DeepPavlov.ai

ODS QA

Финальный проект курса:
DL in NLP Spring 2020

Докладчик: Юрий Зеленский
Куратор: Алексей Сорокин

Команда проекта ODS QA

Илья Сироткин (@Satel1ite) - основной вклад

Екатерина Карпова (@vengodelsur) - автор изначальной идеи

Юрий Зеленский (@yzelensky) - организация, идеи, доклад

Вадим Нарейко (@vnareyko) - консультации

Алексей Сорокин - куратор

Вехи

“Автоответчик ODS” в slack #ods_pet_projects

- **Open domain QA != SQuAD**
- **Open domain QA ещё не с нами**

Кластеризация вопросов

- “Родная” задача для USE с опубликованным baseline (Quora, AskUbuntu)
- ODS Community - сообщество высокой культуры

Классификация вопросов по каналам в Slack ODS:

- сравнительный анализ современных моделей
- робот рекомендующий тематический канал

https://arxiv.org/abs/1907.04307

ekatkarp 9:55 PM

Wednesday, March 18th

Привет!

В ожидании уточнений про то, как будет проходить наш онлайн-формат, приглашаем и вас тоже рассказывать про идеи проектов. Пока — ещё один проект во имя сообщества под уже мелькавшим кодом "**ODS QA**".

Вас должна воодушевить идея, если вы грустите, когда:

- пишут по много раз одни и те же вопросы в канал, ну можно же по слаку поискать!
- задаёшь вопрос, а в ответ только «было уже, поищите», да пытал я уже этот поиск по слаку, ну тяжко ему с русской морфологией!
- хочется попилить что-нибудь NLPшное для души, а на кагле одни картиночки!

Мы хотим сделать «автоответчик» для ODS. Это значит, что можно:

- получить плюшечки от сообщества, как это всегда бывает с проектами во славу ODS
- покрутить модные модельки (yorko не прочь поделиться опытом успешнейшего гоняния альбертов на TPU)
- да что угодно в целом, хотите попробовать написать модель на ассемблере или сделать демку с анимацией на HTML5 — мы чужие вкусы не осуждаем...
- и отдельная интересная часть, с которой всё начнётся: поразвлекаться с краудсорсингом. Нам под силу сделать из нашего чудесного дампа слака датасет, достойный стать Новым Важным Бенчмарком в Question Answering.

Пока неизвестно, что может встать на нашем пути: саркастические ответы? Грязная ложь в комментариях, противоречащая документации pytorch? Расползание обсуждения по нескольким тредам? Так что попутно помимо вопросов-ответов может понадобиться ещё много какая разметка, идущая рука об руку с новыми задачами, так что не волнуйтесь, тем для курсовых проектов на nlpшных курсах хватит всем.

Ставьте ➕, кому интересно поучаствовать. Тогда я не позволю вам пропустить ссылочку на репозиторий:) (edited)

➕ 16 🔥 1

+1 **10 replies** Last reply 4 months ago

https://opendatascience.slack.com/archives/CJW0A6U78/p1584557756041500

Вехи

“Автоответчик ODS” в slack #ods_pet_projects

- Open domain QA != SQuAD
- Open domain QA ещё не с нами

Кластеризация вопросов

- **“Родная” задача для USE с опубликованным baseline (Quora, AskUbuntu)**
- **ODS Community - сообщество высокой культуры**

