

краудфандинг в благотворительности

факторы успеха

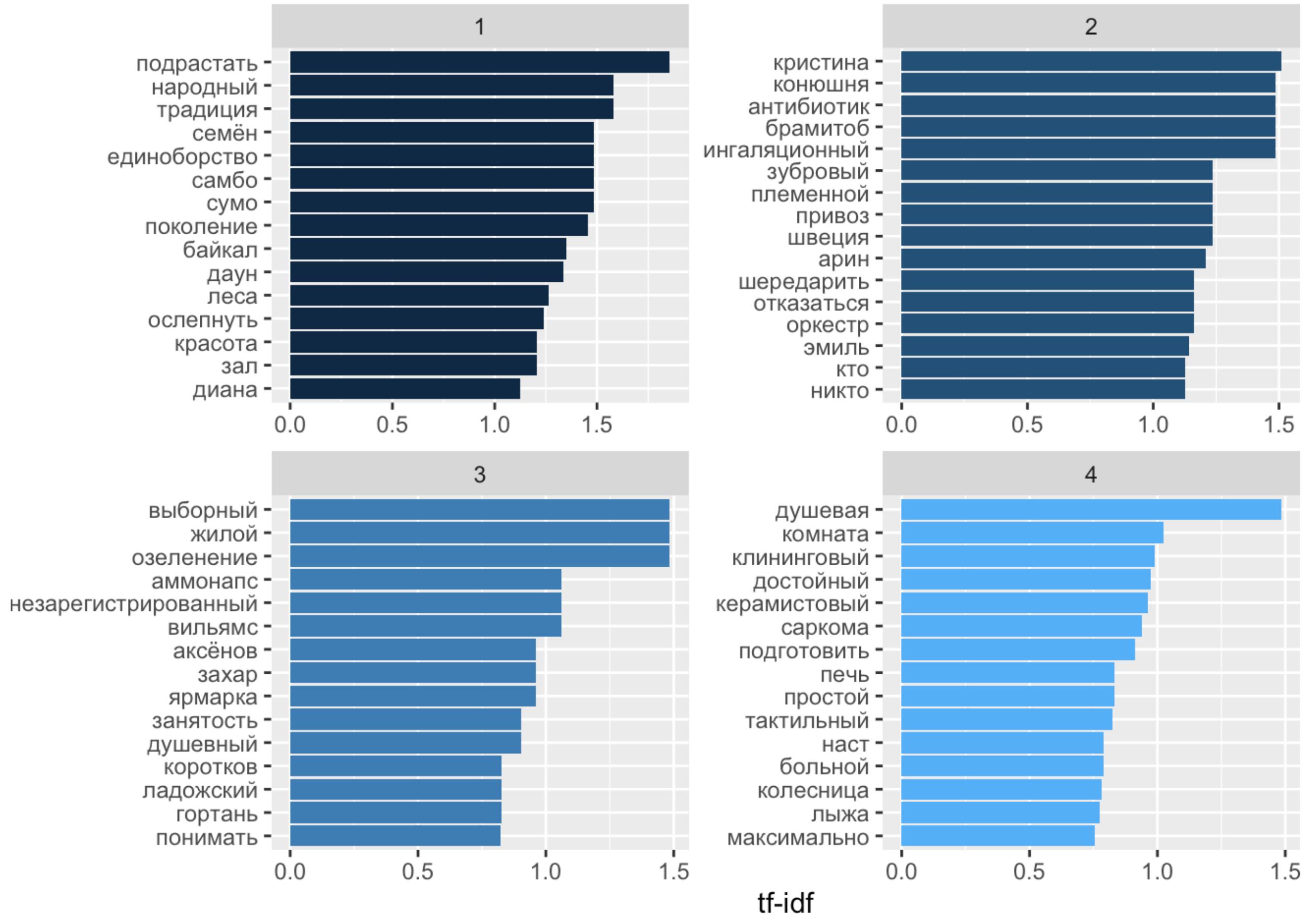
Лиза Кадетова

до этого в R

Есть ли связь между
частотностью слов и итогом
сбора - по рангам?

Кажется - нет —> цель:

посмотреть связь между
темами и успехом



Тексты не рубрицированы

→

Задачи в части МО:

1. Кластеризовать
2. Разметить
3. Обучить модель классифицировать

проблемы

- небольшая коллекция (2038 текстов)
- все документы близки по тематике и лексике: социальная проблематика, бездомные животные vs бездомные люди, брошенные дети - брошенные старики, стоп-слова не решают проблему
- границы спорные даже для людей (я и мои коллеги путались)
- а-ля публицистика - метафоры и фигуры речи: дети-бабочки, солнечные дети, кошачьи дети, узники ПНИ, чтобы Петя снова смог играть в футбол (не буквально)

тиปичные случаи (реальные)

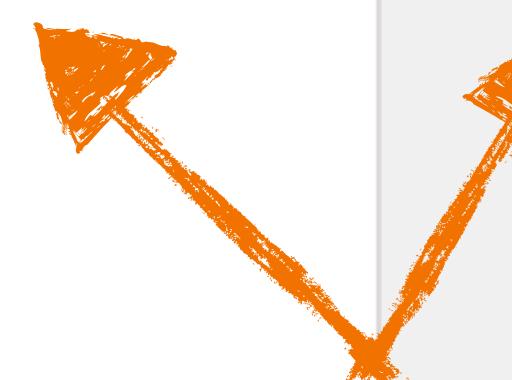
- Строительство ночлежного дома для **бездомных**.
- Строительство приюта для **бездомных** животных.
- Малыш остался **без матери**. **Медвежонка** выходили и выкормили.
- Подарим жизнь маленькому **Мишутке!** (лечение мальчика)
- "Право на помощь". 20 юристов-добровольцев помогают самым незащищенным - **старикам и детям**.



