Önálló projekt 2

Formális és program nyelvek elemzése gépi tanulási modellekkel

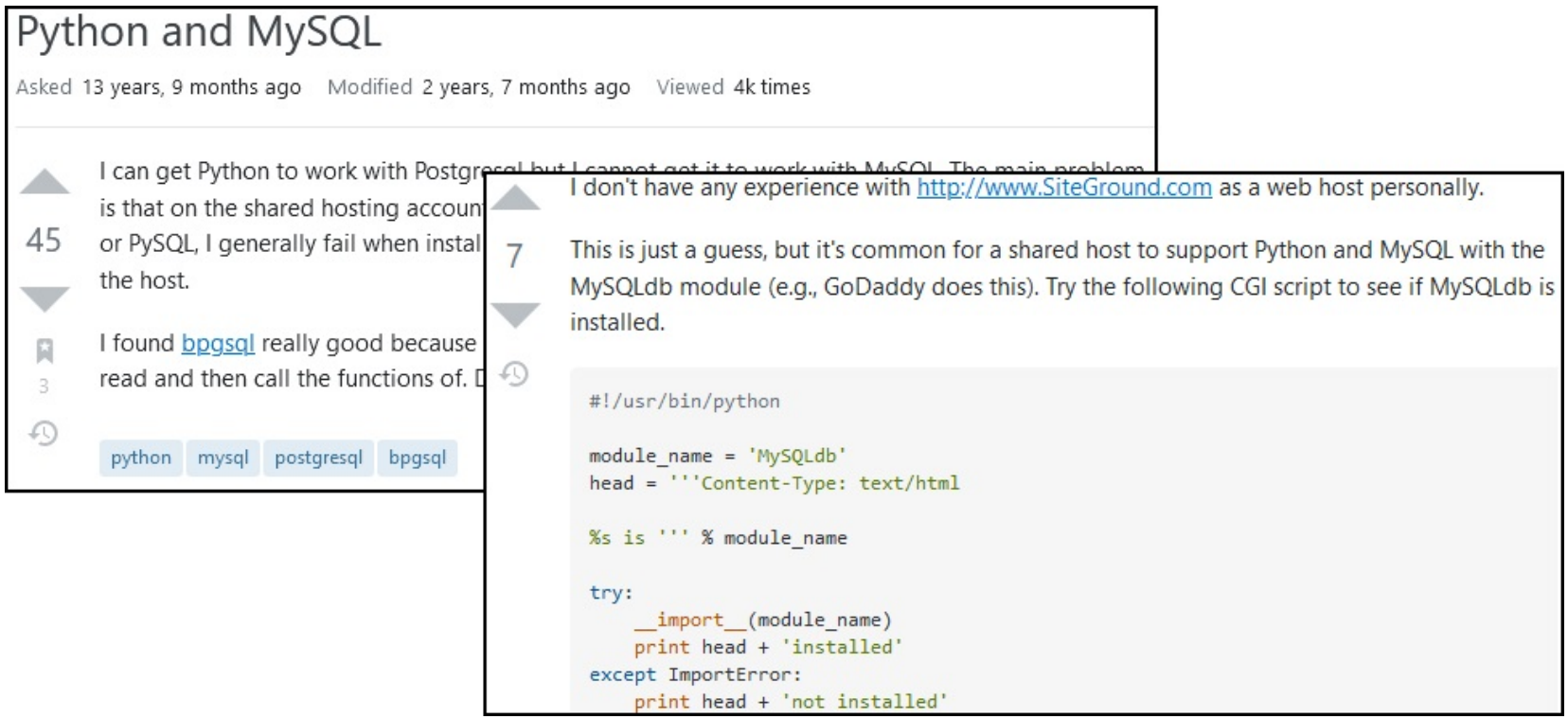
Sisák László Sándor

Témavezető: Lukács András

Feladat

Feladat: Python kódok bináris klasszifikációja gépi tanulás segítségével.

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



Python and MySQL

Asked 13 years, 9 months ago Modified 2 years, 7 months ago Viewed 4k times

45

I can get Python to work with Postgresql but I cannot get it to work with MySQL. The main problem is that on the shared hosting accoun
or PySQL, I generally fail when instal
the host.

I found bpgsql really good because
read and then call the functions of. D

3

python mysql postgresql bpgsql

7

I don't have any experience with http://www.SiteGround.com as a web host personally.