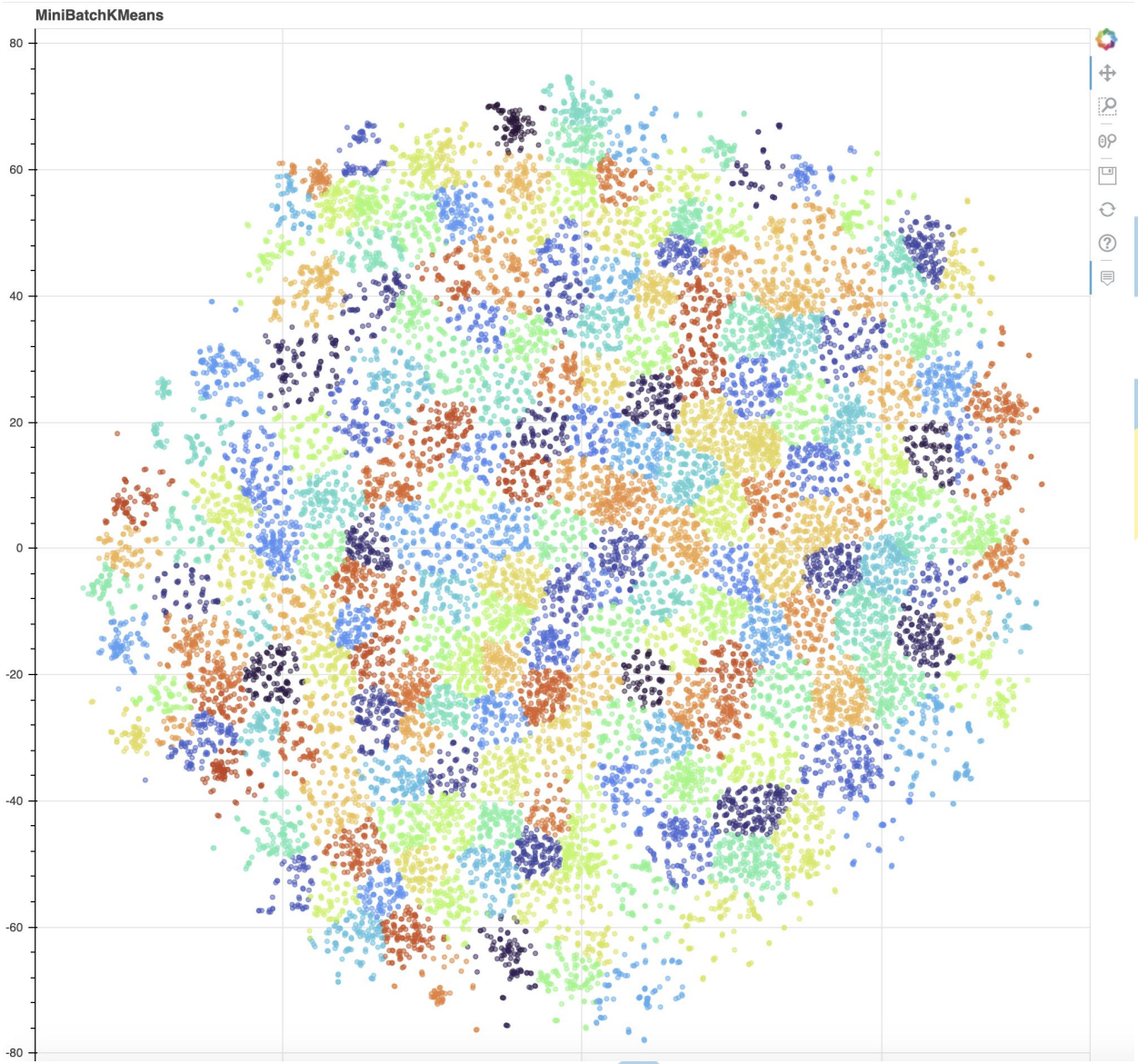
Классификация вопросов по каналам в Slack ODS:

- сравнительный анализ современных моделей
- робот рекомендующий тематический канал