Благотворительность

Бабушки на карантине

Проект направлен на оказание помощи пожилым людям, которые вынуждены находиться дома из-за пандемии коронавируса.



72 %

50 900 ₽



Благотворительность

Зона для общения людей и диких животных

Сбор средств на постройку вольерного комплекса для псовых в рамках постройки приюта для содержания постояльцев центра реабилитации и реинтродукции диких животных "Сирин"

14 %

144 655 ₽

Бабушки на карантине

Проект направлен на оказание помощи пожилым людям, которые вынуждены находиться дома из-за пандемии коронавируса.



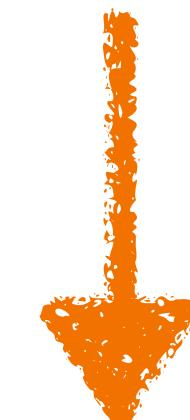
В чем заключается эта помощь? Все очень просто! Мы оказываем гуманитарную, консультативную, добровольческую помощь, проводим различные досуговые мероприятия. В нашу программу включена медико-профилактическая составляющая: оздоровительные и просветительские проекты в области медицины и деменции. Сотрудничаем со стационарными учреждениями (палиативные отделения, больницы, дома-интернаты и пр.).

В период пандемии коронавируса наши подопечные оказались в зоне риска и наиболее подвержены заражению вирусом. В целях профилактики и сохранению их здоровья мы запускаем работу координационного центра, в рамках которого будут закуплены необходимые гуманитарные и лекарственные наборы и переданы самим нуждающимся бабушкам и дедушкам. Для этого нам необходимо собрать 70000 руб.

Давайте вместе позаботимся о пожилых людях!



**вывод 1: на кратких описаниях
получается ерунда, на полных
появляется какая-то осмысленность**



использовано для кластеризации

- CountVectorizer, Tf-IdfVectorizer, embeddings с помощью нескольких моделей rusvectores, сокр. размерности truncated svd.
- С препроцессингом (стоп-слова) и без.
- На униграммах, биграммах, триграммах.
- Модели: kmeans, dbscan, с разным количеством кластеров

**Если несколько кластеров выглядят отлично, то другие кластеры их дублируют
и\или в остальных свалка.**



	Title	Cluster
Рождество-2017 для бедных и бездомных людей	8	
Рождество-2015 для бедных и бездомных людей	8	
Рождественский обед для бездомных - 2014	8	
Поддержите начинающих гребцов с инвалидностью	8	
Бочка равных возможностей для Ивана и Даши	8	
...	...	
«ВДОХ_ВДОХ»/«IN_INBREATHE»	8	
Молодежный развивающий лагерь по хоккею-следж	8	
Здоровое детство детям-сиротам	8	
Печать календаря в поддержку теннисистов	8	
"Юные борцы"	8	

	Title	Cluster
	Подарим жизнь	11
	Создание сайта для помощи животным	11
	Больным собакам жизненно нужен лечебный корм	11
	Сбор средств на коляску для пса Лаки	11
	Социальная ветеринария: новая точка на карте	11

		NaN
	Свой "МЯУ Дом" для кошек и котов приюта	11
	KITTY-park, первый приют для кошек без клеток	11

