Önálló projekt 3

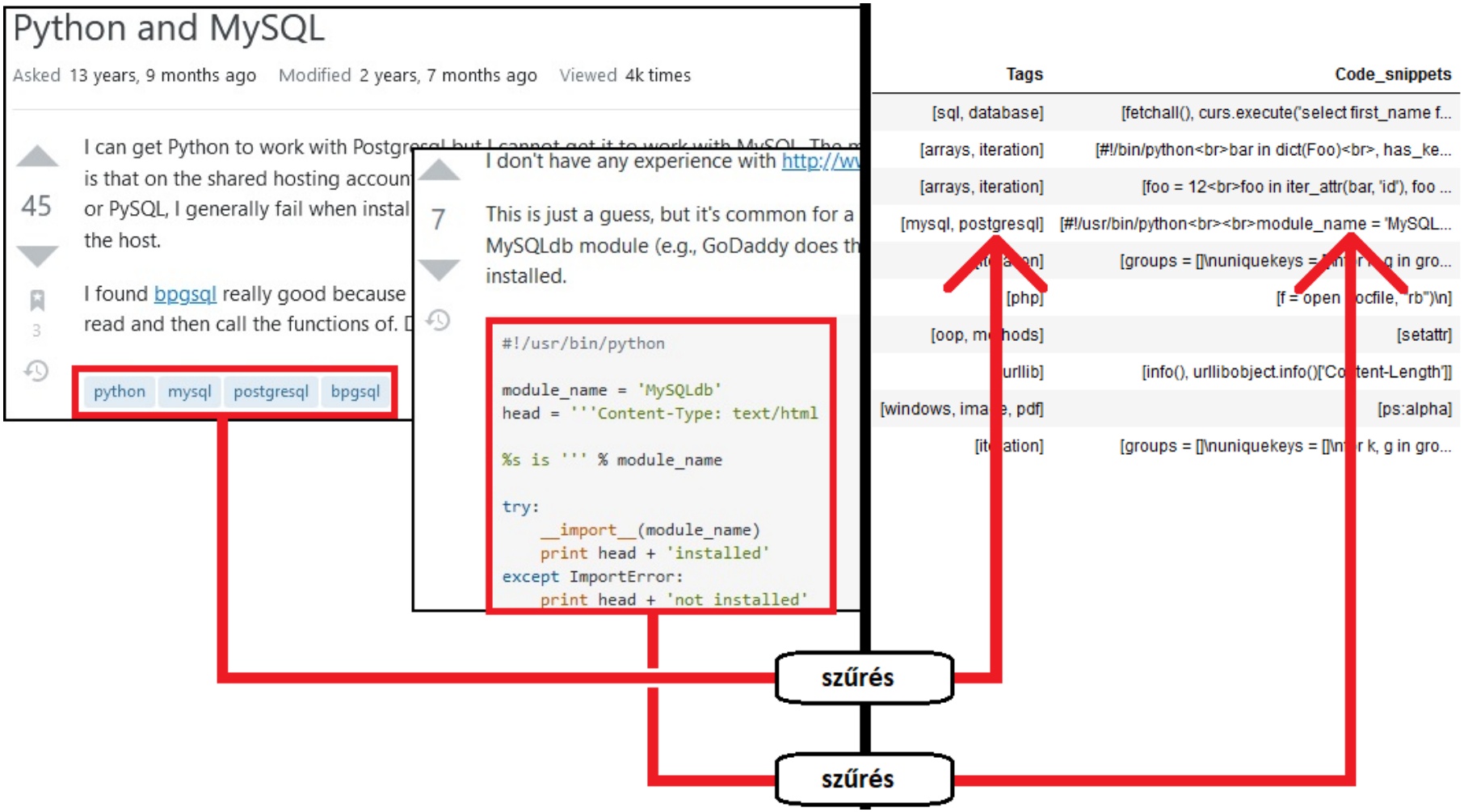
Formális és program nyelvek elemzése gépi tanulási modellekkel

Sisák László Sándor

Témavezető: Lukács András

Adathalmaz

Nyers adat: *Python Questions from Stack Overflow*, 607000 kérdés, 987000 válasz. (forrás: Kaggle)

