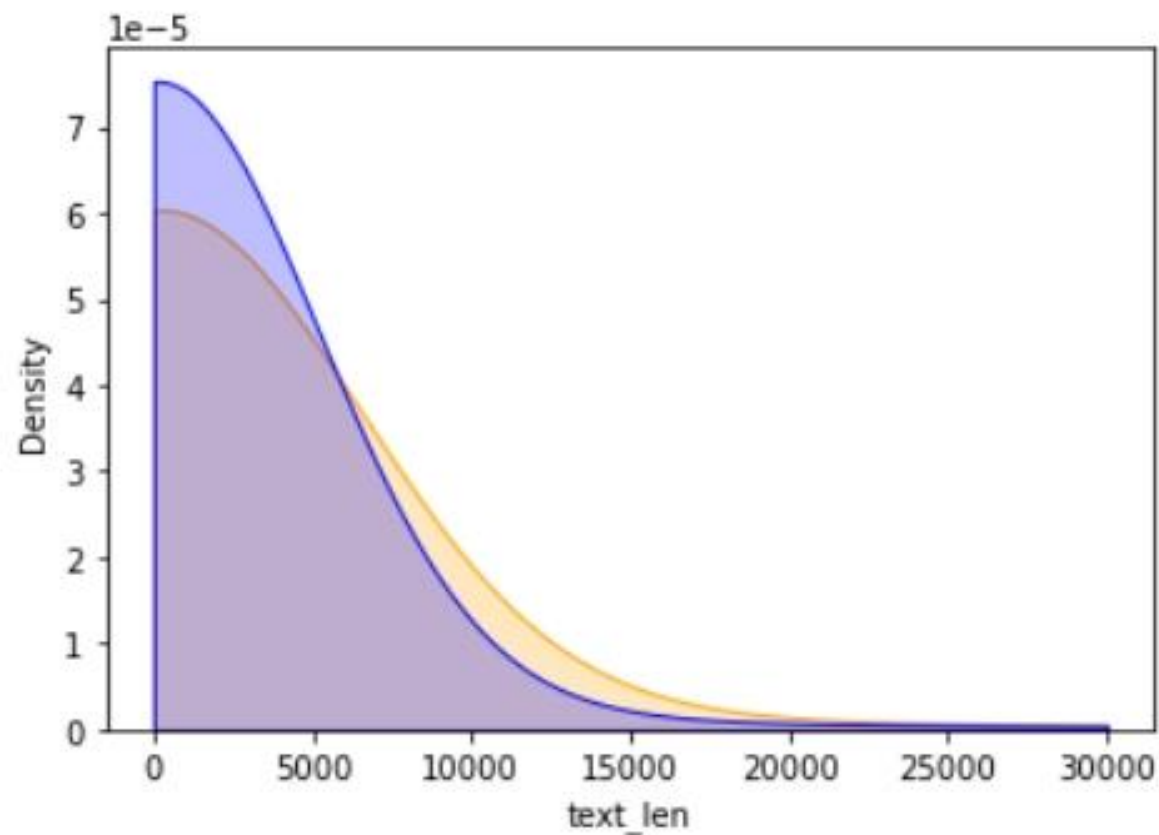
Мера заполненности отчета: сумма длин
всех записей



Признак из regulations_texts.csv

text_len

Объем текста законопроекта



Финальная модель



CatBoost

cat_features:

act_changes_controlling_activities, act_objectives, added_by, added_resp, cluster_kmeans_ngram_6, cluster_kmeans_ngram_8, degree, developer, is_regionally_significant, mineco_solution, okved_list, persons_affected_by_act, publication_day, publication_month, publication_year, regulatory_impact, relations_regulated_by_act, responsible

num_features:

comments_num, dislikes_num, likes_num, na_counts, rep_len_sum, text_len, views_num

Итоги

ROC AUC:

0.92852

Public

0.89862

Private