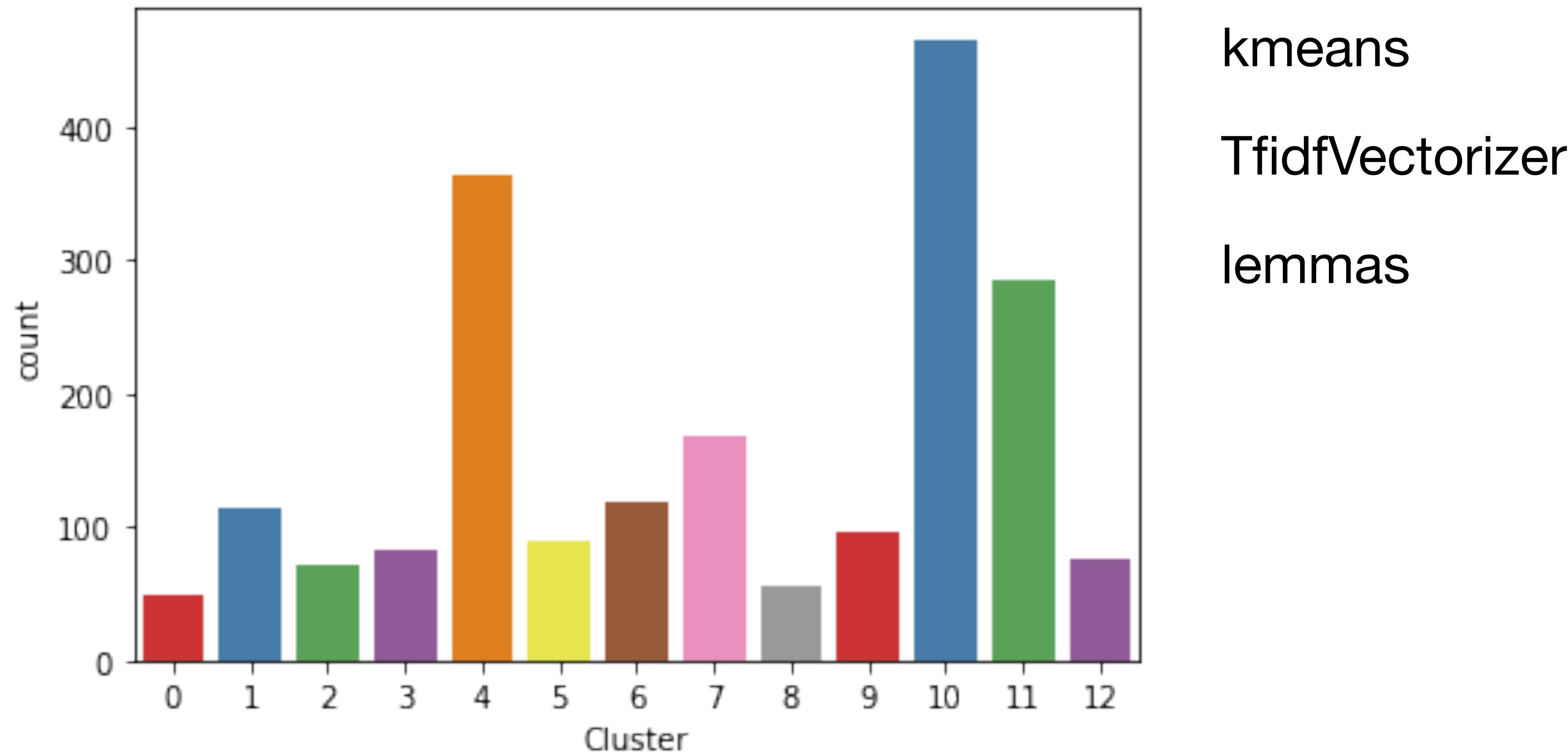
	Title	Cluster
	Календарь реабилитационного центра для птиц	9
	Построим вместе! Печь для приготовления каши	9
	Помощь приюту для собак «Я – живой!»	9
	Шанс для Шани	9
	Компьютеры для детей в малообеспеченные семьи	9

	Сдаем макулатуру - ставим турники!	9
	День Рождения Чебурашки	9
	Чистый воздух	9
	Благотворительная площадка для пожертвований	9

Дети? Мечта? Возможности?

Title	Cluster
Оплата няничек для отказных детей в больницах	12
Строительство домика для собак-инвалидов!	12
Покупаем газель для тех, кому #НечегоНадеть	12
Документальный фильм "Второе дыхание"	12
Вольеры для щенков и собак в приюте	12
...	...
Благотворительный Фестиваль «Версты Победы»	12
Поиск Детей	12
Снова собаки <small>о уходе за животными</small>	12
«Внимание и забота»	12
...	...
Благотворительный турнир по лазертагу	12
Камилла: подарите мне возможность жить дома	12
Я буду танцевать!	10
Лёша хочет ходить	10
Стань Морозом! Подари подарок редким детям!	10
Рождество как Дома	10
Дети должны жить в семье!	10
...	...
Центр адаптации для выпускников детских домов	10
"Шаг навстречу" детям ДЦП	10
Title	Cluster
Молодые и перспективные: поддержки детей-сирот	4
Школа журналистики для Ани	4
Клуб для молодых мам "Второе дыхание"	4
Я все могу!	4
Особая йога	4
...	...
Пасхальная радость для детей с инвалидностью	4
Фонд-приют "Счастье в руки" Ремонт для собак.	4
Согреем «Дом надежды на Горе»!	4