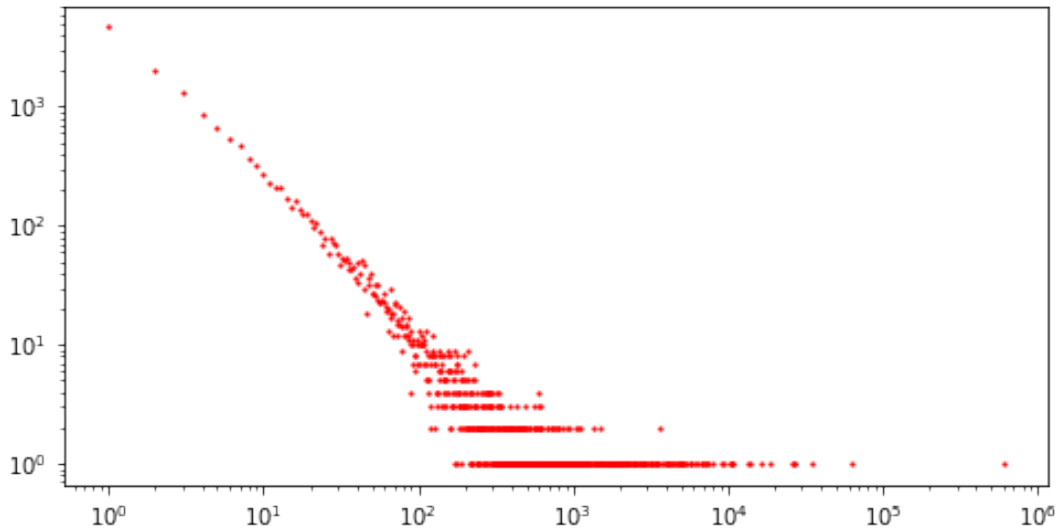


Az adat megtisztítása

A 32 leggyakoribb címkét hagytam meg.

A (tokenizálva) 10 tokennél rövidebb kódokat elhagytam.

Az így kapott adathalmaz 348000 adatpontot tartalmaz.



A címkék gyakorisága. (x, y) pont azt jelöli, hogy y különböző címke van az adathalmazban, amely pontosan x különböző kérdésnél szerepel.

Az adat kiegyensúlyozása

naív oversampling:

$$H_i = \{(x, y) \in D_{\text{train}} : y(i) = 1\}$$

$$N = max\{|H_j| : j \text{ címke}\}$$

$$\forall i \text{ címkére} : k_i = \left\lfloor \frac{N - |H_i|}{|H_i|} \right\rfloor, \quad p_i = \frac{N - (k_i + 1)|H_i|}{|H_i|}$$

H_i minden adatpontját k_i-szer duplikáljuk, majd mindegyiket p_i valószínűséggel még egyszer duplikáljuk. A kapott tanítóadat 1130000 adatpontot számlál.

Az adat kiegyensúlyozása

konvex optimalizálás: Minden adatpontra felveszünk egy változót (x), és megoldjuk a következő feladatot:

