точка

банк для предпринимателей
и предприятий

Как перестать кидать ноутбуки по почте

откуда инфо:

собесы и новые люди в команде

откуда инфо:
собесы и новые люди в команде

что хотим донести:
налаженные процессы разработки улучшают жизнь

откуда инфо:
собесы и новые люди в команде

что хотим донести:
налаженные процессы разработки улучшают жизнь

что расскажем:
что такое налаженные процессы разработки и как их наладить

1. Данные в ML-проекте
2. Безопасность
3. Воспроизводимость
4. Код
5. МЛ-разработка
6. Инфра
7. Команда

точка

Данные в ML-проекте

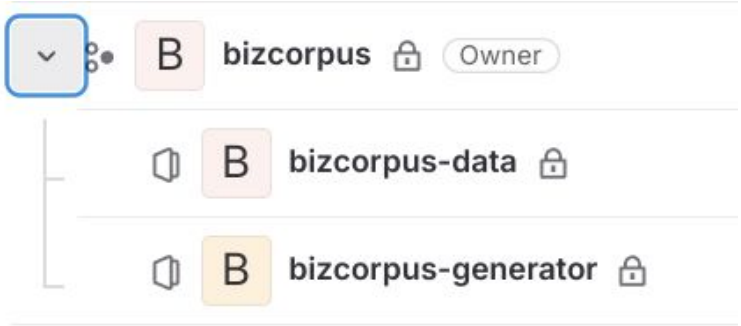
данные > безопасность > воспроизводимость > код > мл-разработка > инфра > команда


CSV

точка

точка

точка



.gitmodules 81 B

```
1  [submodule "bizcorpus-data"]
2          path = bizcorpus-data
3          url = ../bizcorpus-data.git
4
```

1. Как-то парсить данные

точка

1. Как-то парсить данные
2. При парсинге сохранять/формировать айдишники

точка

1. Как-то парсить данные
2. При парсинге сохранять/формировать айдишники
3. Проставлять внутренние айдишники

точка

1. Как-то парсить данные
2. При парсинге сохранять/формировать айдишники
3. Проставлять внутренние айдишники
4. Проставлять внутренние айдишники единообразно

1. Как-то парсить данные
2. При парсинге сохранять/формировать айдишники
3. Проставлять внутренние айдишники
4. Проставлять внутренние айдишники единообразно
5. Иметь единый формат хранения данных и либу для подгрузки

- Структуры данных
  - Материнская структура

```python
class BaseCorpusModel(BaseModel):
    internal_id: str  # ID определяет парсер: если возможно, берёт его из источника; если невозможно - генерирует через хэш по контенту
    title: str | None  # заголовок элемента (если есть)
    source: str  # название источника
    tags: frozenset[str]  # произвольный список тегов к элементу (нужно ли в отдельное поле выносить breadcrumbs-ы?)
    url: str | None  # ссылка на элемент в изначальном источнике
    created_at: datetime | None  # время создания элемента в источнике
    updated_at: datetime | None  # время обновления элемента в источнике
    author: str | None  # колобок, откуда ты это сказал?

    @computed_field
    def id(self) -> str:
       return f'{self.source}/{self.internal_id}'
```

0. SQL database

точка

0. SQL database

---

1. .txt / .json (поотдельности)

точка

0. SQL database

---

1. .txt / .json (поотдельности) - **убьёт файловую систему**

точка

0. SQL database

---

1. .txt / .json (поотдельности) - **убьёт файловую систему**
2. .csv

точка

0. SQL database

---

1. .txt / .json (поотдельности) - **убьёт файловую систему**
2. .csv - **нужно помнить разделитель**
3. .jsonl

0. SQL database

---

1. .txt / .json (поотдельности) - **убьёт файловую систему**
2. .csv - **нужно помнить разделитель**
3. .jsonl
4. .parquet
    - поддерживает структуры данных
    - партицирование
    - сжатие (очень хорошее сжатие)

https://arrow.apache.org/docs/python/parquet.html

0. SQL database

---

1. .txt / .json (поотдельности) - **убьёт файловую систему**
2. .csv - **нужно помнить разделитель**
3. .jsonl
4. .parquet
   - поддерживает структуры данных
   - партицирование
   - сжатие (очень хорошее сжатие)

https://arrow.apache.org/docs/python/parquet.html

точка

README

<repo-name>-data

- как подгрузить данные?
- * описание данных
- * объём

<repo-name>-generator

- как запустить генерацию/парсинг?

Datasets

🤗 Datasets is a library for easily accessing and sharing datasets for Audio, Computer Vision, and Natural Language Processing (NLP) tasks.

Load a dataset in a single line of code, and use our powerful data processing methods to quickly get your dataset ready for training in a deep learning model. Backed by the Apache Arrow format, process large datasets with zero-copy reads without any memory constraints for optimal speed and efficiency. We also feature a deep integration with the Hugging Face Hub, allowing you to easily load and share a dataset with the wider machine learning community.

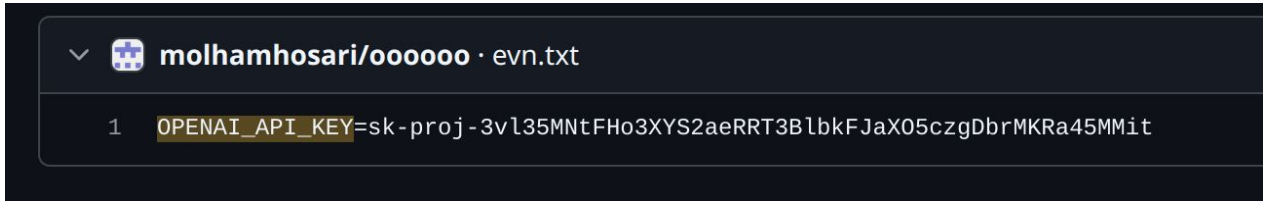
https://huggingface.co/docs/datasets/index

точка

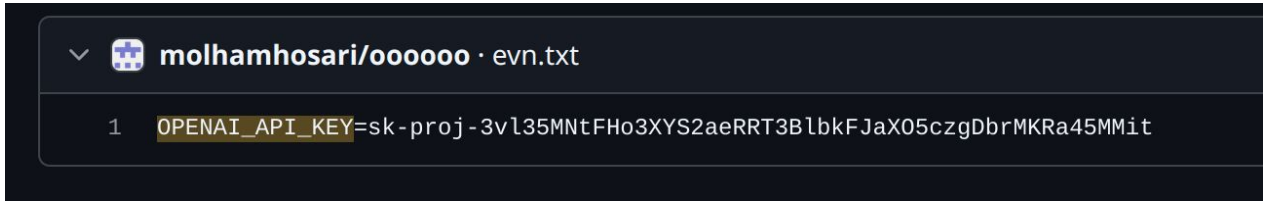
Безопасность

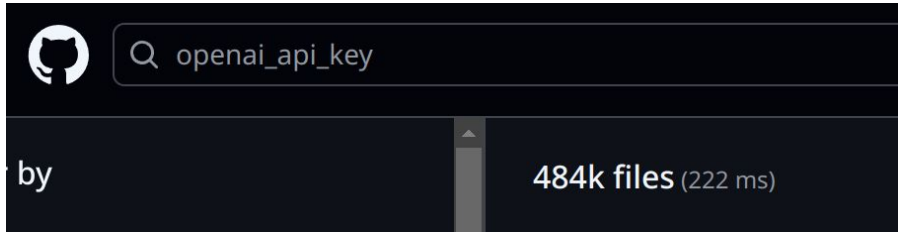
данные > **безопасность** > воспроизводимость > код > мл-разработка > инфра > команда

точка


molhamhosari/oooooo · evn.txt
1 OPENAI_API_KEY=sk-proj-3vl35MNtFHo3XYS2aeRRT3BlbkFJaXO5czgDbrMKRa45MMit

точка


molhamhosari/oooooo · evn.txt
1 OPENAI_API_KEY=sk-proj-3vl35MNtFHo3XYS2aeRRT3BlbkFJaXO5czgDbrMKRa45MMit


openai_api_key
by
484k files (222 ms)

точка

секреты не должны лежать в проекте

секреты не должны лежать в проекте

* бейте себя каждый раз, когда пишете секрет в коде

секреты не должны лежать в проекте

* бейте себя каждый раз, когда пишете секрет в коде

* самый простой (и безопасный) способ - читать секреты из environment variables

секреты не должны лежать в проекте

* бейте себя каждый раз, когда пишете секрет в коде

* самый простой (и безопасный) способ - читать секреты из environment variables

* если в проекте есть .env - 300 раз проверьте, что он в .gitignore

секреты не должны лежать в проекте

* бейте себя каждый раз, когда пишете секрет в коде

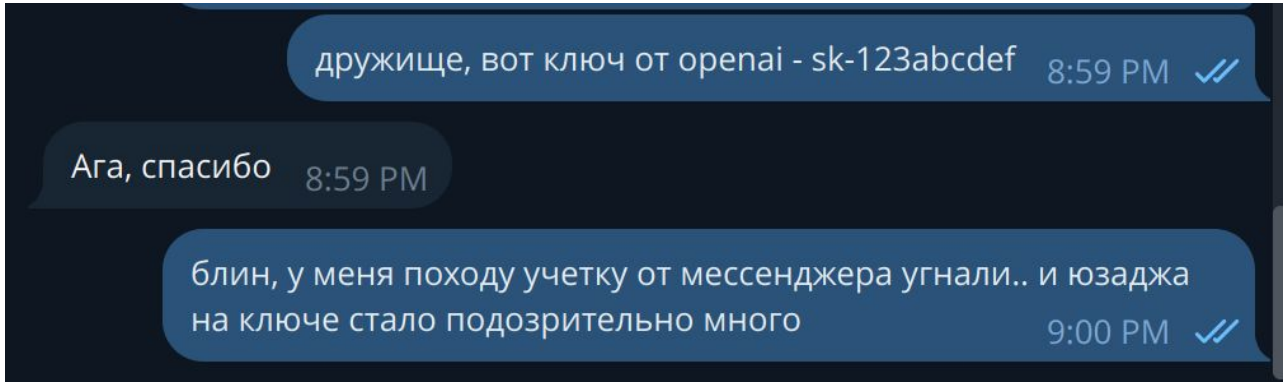
* самый простой (и безопасный) способ - читать секреты из environment variables

* если в проекте есть .env - 300 раз проверьте, что он в .gitignore

* ещё более безопасный способ - использовать корпоративный Vault

https://www.vaultproject.io/

точка


дружище, вот ключ от openai - sk-123abcdef 8:59 PM
Ага, спасибо 8:59 PM
блин, у меня походу учетку от мессенджера угнали.. и юзаджа на ключе стало подозрительно много 9:00 PM

