Дмитрий Кузюрин

Решение задачи

Узнай героев сериала "Друзья" с первых слов

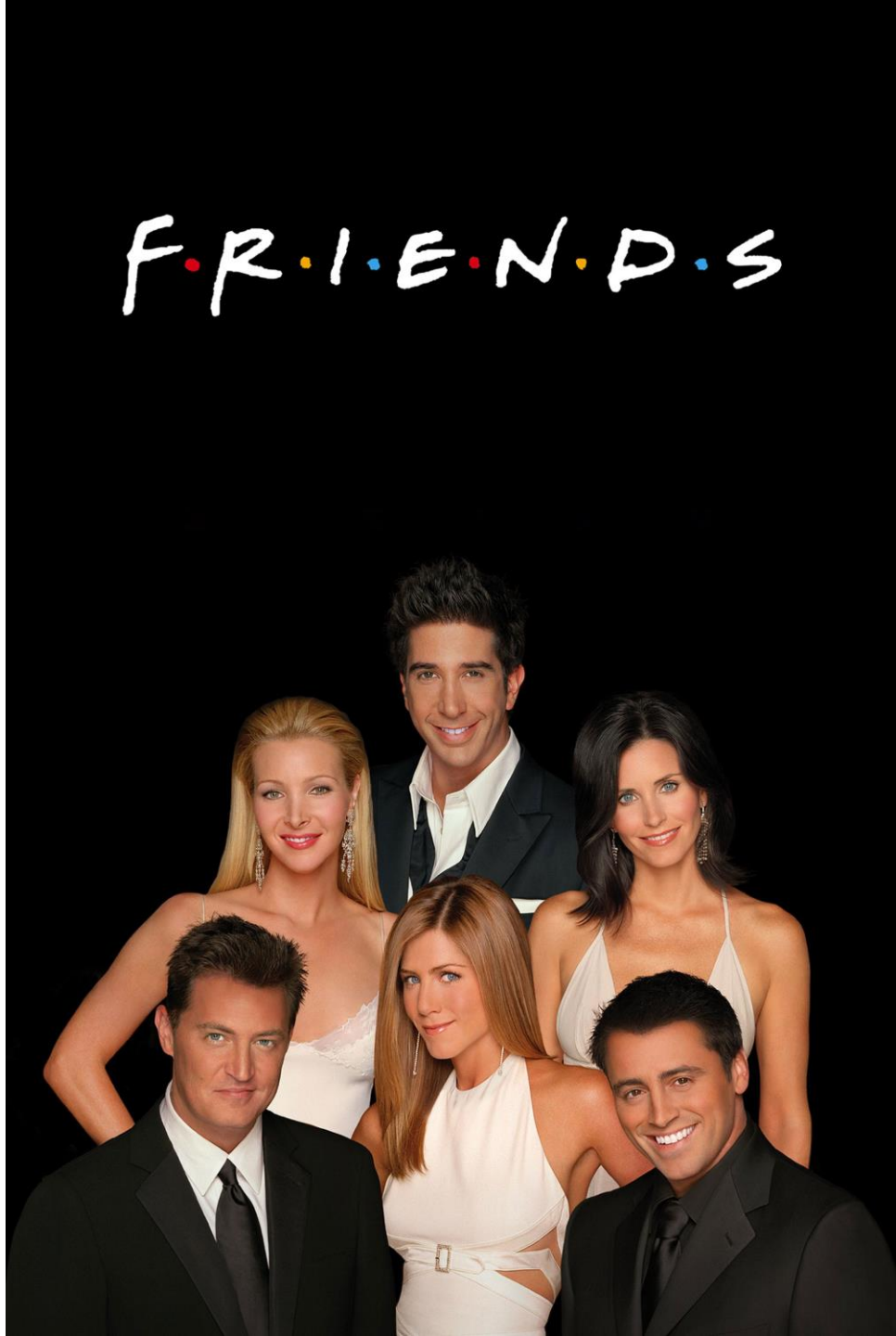
RuCode 4.0, ноябрь 2021

https://www.kaggle.com/c/friends-classification

dkuzyurin@gmail.com

@dkuzyurin

RUCODE festival



Кто автор первой реплики?