$$x \geq \mathbf{1}$$

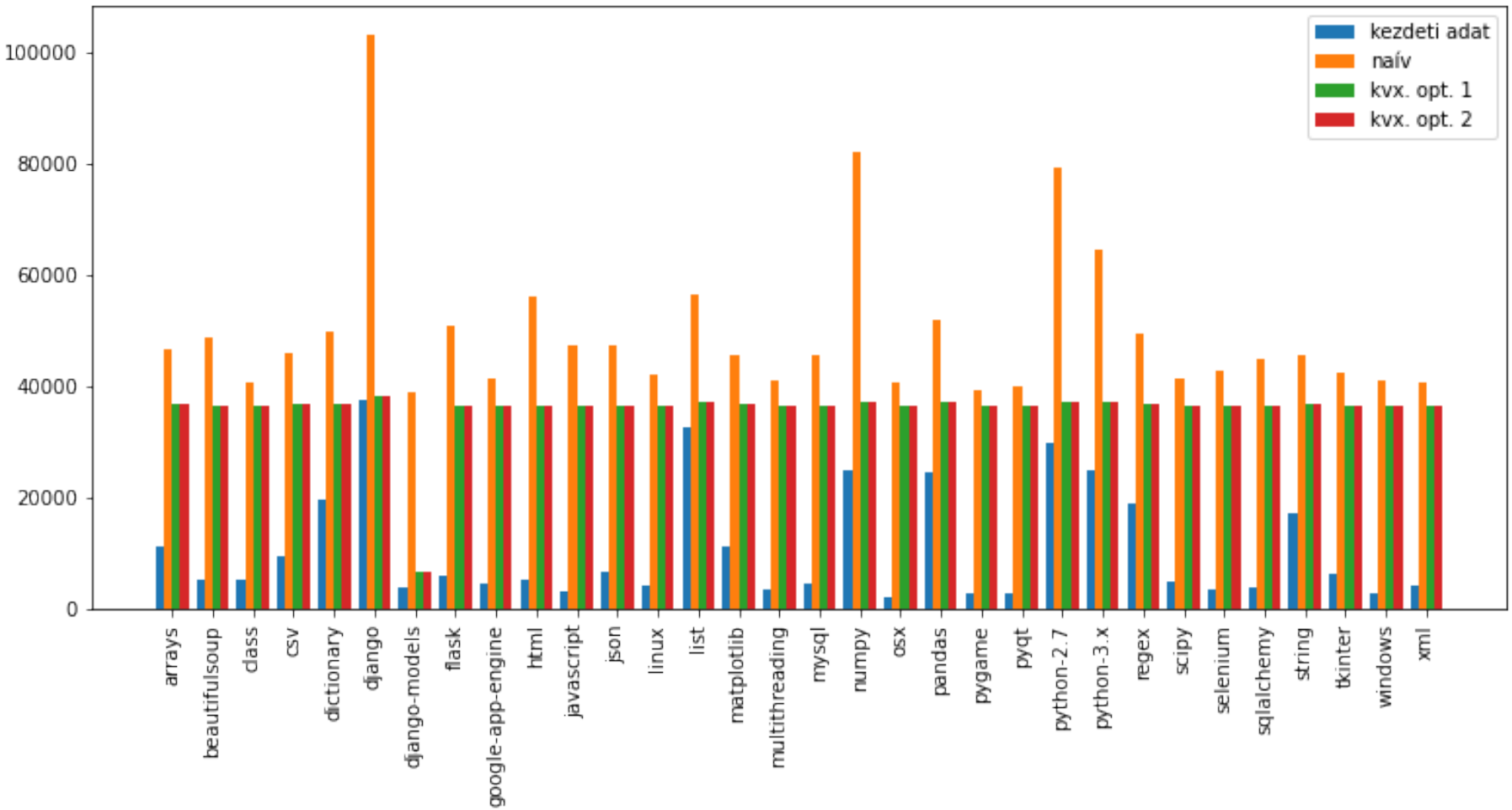
$$\mathbf{1} \cdot |Mx - \mathbf{N}| + \lambda \cdot \max(x) \to \min$$

$M(i,j) = 1$, ha i címke szerepel j adatponton, egyébként 0

$\lambda = 1000$: 836000 adatpont, minden adatpont legfeljebb 18-szor szerepel.

$\lambda = 10000$: 850000 adatpont, minden adatpont legfejebb 12-szer szerepel.