точка

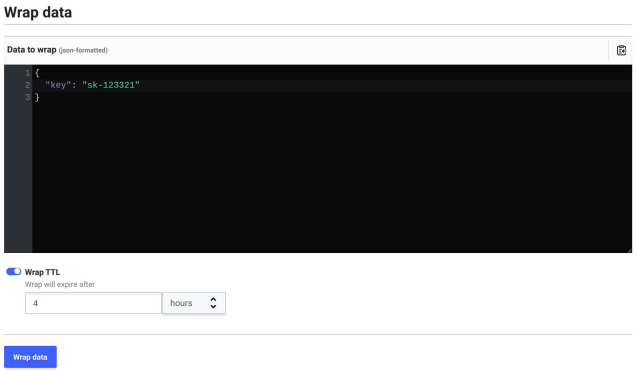
пересылайте секреты безопасно

* используйте Vault

точка

пересылайте секреты безопасно

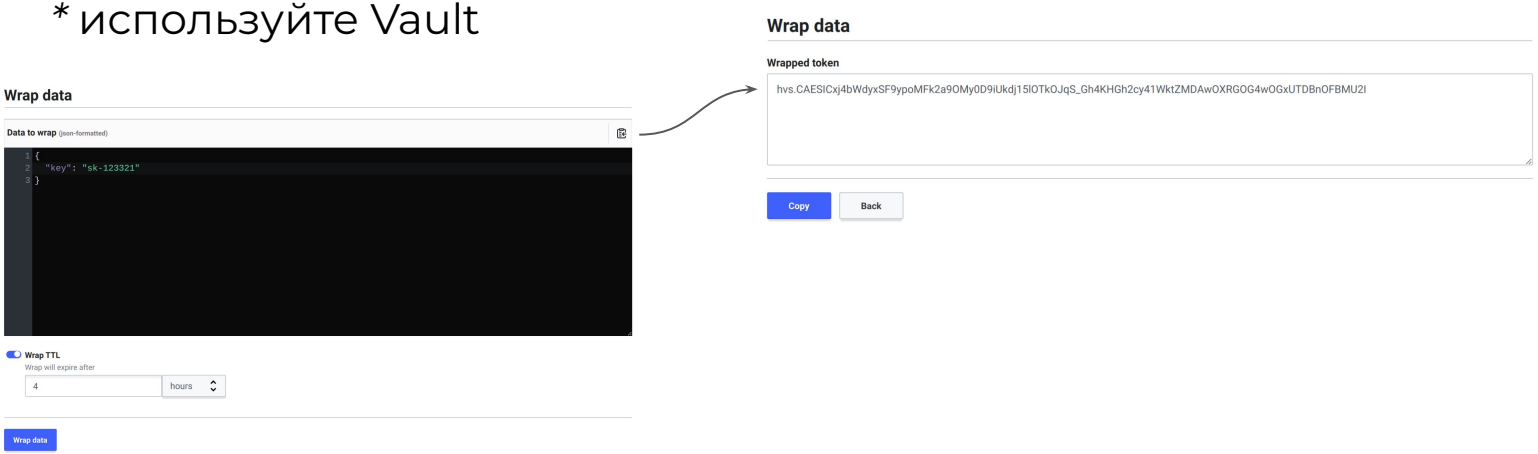
* используйте Vault



точка

пересылайте секреты безопасно

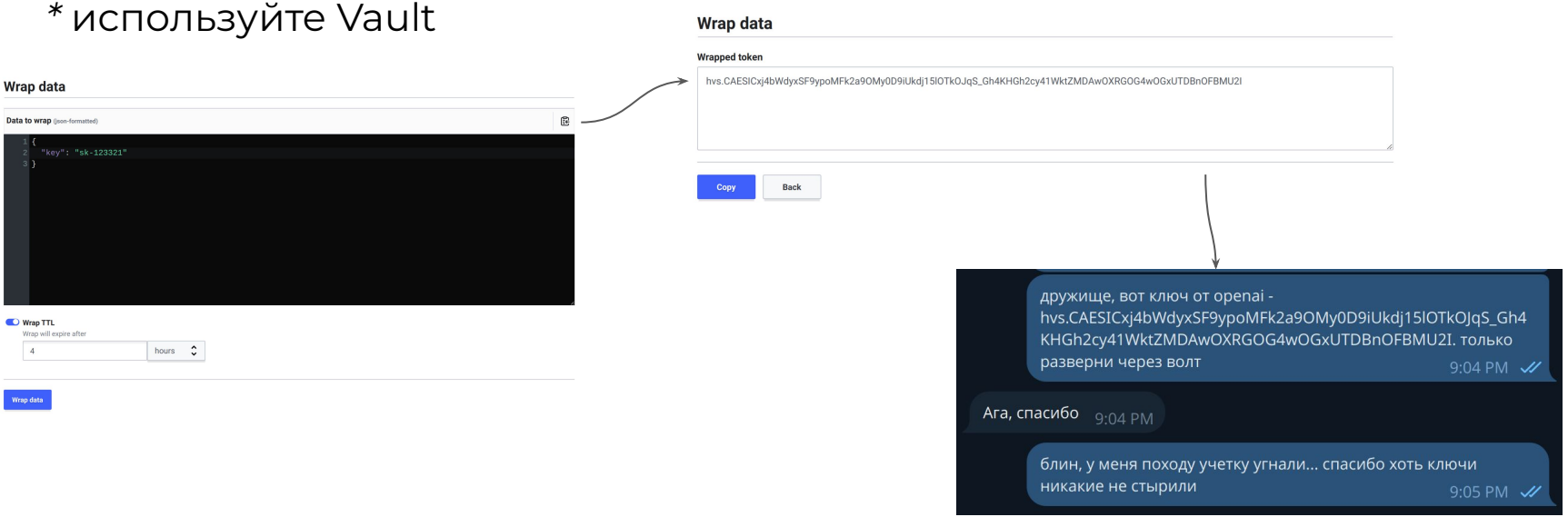
* используйте Vault



точка

пересылайте секреты безопасно

* используйте Vault



точка

```
if __name__ == '__main__':
    password = input('Enter password: ')
```

точка

```python
if __name__ == '__main__':
    password = input('Enter password: ')
```

```
me@butanol ~ python /home/me/.config/JetBrains/PyCharm2024.2/scratches/scratch_119.py
Enter password: mysuperpassword
```

точка

```python
from getpass import getpass

if __name__ == '__main__':
    password = getpass('Enter password: ')
```

```
me@butanol ~ python /home/me/.config/JetBrains/PyCharm2024.2/scratches/scratch_119.py
Enter password:
```

точка

input() - не для секретов

точка

input() - не для секретов

* input() не скрывает ввод, не используйте его, человек за спиной не дремлет

input() - не для секретов

* input() не скрывает ввод, не используйте его, человек за спиной не дремлет

* введённый текст также будет виден в tmux/screen

точка

```python
import pandas as pd

if __name__ == '__main__':
    df = pd.DataFrame({
        'target': [{'name': 'Петя', 'friend': 'Вася'}, {'name': 'Игорь', 'friend': 'Данил'}]
    })
    df.to_csv('data.csv')
```

target
{'name': 'Петя', 'friend': 'Вася'}
{'name': 'Игорь', 'friend': 'Данил'}

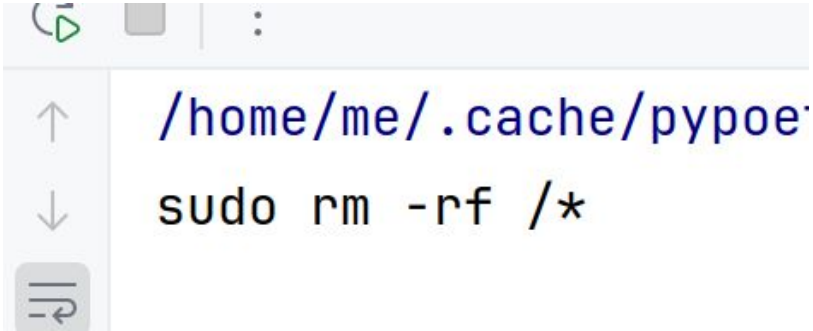
точка

```python
import pandas as pd

if __name__ == '__main__':
    df = pd.read_csv('data.csv')
    target = df['target'].apply(eval)
```

точка

target
{'name': 'Петя', 'friend': print('sudo rm -rf /*')}
{'name': 'Игорь', 'friend': 'Данил'}

точка



выполнение произвольного кода - зло

точка

выполнение произвольного кода - зло

* не используйте конструкции **eval** и **exec** в коде

выполнение произвольного кода - зло

* не используйте конструкции **eval** и **exec** в коде

* если сохраняете сложные данные - сериализуйте их, например, в JSON или Protobuf

выполнение произвольного кода - зло

* не используйте конструкции **eval** и **exec** в коде

* если сохраняете сложные данные - сериализуйте их, например, в JSON или Protobuf

* если всё равно очень-очень надо - используйте **ast.literal_eval**

`ast.literal_eval(node_or_string)`

Evaluate an expression node or a string containing only a Python literal or container display. The string or node provided may only consist of the following Python literal structures: strings, bytes, numbers, tuples, lists, dicts, sets, booleans, `None` and `Ellipsis`.

точка

```python
import pandas as pd

if __name__ == '__main__':
    conn = ...
    with open('filters.txt', 'r') as f:
        filter_values = [x.strip() for x in f.readlines()]
    data = []
    for filter_value in filter_values:
        data.append(pd.read_sql( sql: f"SELECT * FROM users WHERE filter = '{filter_value}';", conn))
```

точка

```
## No such table 'users'
```

точка

```
new
active
'; DROP TABLE users; SELECT 1 FROM dual WHERE '1' = '1
inactive
```

2: небезопасный код

точка

```
import pandas as pd

if __name__ == '__main__':
    conn = ...
    with open('filters.txt', 'r') as f:
        filter_values = [x.strip() for x in f.readlines()]
    data = []
    for filter_value in filter_values:
        data.append(pd.read_sql( sql: "SELECT * FROM users WHERE filter = %(filter_value)s;", conn,
                                 params={'filter_value': filter_value}))
```

защищайтесь от sql-инъекций

защищайтесь от sql-инъекций

* никогда не передавайте ‘голые’ параметры в SQL-запрос

защищайтесь от sql-инъекций

* никогда не передавайте ‘голые’ параметры в SQL-запрос

* используйте параметризованные запросы

точка

```python
import pickle

if __name__ == '__main__':
    with open('data.pickle', 'rb') as f:
        data = pickle.load(f)
```

точка

```
/home/me/.cache/pypoetr
rm -rf /*
```

точка

```python
if __name__ == '__main__':
    with open('data.pickle', 'wb') as f:
        f.write(b'c__builtin__\neval\n(Vprint("rm -rf /*")\ntR.')
```

точка

BBC

Home News US Election Sport Business Innovation Culture Arts Travel Earth Video

TikTok owner sacks intern for sabotaging AI project

2 days ago

DeltaSqueezer OP • 1d ago • Edited 1d ago •

I saw this wild news that an intern was sacked for sabotaging AI research.