лучший результат



признаки

животное
животный который бездомный

приют
построить передержка строительство метр

волвер
забор ДОМ корова

кошка
человек весь здание будка

собака

учреждение семья дом
весь родитель школа
ребята жизнь который

ребёнок мочь волонтёр человек программа

мама детский социальный работника сирота

центр ребята столярный
который изделие навык
швейный молодая возможность

мастерской мастер ребёнок инвалид

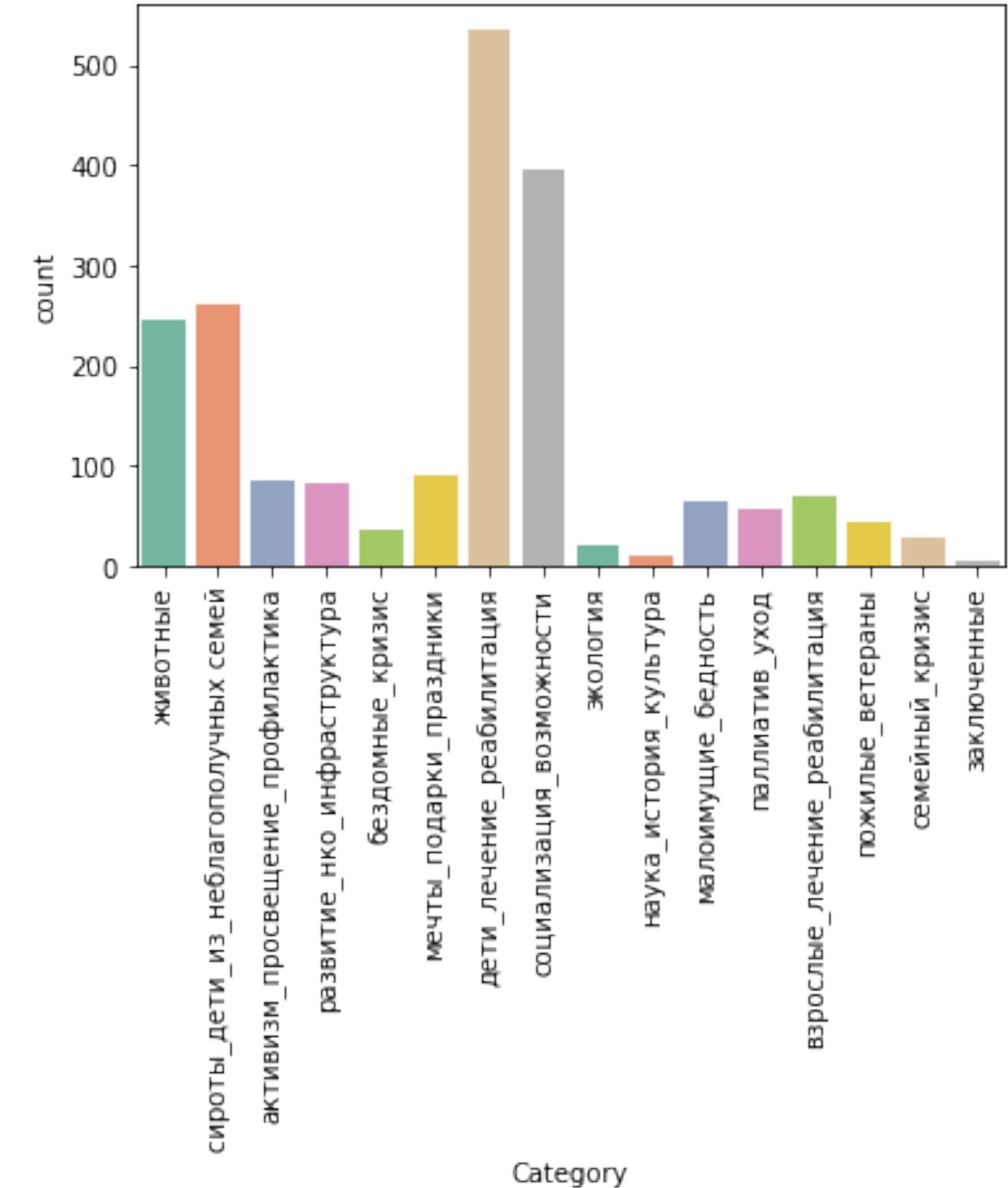
человек работа трудовой творческий занятие

разметка

В итоге размечала вручную.

Получилось 16 категорий.

Неравномерное распределение.



классификация

Попробовала модели:

- LogisticRegression
- MultinomialNB
- TreeClassifier
- RandomForestClassifier
- KNeighborsClassifier
- SVC

Лучший результат у LR.

Классификация

Варианты предобработки

- Станд. + просто стоп-слова, лемматизация или стемминг
- Сильно улучшает MultinomialNB, но он все равно в итоге уступает LR (зато работает очень быстро).
- Регрессию лемматизация и стемминг чуть-чуть улучшают (хотя могут и ухудшать при каких-то других параметрах). Удлиняют время работы векторайзера. Но сет небольшой, так что ОК.

**Оставляю леммы, т.к. все-таки улучшение есть и получается более
внятный вывод признаков на графиках.**

Кластеризация

Векторизация текстов

- CountVectorizer на всех моделях показал себя лучше, чем TfidfVectorizer, поэтому оставляю его.
- Word2vec и сокращение размерности в этот раз не стала пробовать, т.к. при кластеризации они не улучшали результат.
- Биграммы не улучшили классификацию ни с Count, ни с Tfidf, поэтому оставляем униграммы по умолчанию.
- Уточнение $\text{max_df} = 0.50\ 0.30$ и $\text{min_df} = 0.1$ от запуска к запуску могут чуть-чуть улучшать или ухудшать выдачу.