https://arxiv.org/abs/1907.04307

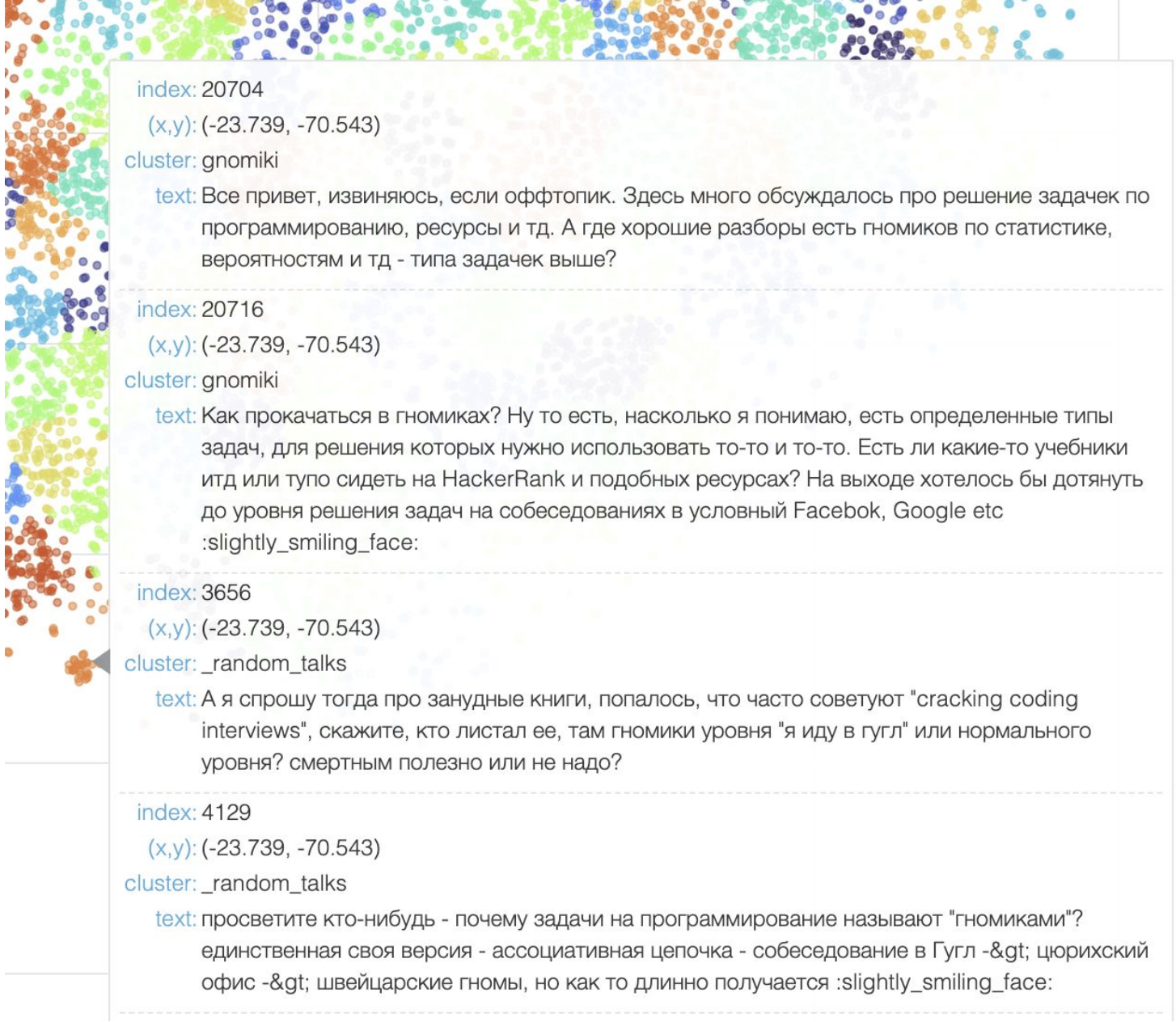
EDA

bokeh
tsna, k-means
embeddings
jittering



EDA

Редкий пример вопросов дубликатов



index: 20704
(x,y): (-23.739, -70.543)
cluster: gnomiki
text: Все привет, извиняюсь, если оффтопик. Здесь много обсуждалось про решение задачек по программированию, ресурсы и тд. А где хорошие разборы есть гномиков по статистике, вероятностям и тд - типа задачек выше?

index: 20716
(x,y): (-23.739, -70.543)
cluster: gnomiki
text: Как прокачаться в гномиках? Ну то есть, насколько я понимаю, есть определенные типы задач, для решения которых нужно использовать то-то и то-то. Есть ли какие-то учебники итд или тупо сидеть на HackerRank и подобных ресурсах? На выходе хотелось бы дотянуть до уровня решения задач на собеседованиях в условный Facebok, Google etc :slightly_smiling_face:

index: 3656
(x,y): (-23.739, -70.543)
cluster: _random_talks
text: А я спрошу тогда про занудные книги, попалось, что часто советуют "cracking coding interviews", скажите, кто листал ее, там гномики уровня "я иду в гугл" или нормального уровня? смертным полезно или не надо?

index: 4129
(x,y): (-23.739, -70.543)
cluster: _random_talks
text: просветите кто-нибудь - почему задачи на программирование называют "гномиками"? единственная своя версия - ассоциативная цепочка - собеседование в Гугл -> цюрихский офис -> швейцарские гномы, но как то длинно получается :slightly_smiling_face:

Вехи

“Автоответчик ODS” в slack #ods_pet_projects

- Open domain QA != SQuAD
- Open domain QA ещё не с нами

Кластеризация вопросов

- “Родная” задача для USE с опубликованным baseline (Quora, AskUbuntu)
- ODS Community - сообщество высокой культуры

Классификация вопросов по каналам в Slack ODS:

- **сравнительный анализ современных моделей**
- **робот рекомендующий тематический канал**

https://arxiv.org/abs/1907.04307

Итоговая задача

Мета-идея:

- сравнительный анализ современных моделей на примере данных дампа ODS с адекватной метрикой выраженной одним числом.