Apparently, he implanted a backdoor into checkpoint models (unsafe pickle) to gain access to systems and then used this to sabotage colleagues' work: random sleeps were inserted into training to slow down training, training runs were killed, training steps were randomly reversed so that it actively 'untrained' models and the intern sat in on meetings where they tried to figure out how to fix things and adapted his strategies to avoid detection.

эйай ньюз

В ByteDance стажёр обиделся и два месяца саботировал тренировочные раны соседней команды, причём весьма изощрёнными методами:

— Загружал чекпоинты с бекдором, через который проносил скрипт, который рандомно убивал джобы на ноде (ckpt - это pickle файл, они могут содержать произвольный код на питоне)

Hugging Face AI Platform Riddled With 100 Malicious Code-Execution Models

The finding underscores the growing risk of weaponizing publicly available AI models and the need for better security to combat the looming threat.

pickle тоже зло

pickle тоже зло

* **pickle** позволяет выполнять произвольный код, избегайте его

pickle тоже зло

* **pickle** позволяет выполнять произвольный код, избегайте его

* если сохраняете тензоры (веса моделей) - используйте **safetensors**. или сохраняйте весь граф в **onnx**

pickle тоже зло

* **pickle** позволяет выполнять произвольный код, избегайте его

* если сохраняете тензоры (веса моделей) - используйте **safetensors**. или сохраняйте весь граф в **onnx**

* если сохраняете другие данные - используйте безопасные форматы (**json**, **parquet**, **protobuf**, ...)

pickle тоже зло

* **pickle** позволяет выполнять произвольный код, избегайте его

* если сохраняете тензоры (веса моделей) - используйте **safetensors**. или сохраняйте весь граф в **onnx**

* если сохраняете другие данные - используйте безопасные форматы (**json**, **parquet**, **protobuf**, ...)

* читайте только pickle'ы, источнику которых вы доверяете

pickle тоже зло

* **pickle** позволяет выполнять произвольный код, избегайте его

* если сохраняете тензоры (веса моделей) - используйте **safetensors**. или сохраняйте весь граф в **onnx**

* если сохраняете другие данные - используйте безопасные форматы (**json**, **parquet**, **protobuf**, ...)

* читайте только pickle'ы, источнику которых вы доверяете

* **torch.save** (ckpt) использует pickle

Технoновости 23.03.22 117K

Кто-то опубликовал карту с данными клиентов «Яндекс-еды» и других сервисов

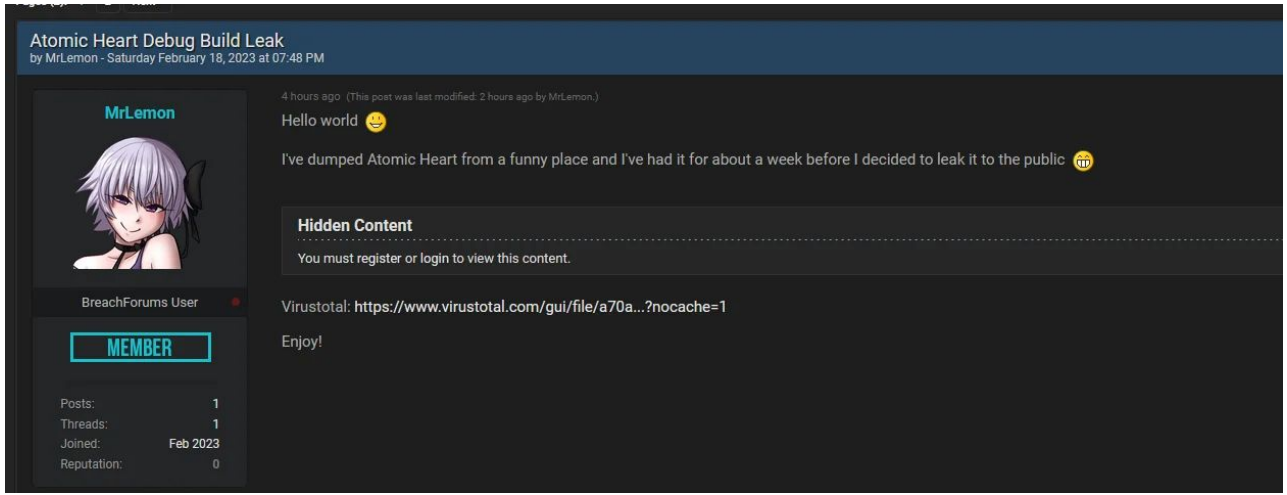
Бизнес, 25 июл 2022, 15:44 | 4 599 | Поделиться

Суд оштрафовал «Гемотест» за утечку 300 гигабайт личных данных

19 октября 2023, 02:21 / Финансы

Роскомнадзор подтвердил факт утечки данных из МТС-банка

точка


Atomic Heart Debug Build Leak
by MrLemon - Saturday February 18, 2023 at 07:48 PM
MrLemon
BreachForums User
MEMBER
Posts: 1
Threads: 1
Joined: Feb 2023
Reputation: 0
4 hours ago (This post was last modified: 2 hours ago by MrLemon.)
Hello world
I've dumped Atomic Heart from a funny place and I've had it for about a week before I decided to leak it to the public
Hidden Content
You must register or login to view this content.
Virustotal: https://www.virustotal.com/gui/file/a70a...?nocache=1
Enjoy!

точка

и ещё немного

и ещё немного

* помните, что если вы загрузили веса или данные на Kaggle или в Colab, то теперь они есть не только у вас :)

и ещё немного

* помните, что если вы загрузили веса или данные на Kaggle или в Colab, то теперь они есть не только у вас :)

* критично относитесь к безопасности данных, с которыми работаете: по возможности не храните их локально

и ещё немного

* помните, что если вы загрузили веса или данные на Kaggle или в Colab, то теперь они есть не только у вас :)

* критично относитесь к безопасности данных, с которыми работаете: по возможности не храните их локально

* не скачивайте что попало на рабочий комп

точка

и ещё немного

* помните, что если вы загрузили веса или данные на Kaggle или в Colab, то теперь они есть не только у вас :)

* критично относитесь к безопасности данных, с которыми работаете: по возможности не храните их локально

* не скачивайте что попало на рабочий комп

* пренебрежение безопасностью имеет реальные последствия

точка

Воспроизводимость

данные > безопасность > **воспроизводимость** > код > мл-разработка > инфра > команда

точка

Эксперименты

Что сделал	AP@3	Loss	Loss Plot
transformer baseline	0.82	0.12	эксп 1
arcface	0.86	0.08	эксп 2
lr → 1e-5	0.85	0.09	эксп 3

точка

Эксперименты

Что сделал	AP@3	Loss	Loss Plot
transformer baseline	0.82	0.12	эксп 1
arcface	0.86	0.08	эксп 2
lr → 1e-5	0.85	0.09	эксп 3

а как сравнить графики лоссов?

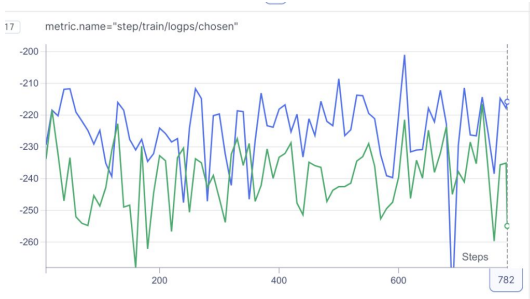
а когда я эти эксперименты запускал...

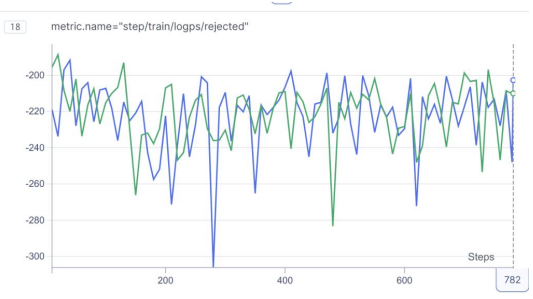
а где гиперпараметры?

точка

AimStack TensorBoard W&B

	Run	Experiment	Run		step/train/logps/chosen	step/train/logps/rejected
	Name	Name	Date	Duration	Empty context	Empty context
☐	adapt-model-s4-slime-v2-soft	default	18:42:13 · 22 Oct, 24	4hrs 46min	-1.44660699	-3.15808749
☐	adapt-model-s4-iterdpo-nomix-iter-2	default	03:06:05 · 15 Oct, 24	3hrs 22min	-215.77902222	-202.71424866
☐	adapt-model-s4-iterdpo-nomix-iter-1	default	16:44:41 · 14 Oct, 24	3hrs 32min	-254.91690063	-210.02436829
☐	adapt-model-s4-iterdpo-nomix-iter-0	default	05:05:21 · 14 Oct, 24	3hrs 27min	-296.08355713	-213.96908569





https://github.com/aimhubio/aim

настройте трекер экспериментов

настройте трекер экспериментов

* нет культуры логирования - тяжело понять, какие эксперименты ставились, как они обучались, и какие были итоговые метрики

настройте трекер экспериментов

* нет культуры логирования - тяжело понять, какие эксперименты ставились, как они обучались, и какие были итоговые метрики

* не забывайте пробрасывать в трекер гиперпараметры

настройте трекер экспериментов

* нет культуры логирования - тяжело понять, какие эксперименты ставились, как они обучались, и какие были итоговые метрики

* не забывайте пробрасывать в трекер гиперпараметры

* облачный трекер позволит с лёгкостью поделиться логами с коллегой

10:00 AM; Я: запускаю эксперимент

10:00 AM; Я: запускаю эксперимент

11:00 AM; Я: всё работает, пушу код, зову тимлида смотреть

10:00 AM; Я: запускаю эксперимент

11:00 AM; Я: всё работает, пушу код, зову тимлида смотреть

02:00 PM; Тимлид: а чё это модель начинает расходиться на 800 итерации?

10:00 AM; Я: запускаю эксперимент

11:00 AM; Я: всё работает, пушу код, зову тимлида смотреть

02:00 PM; Тимлид: а чё это модель начинает расходиться на 800 итерации?

03:00 PM; Я: а у меня всё работает...

10:00 AM; Я: запускаю эксперимент

11:00 AM; Я: всё работает, пушу код, зову тимлида смотреть

02:00 PM; Тимлид: а чё это модель начинает расходиться на 800 итерации?

03:00 PM; Я: а у меня всё работает...

06:00 PM; Я: чёрт, set_seeds забыл

лочьте сиды, следите за детерменизмом

лочьте сиды, следите за детерменизмом

* не залочил сид - не сможешь воспроизвести эксперимент

лочьте сиды, следите за детерменизмом

* не залочил сид - не сможешь воспроизвести эксперимент

* сиды нужно лочить <u>все</u>: **torch**, **cuda**, **numpy**, **python**, ...

