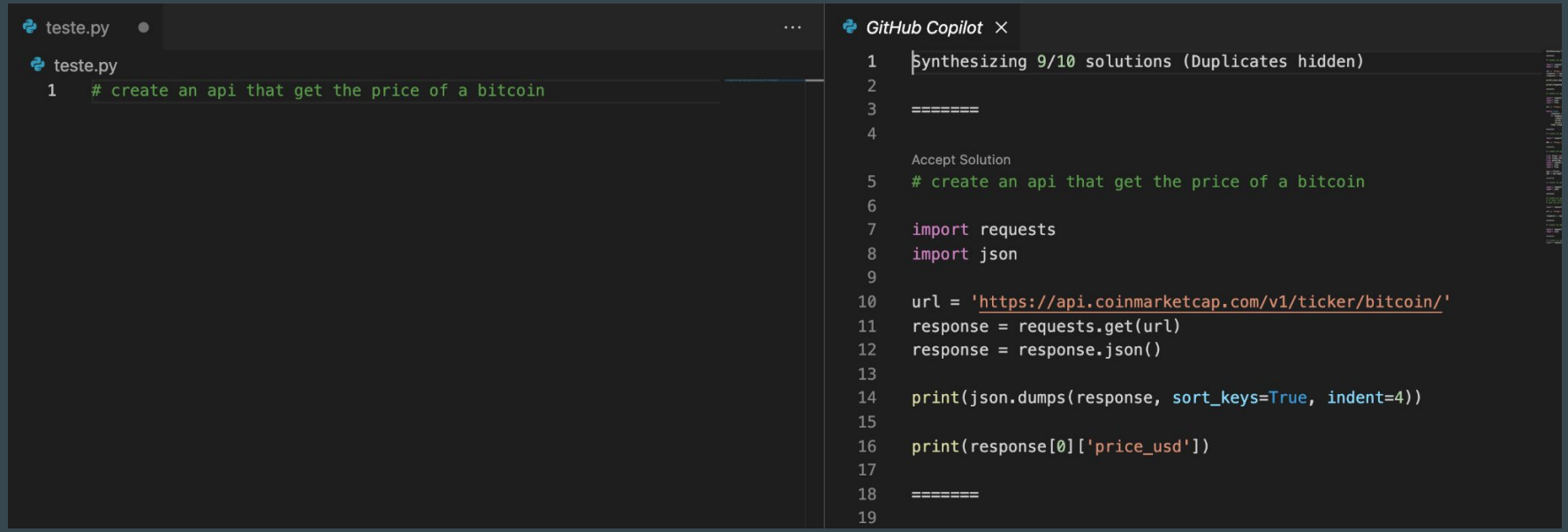
Github CoPilot

...

Gabriel Zezze
Pedro Luiz
Vitor Eller
Desenvolvimento Aberto - 2021/2

O que é ?

- AI Pair Programming - Parceiro de desenvolvimento possibilitado por Inteligência Artificial.
- Foi transformado em uma extensão para o editor de texto Visual Studio Code, recebendo instruções em forma de comentários e a partir do código já desenvolvido.
- A partir das instruções recebidas, ele infere um snippet de código apropriado, se esse snippet não for a melhor solução ele mostra outras 10 alternativas.
- Atualmente está em Technical preview. É possível se aplicar de graça para ter a extensão (experiência pessoal: demoram aproximadamente 3 semanas para ter uma resposta).


teste.py
teste.py
1 # create an api that get the price of a bitcoin
GitHub Copilot
1 Synthesizing 9/10 solutions (Duplicates hidden)
2
3 =======
4
Accept Solution
5 # create an api that get the price of a bitcoin
6
7 import requests
8 import json
9
10 url = 'https://api.coinmarketcap.com/v1/ticker/bitcoin/'
11 response = requests.get(url)
12 response = response.json()
13
14 print(json.dumps(response, sort_keys=True, indent=4))
15
16 print(response[0]['price_usd'])
17
18 =======
19

isso te dá medo?


```
//i dont know how it works but it works
```

```
//i dont know how it works but it works so i dont care
```

```
//i dont know how it works but it works so i dont care :D
```

De onde vem essas sugestões?

- O modelo de IA foi treinado usando o OpenAI Codex, em uma parceria com a OpenAI;
- Como dataset de treinamento foram utilizados códigos em inglês de disponibilidade pública, incluindo os repositórios públicos do GitHub;
- O GitHub argumenta que o modelo foi treinado em repositórios abertos pois redes neurais treinadas em dados públicos é considerado "fair use" pela comunidade.

Dilema Ético 1

Treinar um modelo de uso comercial e não aberto com repositórios abertos, é ético ?

Mas então o CoPilot cópia códigos de outros repositórios?

Resposta curta: não exatamente. Resposta longa: depende...

- De acordo com a definição, o Github CoPilot é, na realidade, um sintetizador de código. Sendo assim, é idealmente improvável encontrar códigos iguais aos presentes no dataset.
  - De acordo com estudos do próprio GitHub, apenas em 0,1% das vezes a sugestão possui snippets iguais ao do dataset de treinamento.
  - Esses casos acontecem apenas quando o contexto dado não é suficiente, ou quanto a solução para o problema for universal (por exemplo uma função de soma).

Mas então qual o problema? (ou melhor, quais as polêmicas?)

Dilema Ético 2

Tendo em vista as diferentes licenças que existem os códigos que derivam deste código com certa licença deve respeitar suas regras de distribuição e uso.

Mas e ao usar o Github CoPilot, qual licença devo respeitar?

Ponto de vista Github

Segundo o site do Github o código gerado é de completa propriedade do desenvolvedor, igual seria com o uso de qualquer outra ferramenta como um editor de texto ou um compilador.

Ponto de vista comunidade

Segundo a FSF (Free software foundation) o CoPilot apesar de nao ser ilegal nao é justo ja que usa trabalho de outros desenvolvedores para desenvolver uma ferramenta de uso comercial exclusivo e nao aberto do GitHub.

Problemas

- Como fazer a conexão entre o código de origem e o código gerado para saber a licença que deve ser utilizada?
- E se forem utilizados vários códigos de origem com licenças diferentes, qual usar?
- Não existe legislação para isso.
- Termos de uso do GitHub

Dilema 3

O Github Copilot é seguro ?

Aparentemente não...

- Estudo diz que 40% das vezes o CoPilot produziu código com falhas de seguranças. (https://www.theinsaneapp.com/2021/09/github-copilot-generated-40-percent-insecure-code.html)

E o Github com isso tudo?

- Todo esse contexto é muito novo! É natural que existam problemas e discussões :D
- Apesar dos dilemas e problemas, o Github vem se mostrando super aberto ao diálogo!
- O Github vem incentivando inclusive conversas sobre fairuse de dados públicos para treinamento de modelos de Machine Learning