This is just a guess, but it's common for a shared host to support Python and MySQL with the MySQLdb module (e.g., GoDaddy does this). Try the following CGI script to see if MySQLdb is installed.

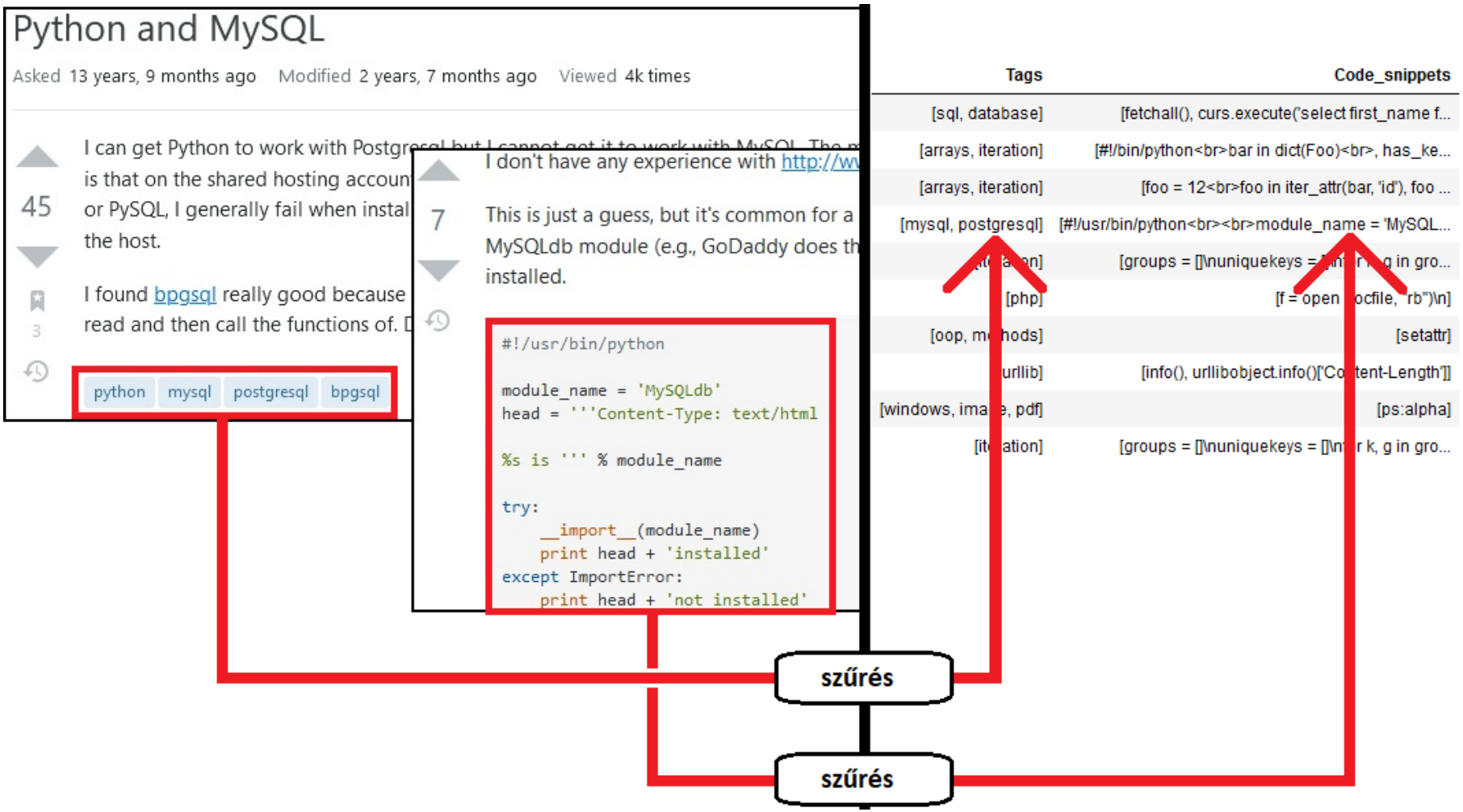
```
#!/usr/bin/python

module_name = 'MySQLdb'
head = '''Content-Type: text/html

%s is ''' % module_name

try:
    __import__(module_name)
    print head + 'installed'
except ImportError:
    print head + 'not installed'
```