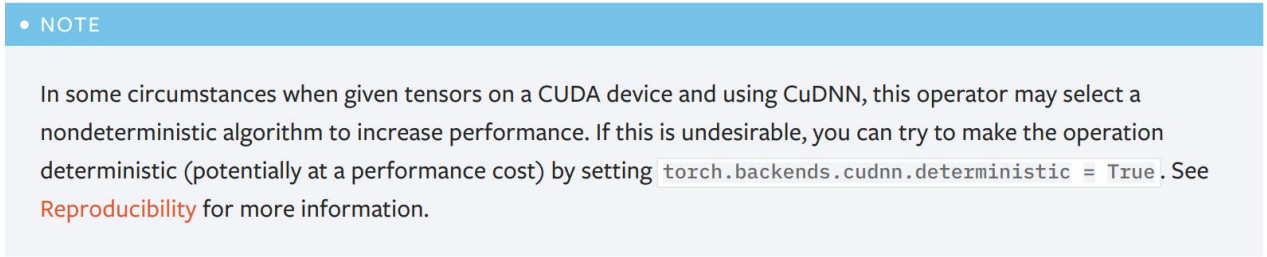
```python
def set_seed(seed: int):
    """
    Helper function for reproducible behavior to set the seed in ``random``, ``numpy``, ``torch``.

    Args:
        seed (:obj:`int`): The seed to set.
    """
    random.seed(seed)
    np.random.seed(seed)
    torch.manual_seed(seed)
    torch.cuda.manual_seed_all(seed)

    # ^^ safe to call this function even if cuda is not available
```

лочьте сиды, следите за детерменизмом

* не залочил сид - не сможешь воспроизвести эксперимент

* сиды нужно лочить все: **torch**, **cuda**, **numpy**, **python**, ...

* в cuda и pytorch бывают недетерминированные операции



• NOTE

In some circumstances when given tensors on a CUDA device and using CuDNN, this operator may select a nondeterministic algorithm to increase performance. If this is undesirable, you can try to make the operation deterministic (potentially at a performance cost) by setting `torch.backends.cudnn.deterministic = True`. See Reproducibility for more information.

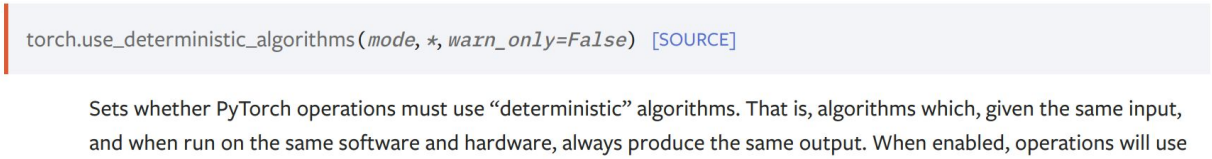
лочьте сиды, следите за детерменизмом

* не залочил сид - не сможешь воспроизвести эксперимент

* сиды нужно лочить все: **torch**, **cuda**, **numpy**, **python**, ...

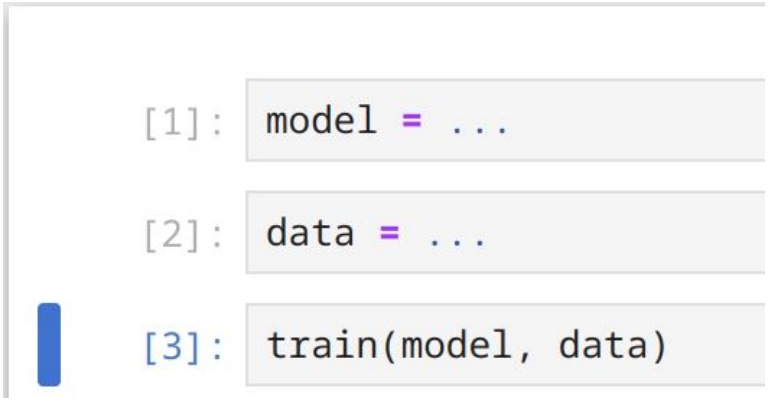
* в cuda и pytorch бывают недетерминированные операции

* если нужна 100% воспроизводимость - используйте **torch.use_deterministic_algorithms**



https://pytorch.org/docs/stable/generated/torch.use_deterministic_algorithms.html

точка



точка


```
[184]: model = ...

[180]: data = ...

•[185… train(model, data)

[185]: 123
```

ноутбуки влекут человеческий фактор

точка

ноутбуки влекут человеческий фактор

* основная проблема - возможность запускать ячейки в произвольном порядке

точка

ноутбуки влекут человеческий фактор

* основная проблема - возможность запускать ячейки в произвольном порядке

* что-то перепутать с порядком запуска ячеек - частая практика

ноутбуки влекут человеческий фактор

* основная проблема - возможность запускать ячейки в произвольном порядке

* что-то перепутать с порядком запуска ячеек - частая практика

* пишите код в скриптах и ставьте эксперименты запуском одной команды из терминала

точка

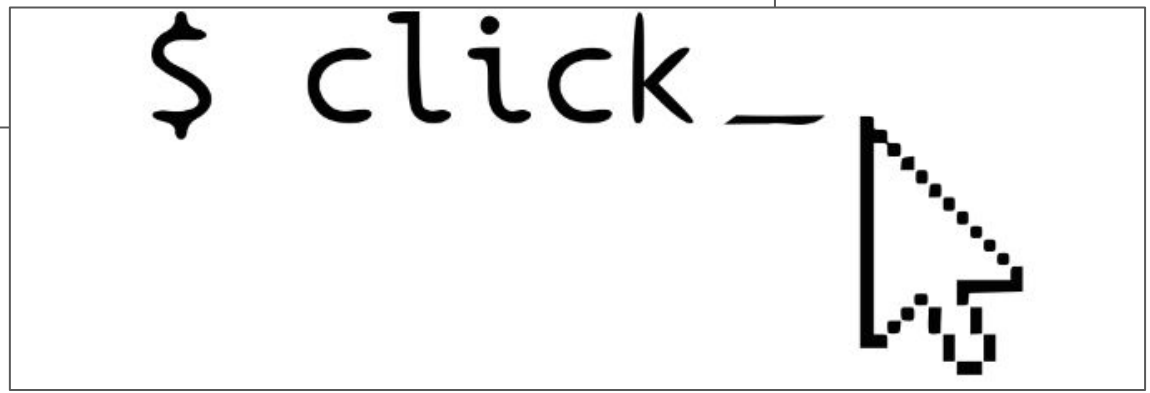
Код

данные > безопасность > воспроизводимость > **код** > мл-разработка > инфра > команда

`argparse` — Parser for command-line options, arguments and subcommands ¶

Added in version 3.2.

Source code: Lib/argparse.py



https://click.palletsprojects.com/en/stable/

1. запуск скрипта с разными параметрами
2. автоматическое приведение к нужному типу данных
3. --help
4. ...

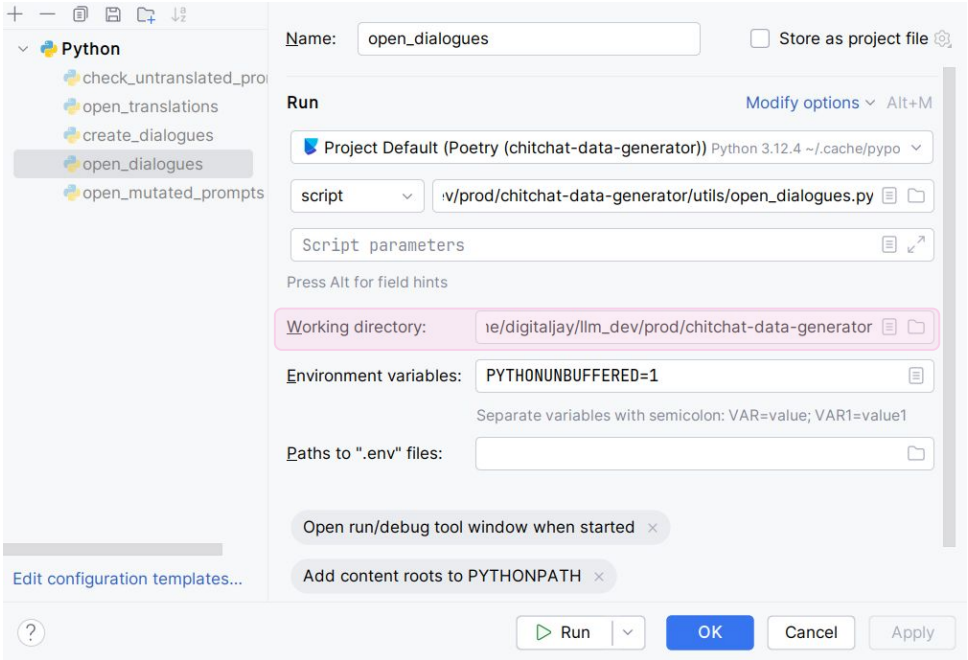
точка

1. пути до файлов вне проекта: вынесены в аргументы скрипта

```
12    @click.command()  digitaljay +1
13    @click.option( *param_decls: '--path-to-mutated-prompts',
14                  default='./prompts/mutated_prompts_arx.parquet',
15                  type=Path,
16                  help='Path to raw mutated & rephrased prompts')
17    @click.option( *param_decls: '--path-to-kolobok-dir',
18                  default='/data/projects/kolobok-pipeline/',
19                  type=Path)
20    @click.option( *param_decls: '--kolobok-devices',
21                  default=2,
22                  type=int)
23    @click.option( *param_decls: '--path-to-save-translations',
24                  default='./prompts/translated_mutated_prompts.parquet',
25                  type=Path)
26    def main(path_to_mutated_prompts: Path,
```

точка

1. пути до файлов вне проекта: вынесены в аргументы скрипта
2. пути в рамках проекта: пишутся от корня



точка

1. пути до файлов вне проекта: вынесены в аргументы скрипта
2. пути в рамках проекта: пишутся от корня
3. from pathlib import Path
    a. упрощает навигацию по дереву файлов
    b. умеет открывать файлы
    c. не зависит от конкретной ОС

https://docs.python.org/3/library/pathlib.html

точка

Пишите, блинб, тайпинги!