Постановка:

- классификации принадлежности **вопроса** к slack каналу ODS

Метрика:

- Pairwise macro average multiclass ROC AUC score (pwROC-AUC).
- A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems https://link.springer.com/article/10.1023/A:1010920819831

Данные

- Дамп slack ODS с 2015-03-12 по 2020-01-14
- Выбраны треды
- Начало треда - вопрос, сообщения - ответы
- Фильтр по наличию знака вопроса
- Фильтр по каналам - 100+ вопросов

	Train	Val	Test
Вопросы	215,054	71,685	71,685
Слова	10,765,047	3,587,705	3,622,738
Классы	79	79	79

https://opendatascience.slack.com/files/U04URBM8V/FSRD4NVLN/opendatascience_slack_export_mar_12_2015_-_jan_14_2020.zip

картинка
жизнь
строка
случай
проблема
статья
хотя
ничто
ещё
привет
помочь
тема
ответ
нибыть
найти
сталкиваться
сделать
получить
правильно
дело
год
класс
человек
размер
написать
язык
слово
коллега
идти
похожий
сказать
метрика
общий
число
куча
думать
получать
понять
классификация
некоторый
задача
весь
вектор
неделя
подход
время
видеть
вид
последний
интересно
команда
признак
подобный
сильно
пара
изображение
наш
курс
самый
использовать
хотеться
например
проект
выборка
маленький
почему
файл
пробовать
часть
смотреть
пытаться
свой
русский
разный
объект
вроде
месяц
сколько
нормальный
никто
алгоритм
дать
тест
идея
готовый
ваш
оно
смысл
сторона
взять
память
модель
параметр
кажется
функция
посоветовать
работа
вопрос
знать
опыт
небольшой
сетка
читать
выбрать
возможно
результат
ребята
давать
компания
метод
должный
решать
новый
кстати
интересный
ссылка
фич
деньга
пожалуйста
вообще
кроме
сразу
очень
помощь
день
каждый
понятно
нужно
значение
говорить
мочь
поэтому
получаться
обучать
знание
либо
рассказать
канал
поиск
питон
стать
текст
точка
просто
добрый
момент
большой
следующий
способ
стоить
пока
решить
система
брать
возможность
поделиться
вакансия
далёкий
место
друг
делать
писать
мнение
народ
соревнование
насколько
посмотреть
заниматься
качество
ошибка
код
вариант
несколько
хотеть
искать
хороший
дата
тип
начать
чат
подсказать
работать
уровень
именно
датасет
пример
простой
считать
сеть
иметь
обучение
понимать
решение
нужный
спасибо
процесс
количество
попробовать
всякий

EDA словоформы(regex,pymorphy)