Az adat megtisztítása

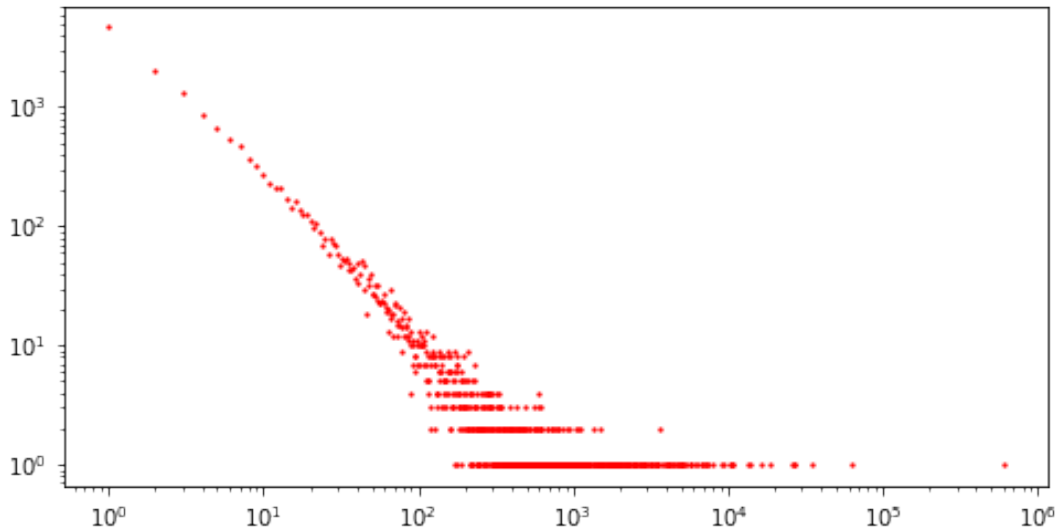


A címkéket gyakoriság, a kódokat szintaktikai helyesség alapján szűrtem.

Az adat megtisztítása

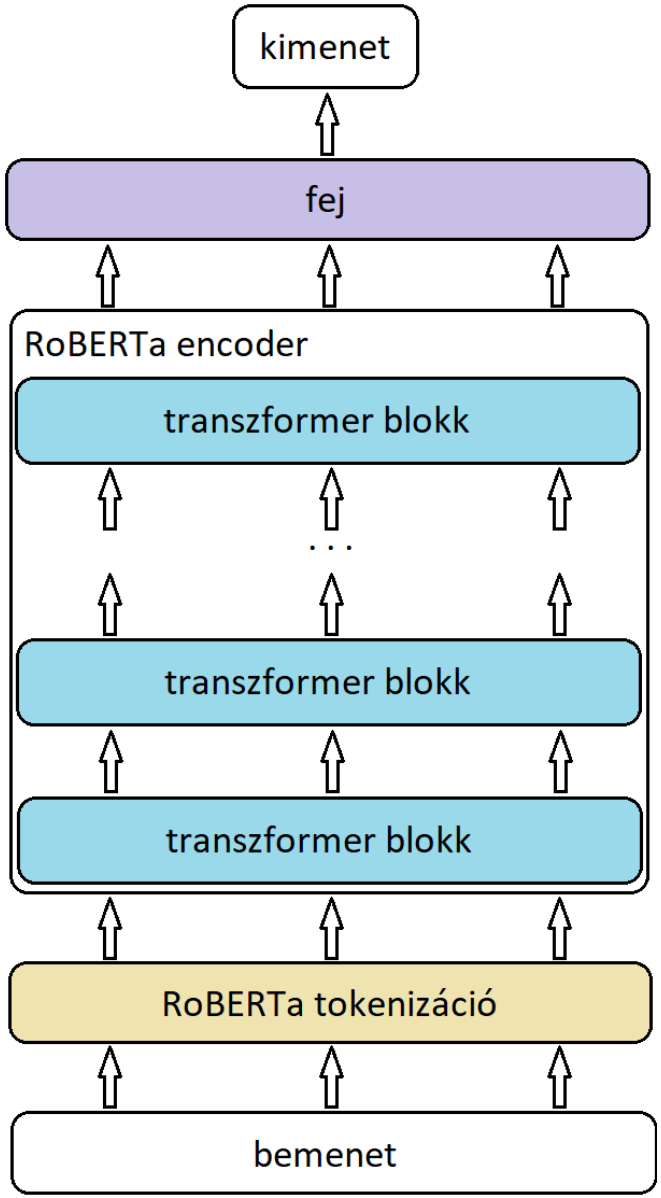
A közel 17000 címkéből csak 207 volt, ami legalább 1000 különböző kérdésnél szerepelt.