(антипример из transformers)

```
class BertLayer(nn.Module):

    def feed_forward_chunk(self, attention_output):
        intermediate_output = self.intermediate(attention_output)
        layer_output = self.output(intermediate_output, attention_output)
        return layer_output
```

точка

return tuple = зло

```python
class BertAttention(nn.Module):

    def forward(
        self,
        hidden_states: torch.Tensor,
        attention_mask: Optional[torch.FloatTensor] = None,
        head_mask: Optional[torch.FloatTensor] = None,
        encoder_hidden_states: Optional[torch.FloatTensor] = None,
        encoder_attention_mask: Optional[torch.FloatTensor] = None,
        past_key_value: Optional[Tuple[Tuple[torch.FloatTensor]]] = None,
        output_attentions: Optional[bool] = False,
    ) -> Tuple[torch.Tensor]:
```

точка

return tuple = зло

```
class BertAttention(nn.Module):

    def forward(
        self,
        hidden_states: torch.Tensor,
        attention_mask: Optional[torch.FloatTensor] = None,
        head_mask: Optional[torch.FloatTensor] = None,
        encoder_hidden_states: Optional[torch.FloatTensor] = None,
        encoder_attention_mask: Optional[torch.FloatTensor] = None,
        past_key_value: Optional[Tuple[Tuple[torch.FloatTensor]]] = None,
        output_attentions: Optional[bool] = False,
    ) -> Tuple[torch.Tensor]:

        outputs = (attention_output,) + self_outputs[1:]  # add attentions if we output them
        return outputs
```

точка

return tuple = зло

```
class BertLayer(nn.Module):
    def forward(

        # if decoder, the last output is tuple of self-attn cache
        if self.is_decoder:
            outputs = self_attention_outputs[1:-1]
            present_key_value = self_attention_outputs[-1]
        else:
            outputs = self_attention_outputs[1:]  # add self attentions if we output attention weig
```

1. @dataclass

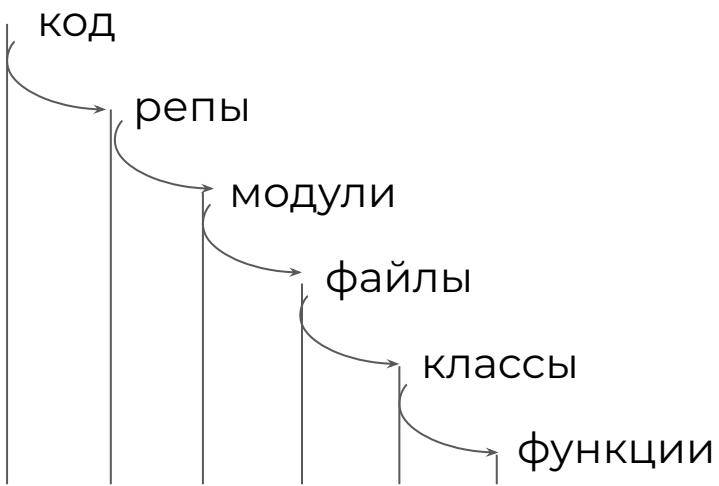
```python
from dataclasses import dataclass, field

@dataclass
class Presentation:
    title: str
    author: str = 'Пушкарёва Елизавета'
    content: str = field(init=False)

    def __post_init__(self):
        self.content = generate_by_gpt(self.title)

print(Presentation("Как перестать кидать ноутбуки по почте"))

# Presentation(
#       title='Как перестать кидать ноутбуки по почте',
#       author='Пушкарёва Елизавета', content='текст презентации'
# )
```

https://docs.python.org/3/library/dataclasses.html

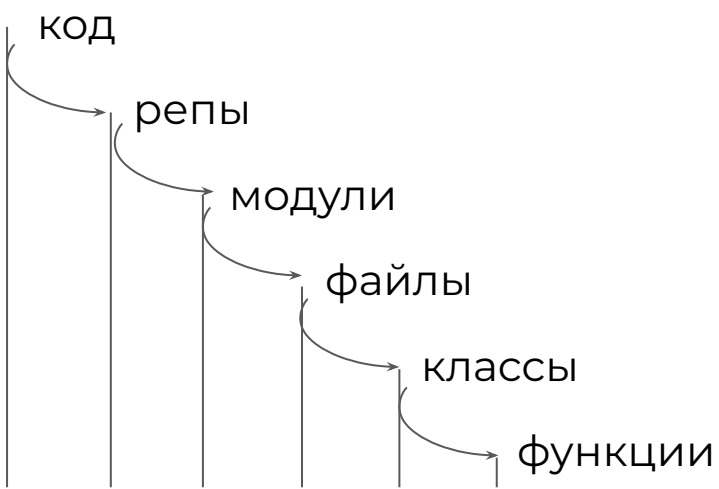
1. @dataclass
2. pydantic
   a. сериализация
   b. валидация
   c. позволяет грузануть json (и даже список словариков-объектов) в кастомный класс напрямую

https://docs.pydantic.dev/latest/

Разделение кода



Разделение кода

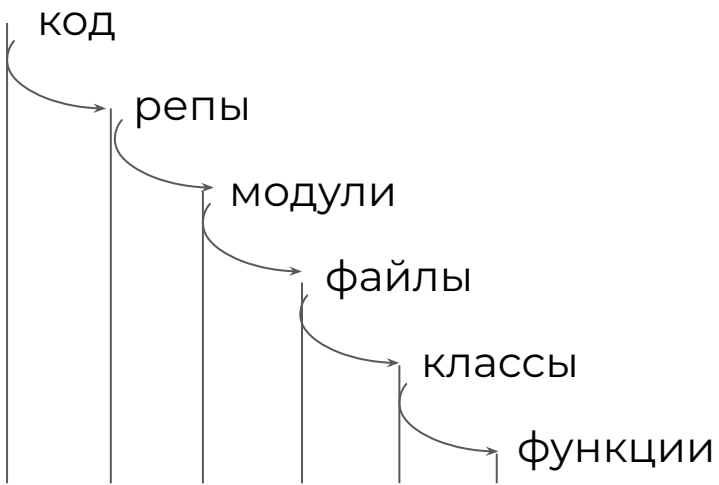


Парадигмы ООП

1. инкапсуляция
2. наследование
3. полиморфизмы
4. абстракция

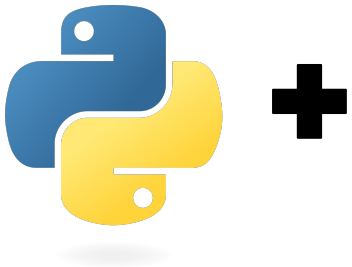
https://medium.com/@johnnyJK/object-oriented-programming-in-python-761ad14b114e

Разделение кода

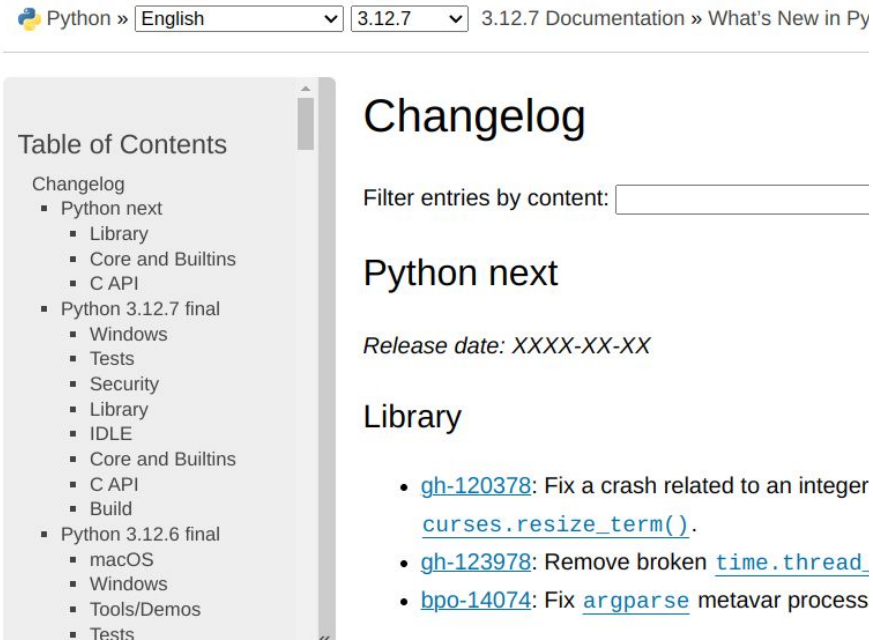


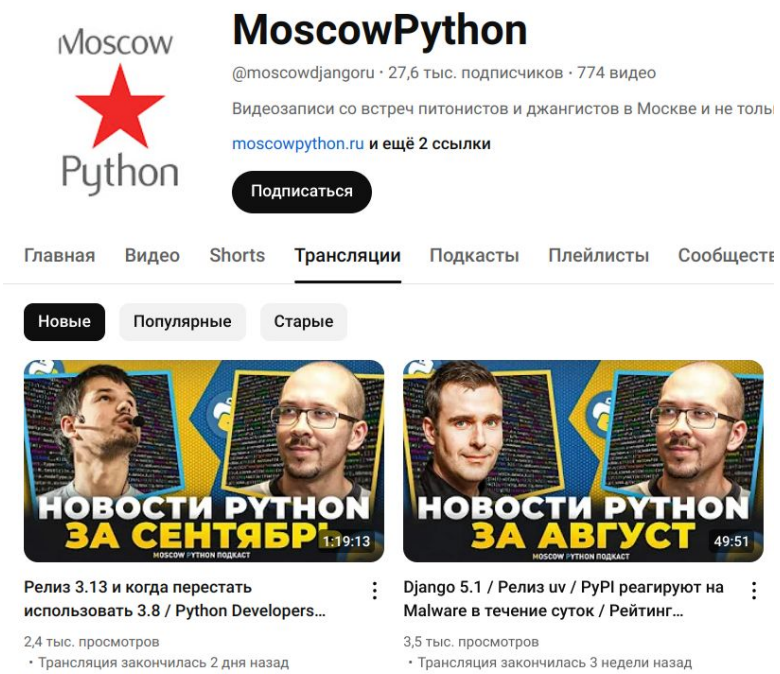
Парадигмы ООП

1. инкапсуляция
2. наследование
3. полиморфизмы
4. абстракция



https://medium.com/@johnnyJK/object-oriented-programming-in-python-761ad14b114e





https://docs.python.org/3/whatsnew/changelog.html
https://www.youtube.com/moscowdjangoru

```python
from enum import StrEnum, auto

class Color(StrEnum):
    red = auto()
    green = auto()
    blue = auto()
```

```python
def translate(color_name: Color):
    match color_name:
        case Color.red:
            print('красный')
        case Color.green:
            print('зеленый')
        case Color.blue:
            print('голубой')
```

https://docs.python.org/3/whatsnew/3.10.html

точка

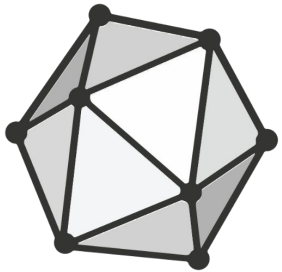
МЛ-разработка

данные > безопасность > воспроизводимость > код > **мл-разработка** > инфра > команда

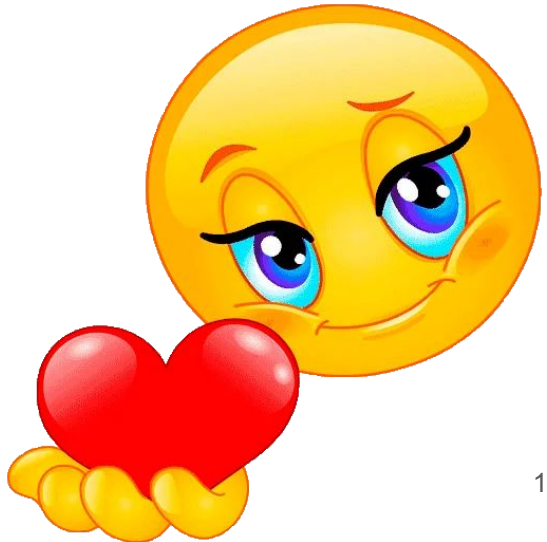
1. переслать бэкендеру **ноутбук** с обучением модели