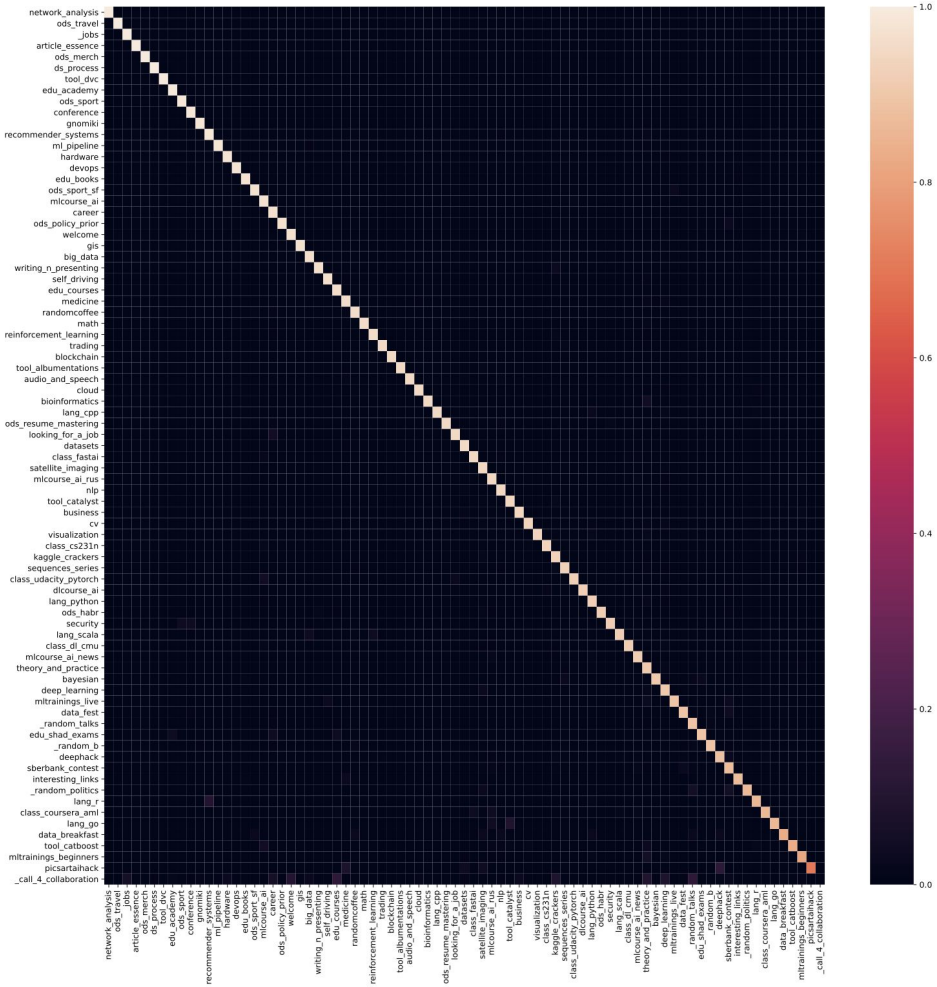
normal_form	forms_count	word_count_sum	forms
использовать	38	6684	использовать использую использует использовал используете используя использу...
хороший	37	4139	хороший хорошие лучший хорошая хорошего лучшие лучших хорошее хороших хорошу...
написать	31	2880	написать написал написано написали напишите написала напишу написаны написан...
реализовать	31	671	реализовать реализовал реализовали реализован реализовано реализована реализ...
найти	29	4902	найти нашел нашёл нашла найду нашли найдет найди найдёт найдены найдешь найд...
сделать	29	6982	сделать сделал сделали сделано сделает сделаю сделаны сделала сделан сделайт...
получить	29	3428	получить получил получили полученных получим получу получит полученный получ...
выбрать	26	1402	выбрать выбрали выбрал выбрала выбранных выбрано выбраны выбран выбери выбра...
построить	26	892	построить построил построили постройте построено построен построишь построен...
обучать	25	1470	обучать обучаю обучающей обучаем обучающую обучал обучали обучающая обучающи...
сохранить	25	421	сохранить сохранил сохранив сохранена сохранено сохраненные сохраненную сохр...
дать	25	10481	данных данные данными дать данном данной дали дайте дал даст даны дало дадут...