классификация

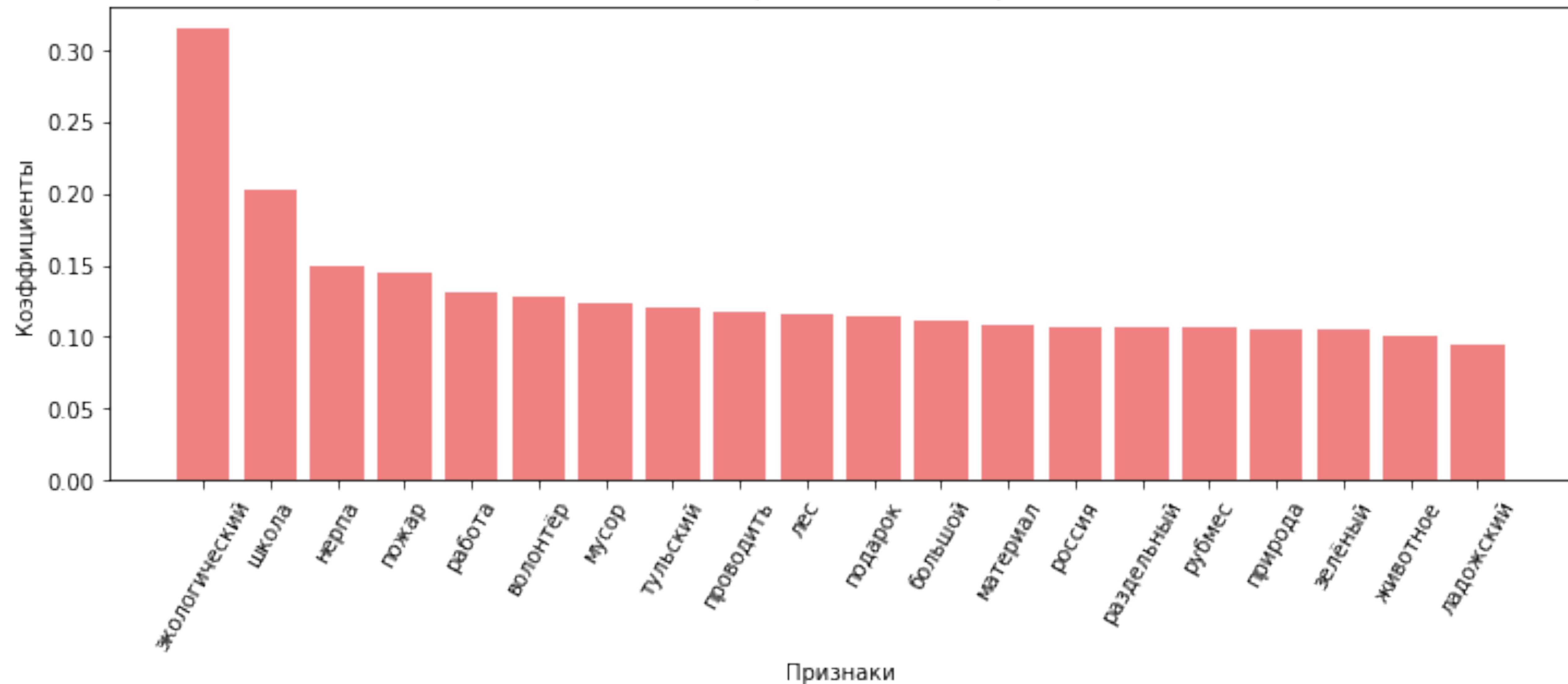
Гиперпараметры LR

- Гиперпараметры для LR были подобраны через gridsearch. Ему в словарь были поданы не все возможные параметры, т.к. это бы заняло слишком много времени.
- В действительности подобранные параметры срабатывают по-разному: в пределах +1:3% они улучшают или ухудшают выдачу.