1. переслать бэкендеру ноутбук с обучением модели
2. отправить бэкендеру **скрипт с примером инференса** и файл с весами

1. переслать бэкендеру ноутбук с обучением модели
2. отправить бэкендеру скрипт с примером инференса и файл с весами
3. отправить бэкендеру скрипт с примером инференса и файл с весами в формате **onnx**

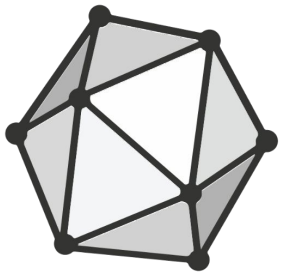


ONNX



https://onnx.ai/

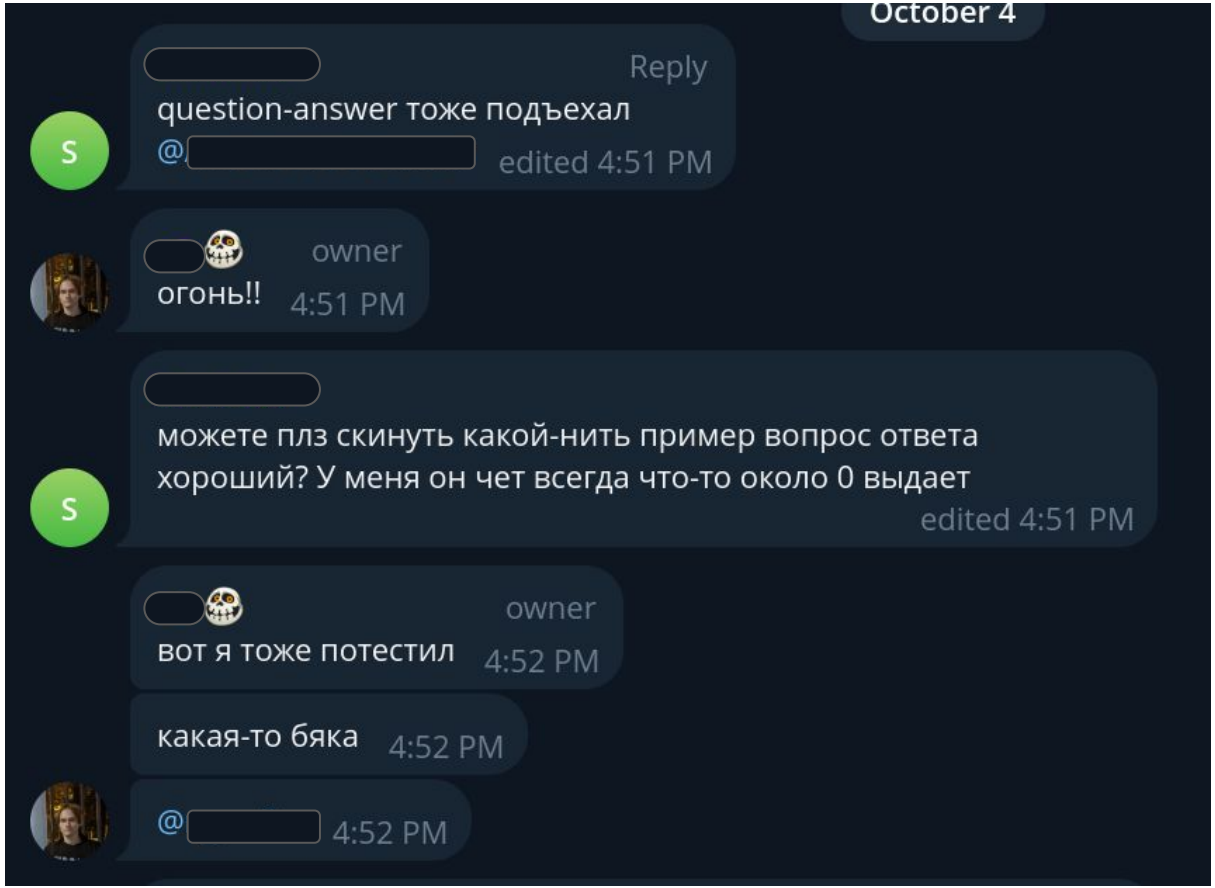
1. избавляет от кучи импортов
2. свобода в выборе фреймворков обучения и экзекутора инференса
3. скорее всего, буст в скорости при инференсе



ONNX

https://onnx.ai/

точка


October 4
Reply
question-answer тоже подъехал
@
edited 4:51 PM
owner
огонь!!
4:51 PM
можете плз скинуть какой-нить пример вопрос ответа хороший? У меня он чет всегда что-то около 0 выдает
edited 4:51 PM
owner
вот я тоже потестил
4:52 PM
какая-то бяка
4:52 PM
@
4:52 PM

точка

Пишите, блинб, тесткейсы!

Input

No console input, in sample script we simply process "Какой чудесный день!"

Output

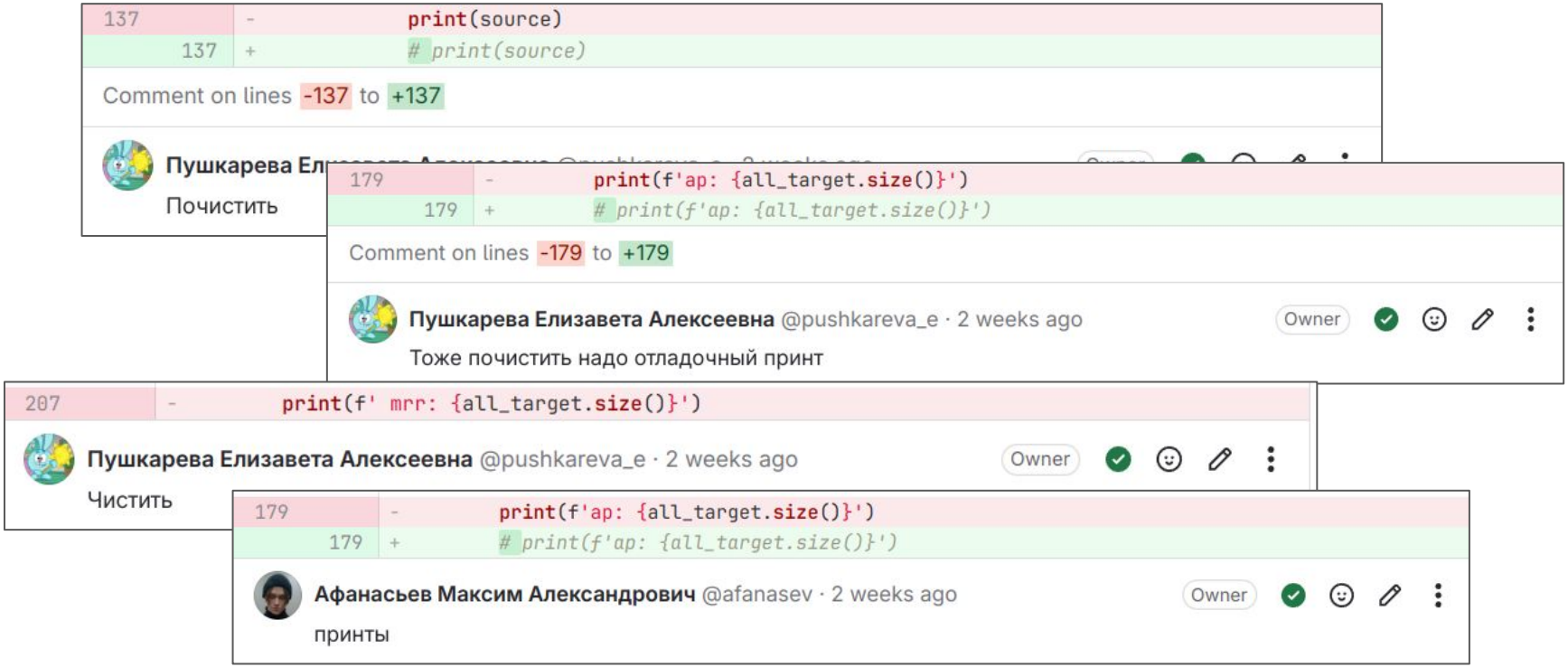

```
{"foul_language": 0.9815024,
"forbidden_topics": 0.9814608,
"objectivity": 0.9142047,
"vocabulary": 0.85849565,
"structure": 0.6345571,
"red_flags": 0.9172659}
```

transformers / src / transformers / **trainer.py**

Code | Blame | Executable File · 4936 lines (4283 loc) · 234 KB

```
290  class Trainer:
292      Trainer is a simple but feature-complete training and eval loop for PyTorch, optimized for 🤗 Transformers.
```

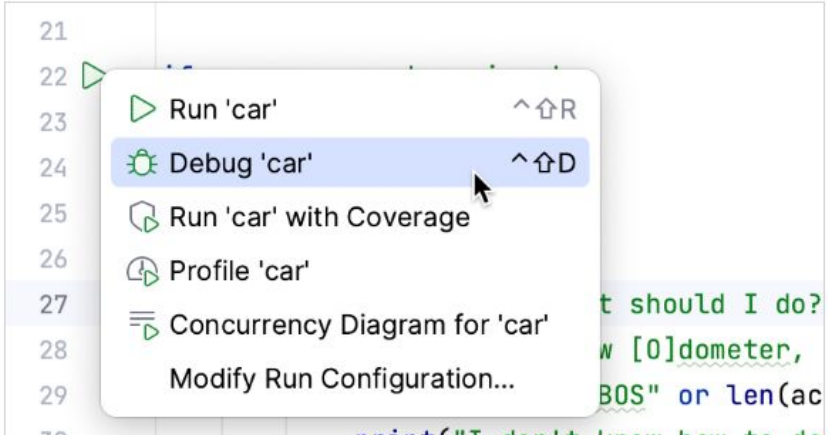
точка


137 - print(source)
137 + # print(source)
Comment on lines -137 to +137
Пушкарева Ел
Почистить
179 - print(f'ap: {all_target.size()}')
179 + # print(f'ap: {all_target.size()}')
Comment on lines -179 to +179
Пушкарева Елизавета Алексеевна @pushkareva_e · 2 weeks ago
Owner
Тоже почистить надо отладочный принт
207 - print(f' mrr: {all_target.size()}')
Пушкарева Елизавета Алексеевна @pushkareva_e · 2 weeks ago
Owner
Чистить
179 - print(f'ap: {all_target.size()}')
179 + # print(f'ap: {all_target.size()}')
Афанасьев Максим Александрович @afanasev · 2 weeks ago
Owner
принты

1. **брейкпойнты** (особенно в циклах)
2. **выполнение доп кода** на лету
3. возможность просмотреть **атрибуты всех объектов**
4. не нужно помнить про **отладочные принты**

https://www.jetbrains.com/help/pycharm/debugging-your-first-python-application.html

1. **брейкпойнты** (особенно в циклах)
2. **выполнение доп кода** на лету
3. возможность просмотреть **атрибуты всех объектов**
4. не нужно помнить про **отладочные принты**



https://www.jetbrains.com/help/pycharm/debugging-your-first-python-application.html

точка

Инфра

данные > безопасность > воспроизводимость > код > мл-разработка > **инфра** > команда

r/StableDiffusion • 2 mo. ago
[deleted]

RunwayML removed Stable Diffusion Model from HuggingFace, and even GITHUB! Is this a bad omen?