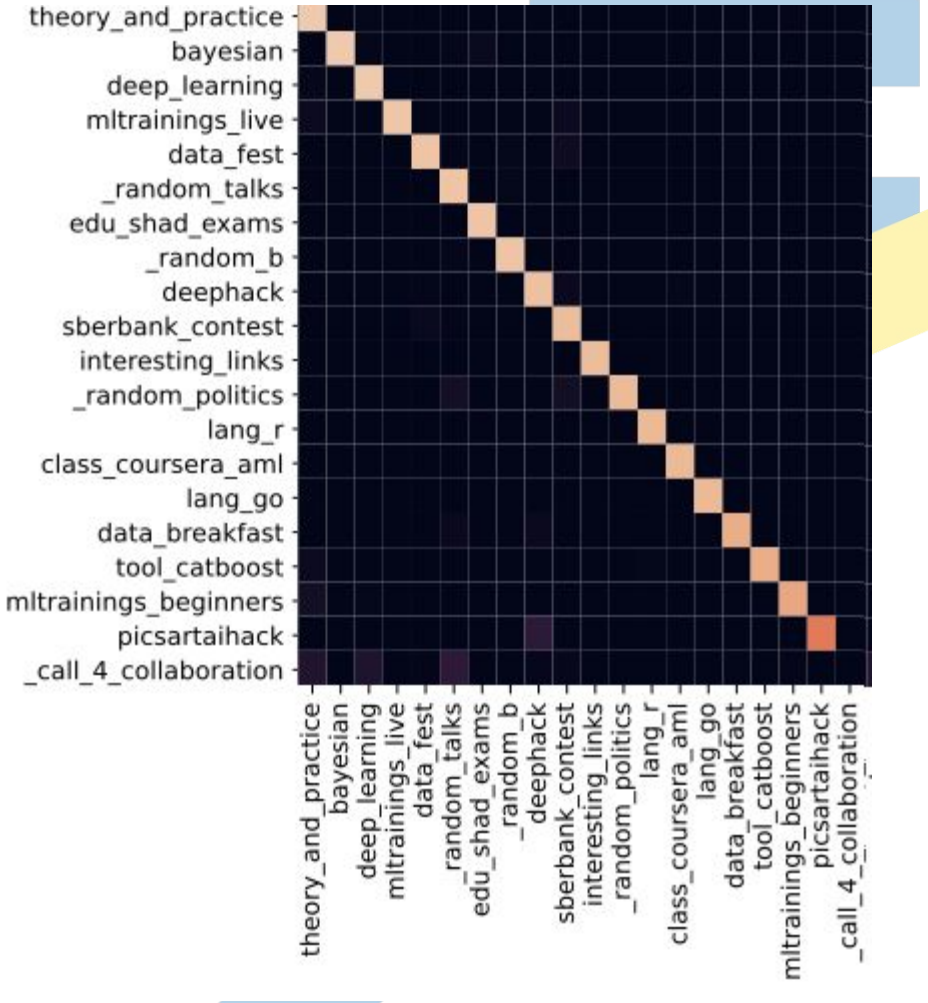
Результаты

Model	pwROC-AUC score
USE_baseline	0.61
USE_finetuned	0.985
SBERT_finetuned	0.991
DeepPavlov ru-SBERT	**0.995**
DeepPavlov mul-SBERT	0.993

https://tfhub.dev/google/universal-sentence-encoder-multilingual/3
https://github.com/UKPLab/sentence-transformers

Результаты


network_analysis
ods_travel
_jobs
article_essence
ods_merch
ds_process
tool_dvc
edu_academy
ods_sport
conference
gnomiki
recommender_systems
ml_pipeline
hardware
devops
edu_books
ods_sport_sf
mlcourse_ai
career
ods_policy_prior
welcome
gis
big_data
writing_n_presenting
self_driving
edu_courses
medicine
randomcoffee
math
reinforcement_learning
trading
blockchain
tool_albumentations
audio_and_speech
cloud
bioinformatics
lang_cpp
ods_resume_mastering
looking_for_a_job
datasets
class_fastai
satellite_imaging
mlcourse_ai_rus
nlp
tool_catalyst
business
cv
visualization
class_cs231n
kaggle_crackers
sequences_series
class_udacity_pytorch
dlcourse_ai
lang_python
ods_habr
security
lang_scala
class_dl_cmu
mlcourse_ai_news
theory_and_practice
bayesian
deep_learning
mltrainings_live
data_fest
_random_talks
edu_shad_exams
_random_b
deephack
sberbank_contest
interesting_links
_random_politics
lang_r
class_coursera_aml
lang_go
data_breakfast
tool_catboost
mltrainings_beginners
picsartaihack
_call_4_collaboration
1.0
0.8
0.6
0.4
0.2
0.0


theory_and_practice
bayesian
deep_learning
mltrainings_live
data_fest
_random_talks
edu_shad_exams
_random_b
deephack
sberbank_contest
interesting_links
_random_politics
lang_r
class_coursera_aml
lang_go
data_breakfast
tool_catboost
mltrainings_beginners
picsartaihack
_call_4_collaboration

Выводы

- Дообучение проанализированных моделей, сильно повысило их результативность на задаче классификации
- Современным моделям не нужна предварительная токенизация
- Даже дообучение требовательно к объему доступной GPU RAM
- Высокие результаты (и USE и SBERT) открывают путь к практическому использованию
- Высокие абсолютные показатели, свидетельствуют о высоком качестве разделения (и разделимости) исходных данных (и ODS - сообщество высокой культуры и эвристика построения датасета сработала)

Дальнейшие исследования и разработки

- Добавить к сравнительному анализу “классические” подходы к topic modelling (LDA, SVM on TF-IDF, etc)
- Добавить к сравнительному анализу другие современные модели, в частности справившись с их ограничениями на объем GPU RAM (ULM-FiT, ELMO)
- Провести hard sample mining анализ
- Понять причину сложностей с классификацией сообщений канала _call_4_colaboration
- Провести слепой тест, на свежих данных, исключить “утечку”
- Разработать Slack-бота, ненавязчиво рекомендующего “правильный” канал
- Вернуться к исходным, более “тяжелым” постановка задачи

Спасибо за внимание

DeepPavlov.ai

Отчет https://gist.github.com/yzelensky/72148d7b4e8ad62d2f480a2e3ccb6228