Az adat kiegyensúlyozása



A modell felépítése

Architektúra: CodeBERTa-small encoder és a klasszifikációt végző fej

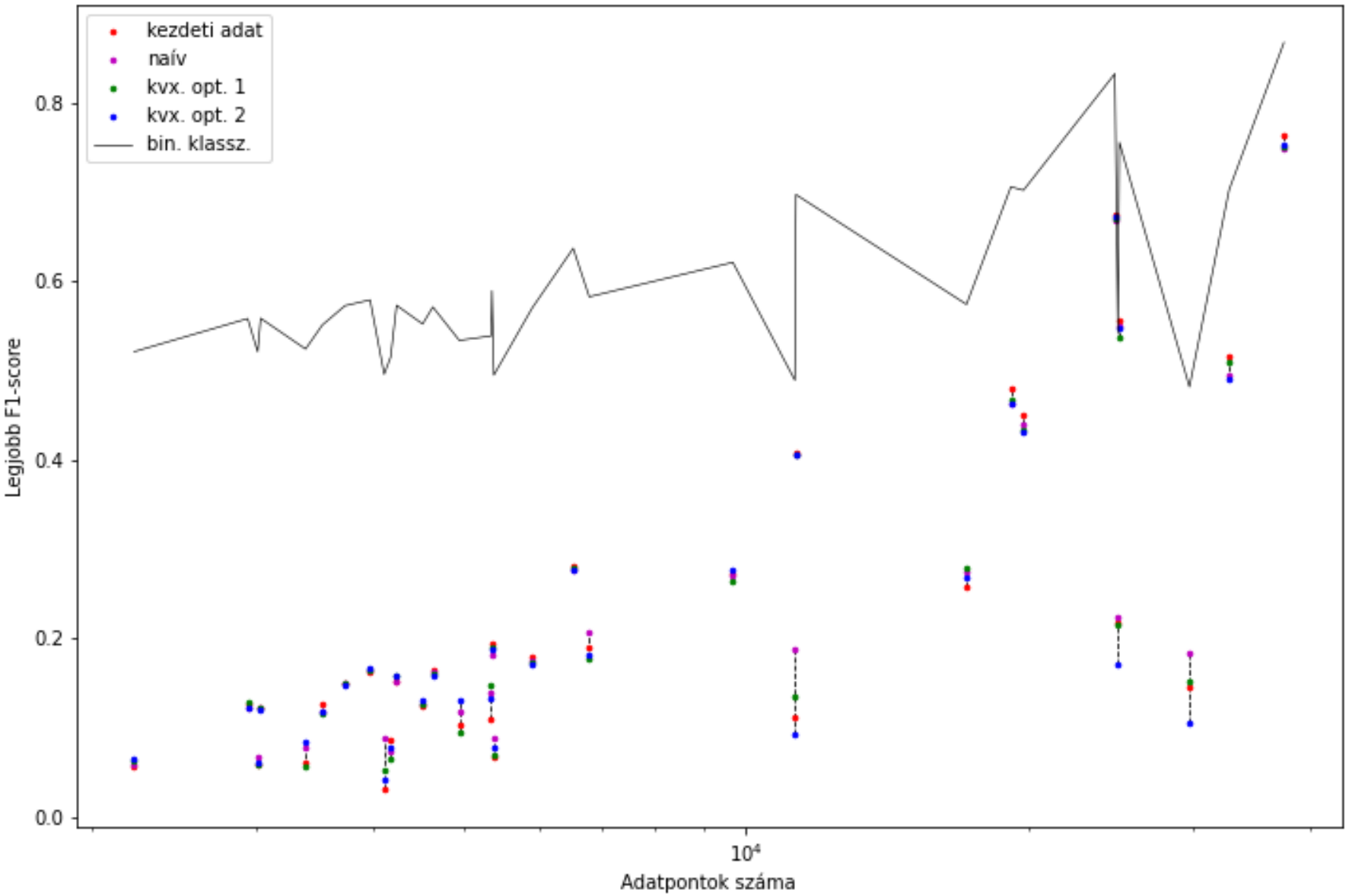
Loss függvény: bináris keresztentrópia

Optimializációs algoritmus: AdamW

Batch size: 16

Implementáció Pytorch modullal

A modell tanítása



A modell értékelése

A különböző címkéken nem ugyanabban az epochban tetőzött az F1-score, ezért nem teljesen egyértelmű, melyik epoch utáni paraméterezést használjuk.
Várakozásaimmal ellentétben a legtöbb címkén a kezdeti adathalmazon tanult modell teljesített a legjobban.

	kezdeti (3)	kezdeti (5)	naív	kvx. opt. 1	kvx. opt. 2
Macro F1	0.219	**0.224**	0.22	0.209	0.216
Micro F1	0.347	**0.356**	0.331	0.321	0.328

Table: A teszthalmazon mért mikro- és makro F1-score különböző paraméterezések mellett.

Kitekintés

Kis adathalmaz, komplex feladat

Apró batchméret

Az AdamW hiperparamétereire nem optimalizáltam

Miért a legjobban kiegyensúlyozott adathalmazon születtek a leggyengébb eredmények?

Köszönöm a figyelmet!