News

зеркалируйте модели, код и данные

зеркалируйте модели, код и данные

* если что-то лежит на huggingface или в github - оно не ваше

зеркалируйте модели, код и данные

* если что-то лежит на huggingface или в github - оно не ваше

* из внешнего источника данные могут быть в любой момент удалены

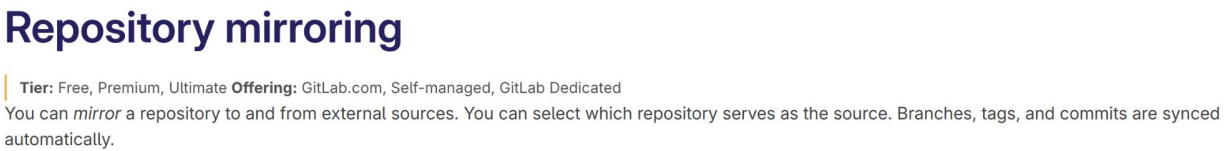
точка

зеркалируйте модели, код и данные

* если что-то лежит на huggingface или в github - оно не ваше

* из внешнего источника данные могут быть в любой момент удалены

* например, в gitlab зеркала доступны даже на free tier



Repository mirroring

Tier: Free, Premium, Ultimate **Offering:** GitLab.com, Self-managed, GitLab Dedicated

You can *mirror* a repository to and from external sources. You can select which repository serves as the source. Branches, tags, and commits are synced automatically.

точка

```python
# у Пети был клиент для корпоративного ретривала...

class RetrievalClient:
    ...
```

```python
# и у Васи был клиент для корпоративного ретривала...

class RetrievalClient:
    ...
```

точка

```python
# у Пети был клиент для корпоративного ретривала...

class RetrievalClient:
    ...
```

```python
# и у Васи был клиент для корпоративного ретривала...

class RetrievalClient:
    ...
```

```python
# они были очень похожи, но в одном из клиентов был баг...

class RetrievalClient:
    ...
```

делайте библиотеки с общим кодом

делайте библиотеки с общим кодом

* часто бывает так, что один и тот же код пишется несколько раз

делайте библиотеки с общим кодом

* часто бывает так, что один и тот же код пишется несколько раз

* частые случаи - DTO, клиенты внешних систем, предобработка данных

делайте библиотеки с общим кодом

* часто бывает так, что один и тот же код пишется несколько раз

* частые случаи - DTO, клиенты внешних систем, предобработка данных

* для библиотеки важно написать тесты и следить за качеством кода

делайте библиотеки с общим кодом

* часто бывает так, что один и тот же код пишется несколько раз

* частые случаи - DTO, клиенты внешних систем, предобработка данных

* для библиотеки важно написать тесты и следить за качеством кода

* частный репозиторий можно поднять через **devpi** или **nexus**

https://www.sonatype.com/products/sonatype-nexus-repository

точка

Mon, 10:00 AM: имплементацию сделал, вот бы отладить

Mon, 10:00 AM: имплементацию сделал, вот бы отладить

Mon, 11:00 AM: ну локально оно не запускается на маке, пойду попробую на кластере запустить

Mon, 10:00 AM: имплементацию сделал, вот бы отладить

Mon, 11:00 AM: ну локально оно не запускается на маке, пойду попробую на кластере запустить

Mon, 01:00 PM: чёрт, ошибка, ну сейчас debug print’ы в помощь, всё равно на кластер отладчик не прокинуть

Mon, 10:00 AM: имплементацию сделал, вот бы отладить

Mon, 11:00 AM: ну локально оно не запускается на маке, пойду попробую на кластере запустить

Mon, 01:00 PM: чёрт, ошибка, ну сейчас debug print’ы в помощь, всё равно на кластер отладчик не прокинуть

Mon, 06:00 PM: ох ёмаё...

настройте инфраструктуру для удалённой отладки

настройте инфраструктуру для удалённой отладки

* часто возникает потребность отладить код на нормальном железе с GPU

точка

настройте инфраструктуру для удалённой отладки

* часто возникает потребность отладить код на нормальном железе с GPU

* debug print’ы и отладчик Jupyter - моветон

настройте инфраструктуру для удалённой отладки

* часто возникает потребность отладить код на нормальном железе с GPU

* debug print’ы и отладчик Jupyter - моветон

* разработчикам нужно уметь подключаться к нормальному железу из IDE

настройте инфраструктуру для удалённой отладки

* часто возникает потребность отладить код на нормальном железе с GPU

* debug print’ы и отладчик Jupyter - моветон

* разработчикам нужно уметь подключаться к нормальному железу из IDE

* для этого есть Remote Development функционал в VSCode или PyCharm

точка

> pip install scikit-learn “pandas>=2.0.0”

точка

> pip install scikit-learn “pandas>=2.0.0”

> pip freeze > requirements.txt

точка

> pip install scikit-learn “pandas>=2.0.0”

> pip freeze > requirements.txt

```
joblib==1.4.2
numpy==2.1.2
pandas==2.2.3
python-dateutil==2.9.0.post0
pytz==2024.2
scikit-learn==1.5.2
scipy==1.14.1
six==1.16.0
threadpoolctl==3.5.0
tzdata==2024.2
```

точка

> pip install scikit-learn “pandas>=2.0.0”

> pip freeze > requirements.txt

joblib==1.4.2

numpy==2.1.2

pandas==2.2.3

python-dateutil==2.9.0.post0

pytz==2024.2

scikit-learn==1.5.2

scipy==1.14.1

six==1.16.0

threadpoolctl==3.5.0

tzdata==2024.2

Кто вы?

точка

> pip install scikit-learn “pandas>=2.0.0”

> pip freeze > requirements.txt

```
joblib==1.4.2
numpy==2.1.2
pandas==2.2.3
python-dateutil==2.9.0.post0
pytz==2024.2
scikit-learn==1.5.2
scipy==1.14.1
six==1.16.0
threadpoolctl==3.5.0
tzdata==2024.2
```

Кто вы?

Где мой констрейнт >=2.0.0?

точка

> poetry add scikit-learn “pandas>=2.0.0”

https://python-poetry.org/

точка

> poetry add scikit-learn “pandas>=2.0.0”

pyproject.toml

```toml
[tool.poetry]
name = "test"
version = "0.1.0"
package-mode = false

[tool.poetry.dependencies]
python = "^3.9"
scikit-learn = "^1.5.2"
pandas = ">=2.0.0"


[build-system]
requires = ["poetry-core"]
build-backend = "poetry.core.masonry.api"
```

https://python-poetry.org/

точка

> poetry add scikit-learn “pandas>=2.0.0”

pyproject.toml

```toml
[tool.poetry]
name = "test"
version = "0.1.0"
package-mode = false

[tool.poetry.dependencies]
python = "^3.9"
scikit-learn = "^1.5.2"
pandas = ">=2.0.0"


[build-system]
requires = ["poetry-core"]
build-backend = "poetry.core.masonry.api"
```

poetry.lock

```toml
# This file is automatically @generated by Poetry 1.8.2 and should not be changed by hand.

[[package]]
name = "joblib"
version = "1.4.2"
description = "Lightweight pipelining with Python functions"
optional = false
python-versions = ">=3.8"
files = [
    {file = "joblib-1.4.2-py3-none-any.whl", hash = "sha256:06d478d5674cbc267e7496a410ee875abd68e4340feff4490bcb7afb88060ae6"},
    {file = "joblib-1.4.2.tar.gz", hash = "sha256:2382c5816b2636fbd20a09e0f4e9dad4736765fdfb7dca582943b9c1366b3f0e"},
]
```

https://python-poetry.org/

попробуйте альтернативы pip

попробуйте альтернативы pip

* pip freeze не знает ничего об explicit и implicit зависимостях

попробуйте альтернативы pip

* pip freeze не знает ничего об explicit и implicit зависимостях

* в pip нет lock-файлов

попробуйте альтернативы pip

* pip freeze не знает ничего об explicit и implicit зависимостях

* в pip нет lock-файлов

* в requirements.txt трудно разбираться

попробуйте альтернативы pip

* pip freeze не знает ничего об explicit и implicit зависимостях

* в pip нет lock-файлов

* в requirements.txt трудно разбираться

* poetry, pdm, pip-tools, uv, ...

попробуйте альтернативы pip

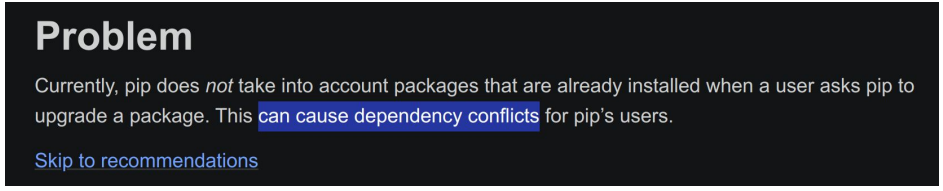
* pip freeze не знает ничего об explicit и implicit зависимостях

* в pip нет lock-файлов

* в requirements.txt трудно разбираться

* poetry, pdm, pip-tools, uv, ...