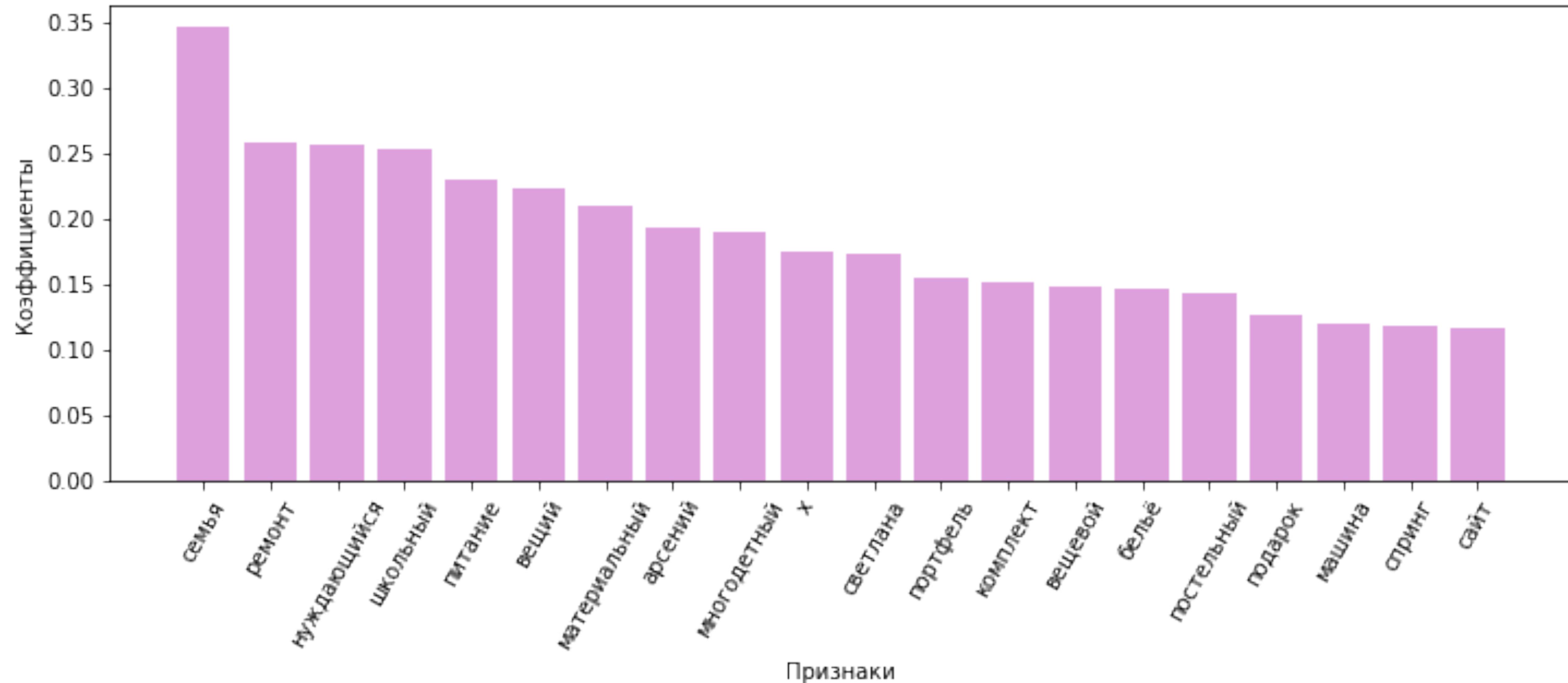
LR + CountVectorizer + prepr. lemmas

	precision	recall	f1-score	support
активизм_просвещение_профилактика	0.50	0.24	0.32	21
бездомные_кризис	1.00	0.67	0.80	9
взрослые_лечение_реабилитация	0.67	0.33	0.44	18
дети_лечение_реабилитация	0.79	0.88	0.83	134
животные	0.97	1.00	0.98	61
заключенные	0.00	0.00	0.00	1
малоимущие_бедность	0.40	0.35	0.38	17
мечты_подарки_праздники	0.67	0.78	0.72	23
наука_история_культура	0.00	0.00	0.00	3
паллиатив_уход	0.82	0.64	0.72	14
пожилые_ветераны	0.78	0.64	0.70	11
развитие_нко_инфраструктура	0.38	0.24	0.29	21
семейный_кризис	0.67	0.57	0.62	7
сироты_дети_из_неблагополучных семей	0.72	0.85	0.78	66
социализация_возможности	0.73	0.81	0.77	99
экология	1.00	0.60	0.75	5
accuracy			0.75	510
macro avg	0.63	0.54	0.57	510
weighted avg	0.74	0.75	0.74	510

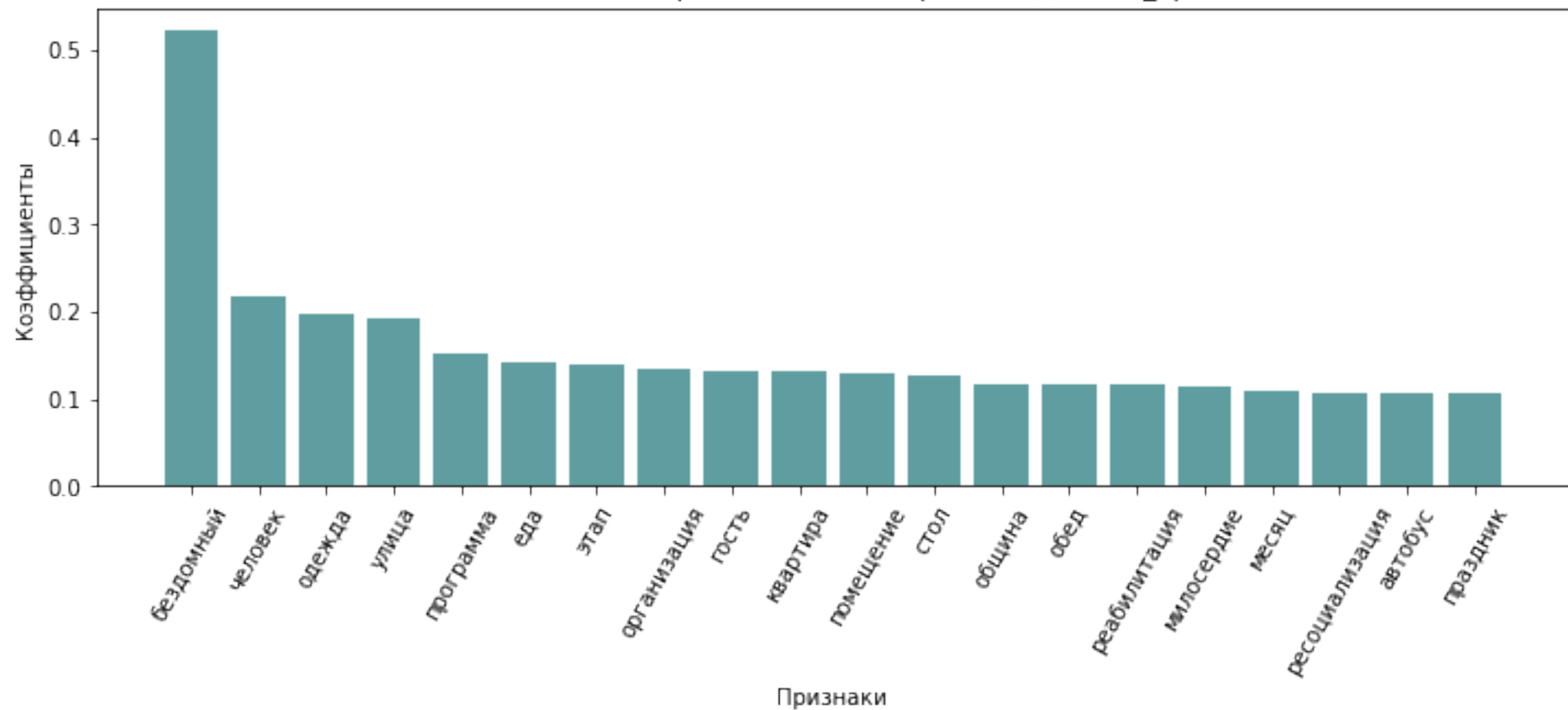
Самые значимые признаки в категории: экология