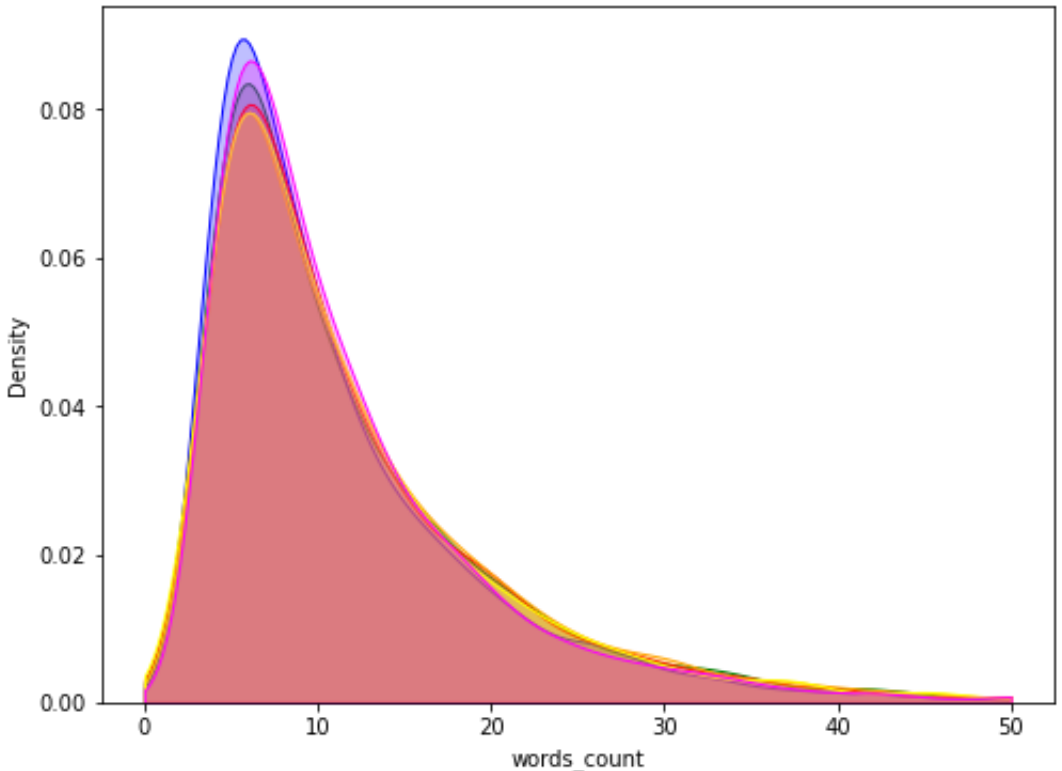
В 35% случаев мы это знаем!

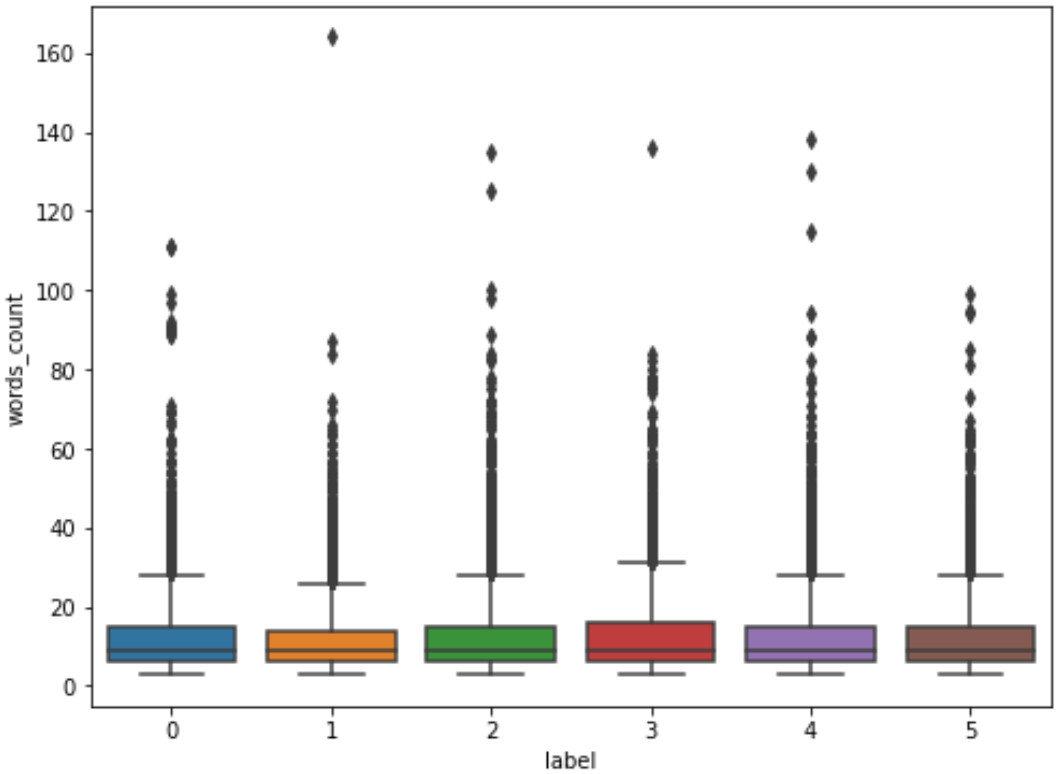
	other_speaker	friend_response	label
2579	Мм-хм. Она… она эмоциональна, но напористая.	Вы знаете, что я собираюсь делать? Я собираюсь...	2
7369	Да, думаю, я неправильно сказал. Знаешь, мы до...	Ой, торт действительно вкусный!	3
14229	О боже.	Четыре буквы: Круг или обруч.	3
14381	Ой, боже, это плохо.	Дорогая, может нам стоит отвезти тебя к врачу.	2
16036	Значит! Ты действительно заставил меня туда по...	О, мы могли бы заниматься этим весь день.	5
16238	Еще одна ложь. Вы заболели!	Росс, мне кажется, ты больше не замужем за нам...	5
18529	Росс! Там твои волнистые черные линии!	Хорошо, теперь, когда Росс знает, можешь расск...	1
19443	Ты нехороший!	Всегда приятно познакомиться с фанатом!	0
22704	Я не смогу съесть еще кусочек.	Мне нужно что-нибудь сладкое.	0