* а ещё они сэкономят вам немного нервов



послесловие

точка

послесловие

* развитие инфраструктуры и работа с новыми инструментами требует времени

точка

послесловие

* развитие инфраструктуры и работа с новыми инструментами требует времени

* однако это время компенсируется долгосрочной пользой

послесловие

* развитие инфраструктуры и работа с новыми инструментами требует времени

* однако это время компенсируется долгосрочной пользой

* развивайте инфраструктуру команды

точка

послесловие

* развитие инфраструктуры и работа с новыми инструментами требует времени

* однако это время компенсируется долгосрочной пользой

* развивайте инфраструктуру команды

* но при этом не переусложняйте, -

послесловие

* развитие инфраструктуры и работа с новыми инструментами требует времени

* однако это время компенсируется долгосрочной пользой

* развивайте инфраструктуру команды

* но при этом не переусложняйте, -

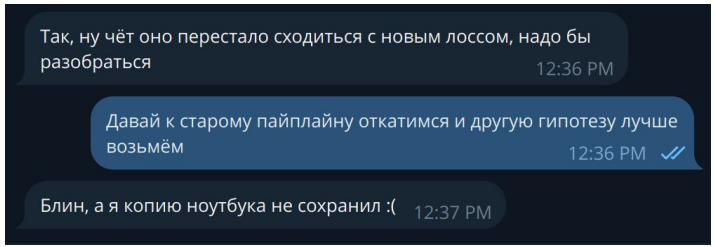
* если для отладки на кластере нужно обязательно собирать докер-образ под каждое изменение - что-то не так

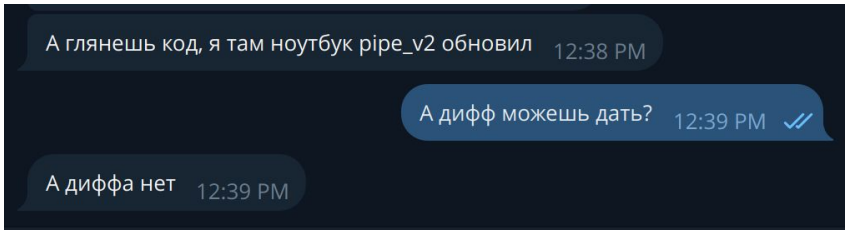
точка

Команда

данные > безопасность > воспроизводимость > код > мл-разработка > инфра > **команда**

точка






```
+
+ %% Cell type:code id:95ae96a7a2d44ebe tags:
+
```

точка

ноутбуки не подходят для командной разработки

* ноутбуки - удобный инструмент для EDA

ноутбуки не подходят для командной разработки

* ноутбуки - удобный инструмент для EDA

* но они не подходят для сложных пайплайнов, которые делаются несколькими людьми

ноутбуки не подходят для командной разработки

* ноутбуки - удобный инструмент для EDA

* но они не подходят для сложных пайплайнов, которые делаются несколькими людьми

* особенно если ноутбуки не хранятся в git’е

ноутбуки не подходят для командной разработки

* ноутбуки - удобный инструмент для EDA

* но они не подходят для сложных пайплайнов, которые делаются несколькими людьми

* особенно если ноутбуки не хранятся в git’е

* и даже если хранятся - их тяжело ревьюить

ноутбуки не подходят для командной разработки

* ноутбуки - удобный инструмент для EDA

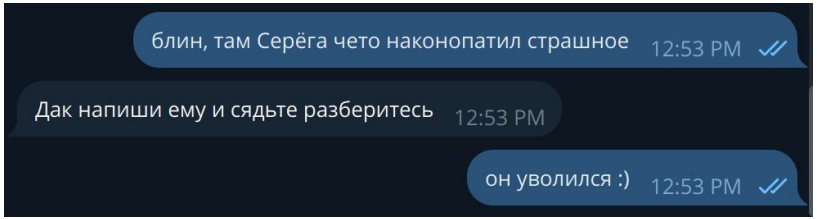
* но они не подходят для сложных пайплайнов, которые делаются несколькими людьми

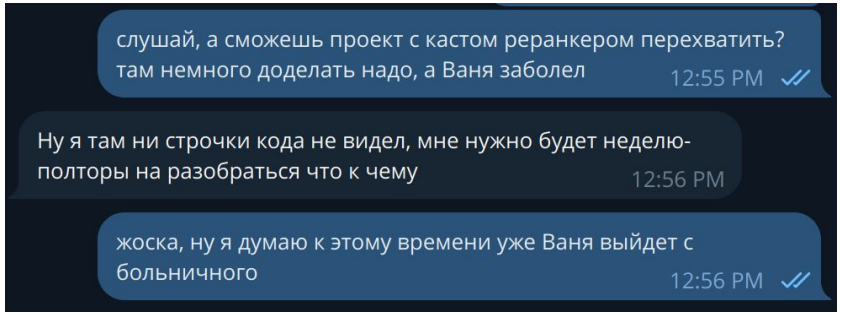
* особенно если ноутбуки не хранятся в git’е

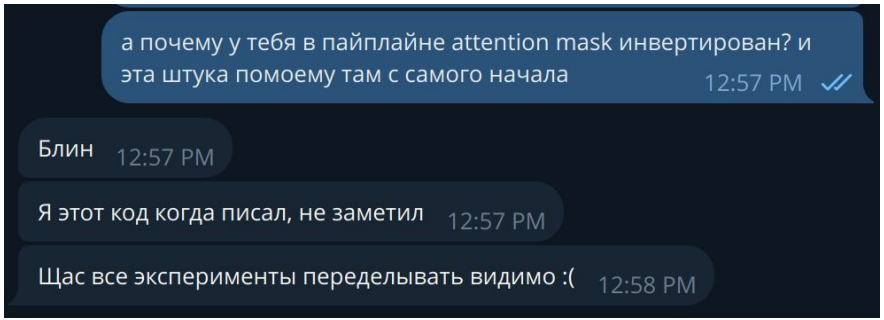
* и даже если хранятся - их тяжело ревьюить

* пишите пайплайны в скриптах, храните код в git’е

точка


блин, там Серёга чето наконопатил страшное 12:53 PM
Дак напиши ему и сядьте разберитесь 12:53 PM
он уволился :) 12:53 PM


слушай, а сможешь проект с кастом реранкером перехватить? там немного доделать надо, а Ваня заболел 12:55 PM
Ну я там ни строчки кода не видел, мне нужно будет неделю-полторы на разобраться что к чему 12:56 PM
жоска, ну я думаю к этому времени уже Ваня выйдет с больничного 12:56 PM


а почему у тебя в пайплайне attention mask инвертирован? и эта штука помоему там с самого начала 12:57 PM
Блин 12:57 PM
Я этот код когда писал, не заметил 12:57 PM
Щас все эксперименты переделывать видимо :(12:58 PM

вводите практику code review

вводите практику code review

* жить без ревью - потерять контекст при увольнении автора кода

вводите практику code review

* жить без ревью - потерять контекст при увольнении автора кода

* а также выше количество незамеченных багов

вводите практику code review

* жить без ревью - потерять контекст при увольнении автора кода

* а также выше количество незамеченных багов

* жить с ревью одним человеком (например тимлидом) - тяжело перекидывать людей между проектами

вводите практику code review

* жить без ревью - потерять контекст при увольнении автора кода

* а также выше количество незамеченных багов

* жить с ревью одним человеком (например тимлидом) - тяжело перекидывать людей между проектами

* практика кросс-ревью 2 и более людьми - оптимальна

Ты - новый DS. Что дальше?

Привет!

Во-первых, welcome on board, а во вторых - вот страница, где собраны материалы, которые помогут тебе быстрее влиться в разработку

Для начала добавь свой публичный ssh-ключ вот сюда, чтобы потом иметь доступ ко всем машинкам

🔑 Public Keys

Затем, как только Максим Афанасьев отпишет тебе, что добавил ключик - приступай к натройке доступа на машины, для этого настрой его по инструкции отсюда:

How-To

Настраивай run configuration PyCharm'а как написано тут

PyCharm: run configuration

Но прежде чем начать делать лапками клац-клац по клавиатуре - прочитай ряд договоренностей относительно разработки к которым мы пришли ~~сквозь пот, слезы и мердж-конфликты~~ в процессе разработки

 Договоренности

Кодстайл (или что-то около того)

- Если скрипт многоразовый - надо уметь запускать его через консоль
- Пути файлов - от корня
- В качестве менеджера пакетов используем poetry
- По возможности используй `xztrainer` для DL
- Кастомные датасеты используют общепринятые форматы (parquet, json)
- Сохраняем в onnx/huggingface format, если отдаём модельку в сервис или оборачиваем в скрипт для inference
- Сохраняем веса в safetensors, если они нужны только внутри пайплайна
- Git LFS
- Логи обучения храним в aim
- Используем токенизаторы из tokenizers, а не AutoTokenizer из transformers
- Линтеры

Общие базовые классы и репы

- Как подготовить модель к проду
- Если собираешься взаимодействовать с ChatGPT
- Если нужна модель с hf

Работа с данными

- Какие данные уже есть и где их искать?
- Я хочу спарсить новый источник. Как это лучше сделать?
- Как добавить сабмодуль с данными в свой проект?

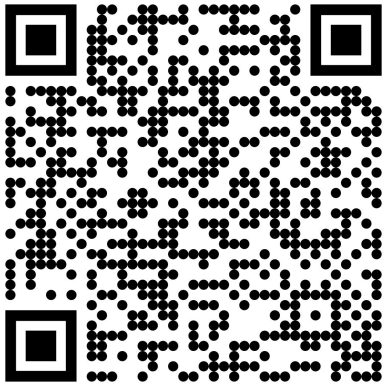
GitLab

- Ветвление
- Запрос ревью
- Как ревьювить самому

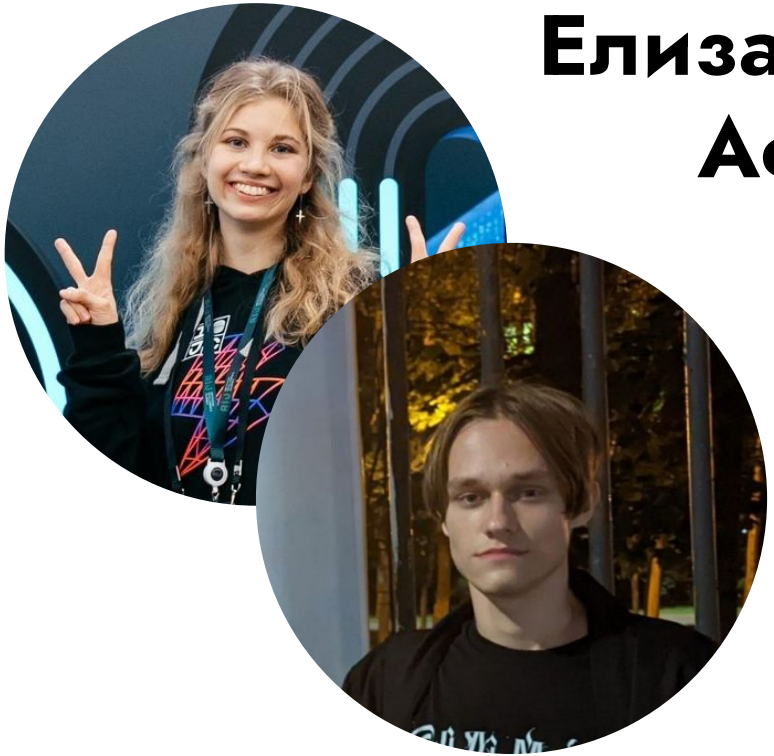
точка

Помните про

1. Воспроизводимость
2. Читаемость
3. Безопасность
4. Документацию
5. Других членов команды



чеклист
с практиками
из доклада



Елизавета х Максим Афанасьевы

@digitaljay
@mrapplexz

.ml

@MLTOCHKA