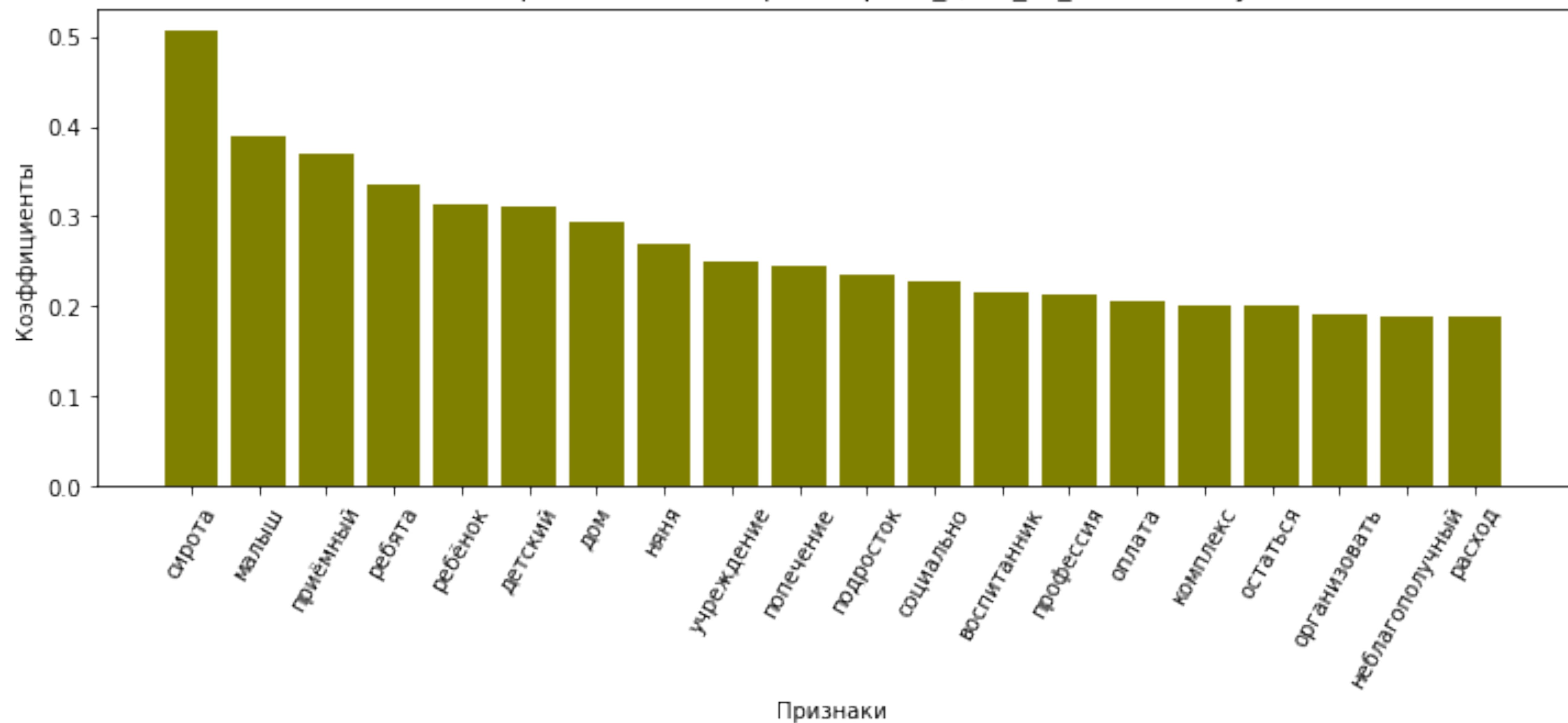
Самые значимые признаки в категории: малоимущие_бедность