	other_speaker	friend_response	label
5187	Четыре буквы: Круг или обруч.	Кольцо, черт возьми, кольцо!	5
5394	Росс, мне кажется, ты больше не замужем за нам...	Эй, кто-то оставил свои ключи. Оооо, в Порше! ...	0
10116	Хорошо, теперь, когда Росс знает, можешь расск...	Начни с того, где.	4
12571	Дорогая, может нам стоит отвезти тебя к врачу.	Нет, мой окулист Ричард! Я не могу пойти к нем...	1
13617	О, мы могли бы заниматься этим весь день.	Ладно, послушай, давай поговорим о том, какой ...	0
15681	Ой, торт действительно вкусный!	Ну ладно, понимаете? Дела уже налаживаются!	4
17159	Мне нужно что-нибудь сладкое.	Кто-нибудь хочет смотреть телевизор?	4
19862	Всегда приятно познакомиться с фанатом!	Итак, что вы здесь делаете?	2
22848	Вы знаете, что я собираюсь делать? Я собираюсь...	А вы думали, она нам помешает! Так почему бы т...	3

	other_speaker	friend_response	label	other_speaker_label
2579	Мм-хм. Она… она эмоциональна, но напористая.	Вы знаете, что я собираюсь делать? Я собираюсь...	2	-1
5187	Четыре буквы: Круг или обруч.	Кольцо, черт возьми, кольцо!	5	3
5394	Росс, мне кажется, ты больше не замужем за нам...	Эй, кто-то оставил свои ключи. Оооо, в Порше! ...	0	5
7369	Да, думаю, я неправильно сказал. Знаешь, мы до...	Ой, торт действительно вкусный!	3	4
10116	Хорошо, теперь, когда Росс знает, можешь расск...	Начни с того, где.	4	1
12571	Дорогая, может нам стоит отвезти тебя к врачу.	Нет, мой окулист Ричард! Я не могу пойти к нем...	1	2
13617	О, мы могли бы заниматься этим весь день.	Ладно, послушай, давай поговорим о том, какой ...	0	5
14229	О боже.	Четыре буквы: Круг или обруч.	3	-1
14381	Ой, боже, это плохо.	Дорогая, может нам стоит отвезти тебя к врачу.	2	4
15681	Ой, торт действительно вкусный!	Ну ладно, понимаете? Дела уже налаживаются!	4	3

Количество слов во friend_response





Токенизация

Оставляем стоп-слова

Лемматизация и стемминг

pymystem3.Mystem

Знаешь, я думаю, ты сможешь ее забрать.

знаешь я думаю ты сможешь ее забрать

знать я думать ты смочь она забирать

nltk.stem.snowball.SnowballStemmer

Знаешь, я думаю, ты сможешь ее забрать.

знаешь я думаю ты сможешь ее забрать

знаеш я дума ты сможеш е забра

Финальная модель

CatBoost

text_features:
other_speaker, friend_response

Проверка на совпадение с автором первой реплики

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Dmitry Kuzyurin			0.39826	29	12d
	Your Best Entry					
	Your submission scored 0.39826, which is an improvement of your previous score of 0.37229. Great job! Tweet this					
2	reg0×00			0.35930	3	2d
3	♂Deep♂ Neural			0.35714	3	14d
4	Бобы			0.35497	92	2d
5	Ivan V. Savkin			0.35281	16	13d

#	Δpub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲6	Andrew Argatkiny			0.33879	1	14d
2	▲7	катбуст: скоростной мангуст			0.33536	2	17d
3	▼1	reg0×00			0.33460	3	2d
4	▼3	Dmitry Kuzyurin			0.32812	29	12d
5	▼2	♂Deep♂ Neural			0.32736	3	14d

Пути улучшения:

1. Иногда известно не только кто обращается к герою, но и к кому обращается сам герой

2. Поэксперементировать с обработкой текста внутри CatBoost: ngram (1,2), макс.объем словаря и т.д.

3. Ансамбль моделей (CatBoost + LogisticRegression и т.д.)