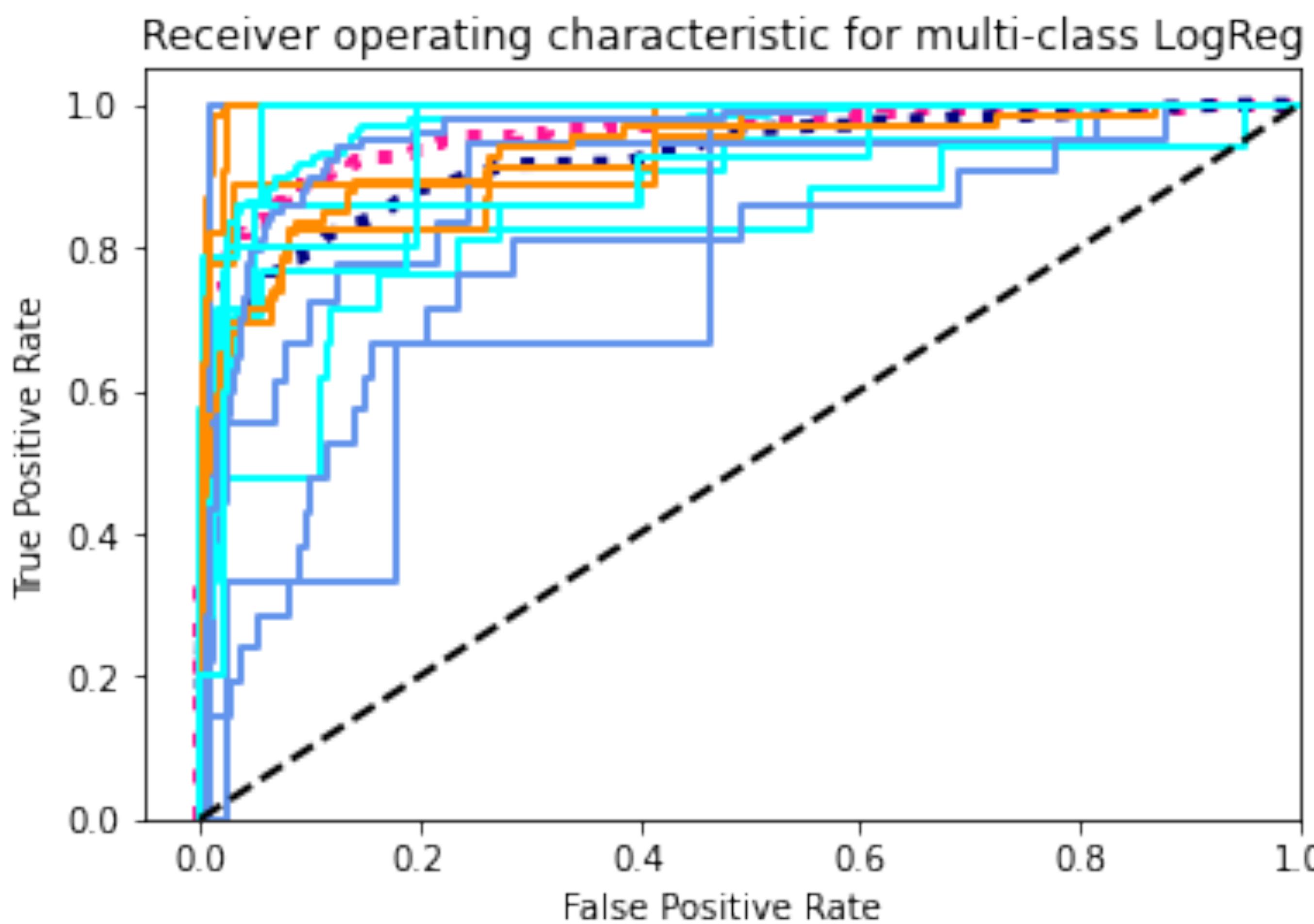
Самые значимые признаки в категории: бездомные_кризис



Самые значимые признаки в категории: сироты_дети_из_неблагополучных семей



Confusion matrix



- micro-average ROC curve (area = 0.96)
- macro-average ROC curve (area = 0.92)
- ROC curve of class 0 (area = 0.86)
- ROC curve of class 1 (area = 0.95)
- ROC curve of class 2 (area = 0.89)
- ROC curve of class 3 (area = 0.97)
- ROC curve of class 4 (area = 0.99)
- ROC curve of class 5 (area = 0.99)
- ROC curve of class 6 (area = 0.85)
- ROC curve of class 7 (area = 0.93)
- ROC curve of class 8 (area = 0.78)
- ROC curve of class 9 (area = 0.93)
- ROC curve of class 10 (area = 0.99)
- ROC curve of class 11 (area = 0.78)
- ROC curve of class 12 (area = 0.98)
- ROC curve of class 13 (area = 0.93)
- ROC curve of class 14 (area = 0.96)
- ROC curve of class 15 (area = 0.95)

далше

- проверить на внешних текстах
- научить предсказывать успех (сумму) по категории

charity.planeta.ru

github.com/Liza-Kadetova/Project-1