A kapott adathalmaz 492000 adatpontot tartalmazott.



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

A modell felépítése



Tokenizáció: A bemeneti stringet egészértékű vektorrá alakítja.

Encoder: A tokenizált bemenetből vektorreprezentációt állít elő. Az így kapott vektor a bemenet jellemzőit írja le.

A tokenizációhoz és a beágyazás előállításához az előtanított RoBERTa transzformer alapú modellt használtam.

Fej: Egy sűrű réteg, ami a beágyazást klasszifikálja.

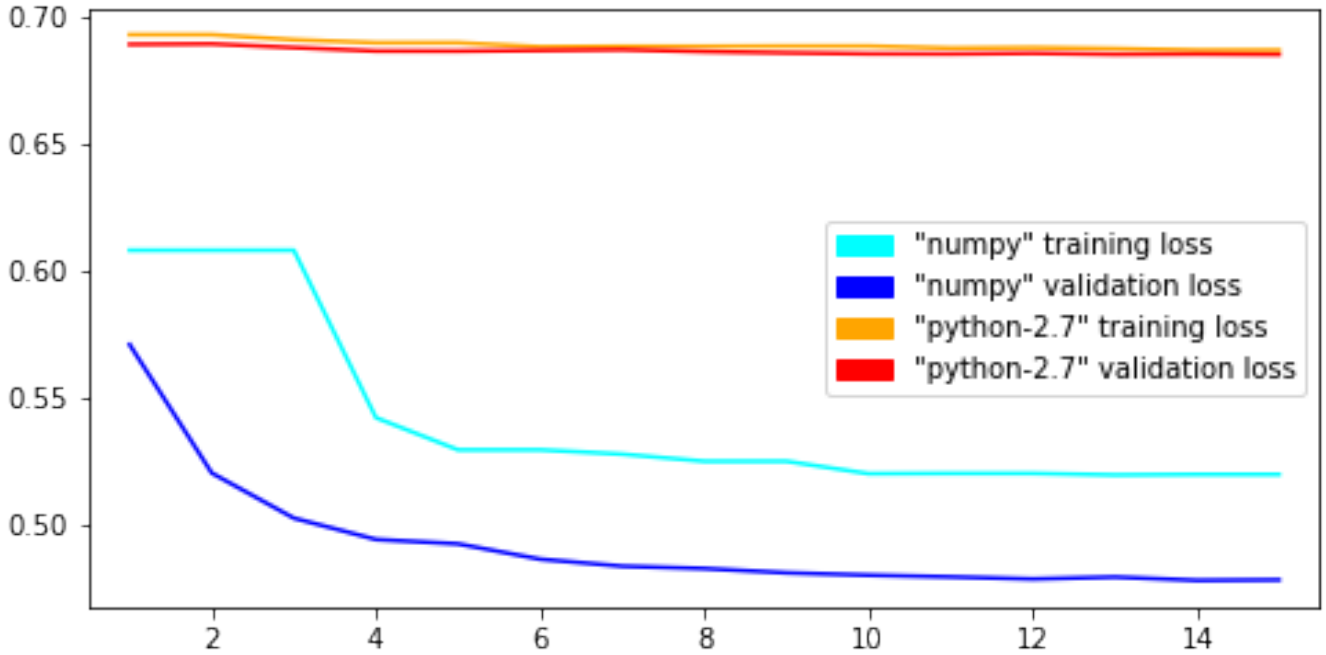
A modell tanítása

A modell csak az öt leggyakoribb címkén tanult. Az adathalmaz ferdeségét korrigálnom kellett.

Címke	Össz. előfordulás	Előfordulás arány	Tanítóadat mérete
django	55105	11,19%	88168
python-2.7	37313	7,58%	59701
python-3.x	33387	6,78%	53419
numpy	30759	6,25%	49214
pandas	28227	5,73%	45163

A modell értékelése

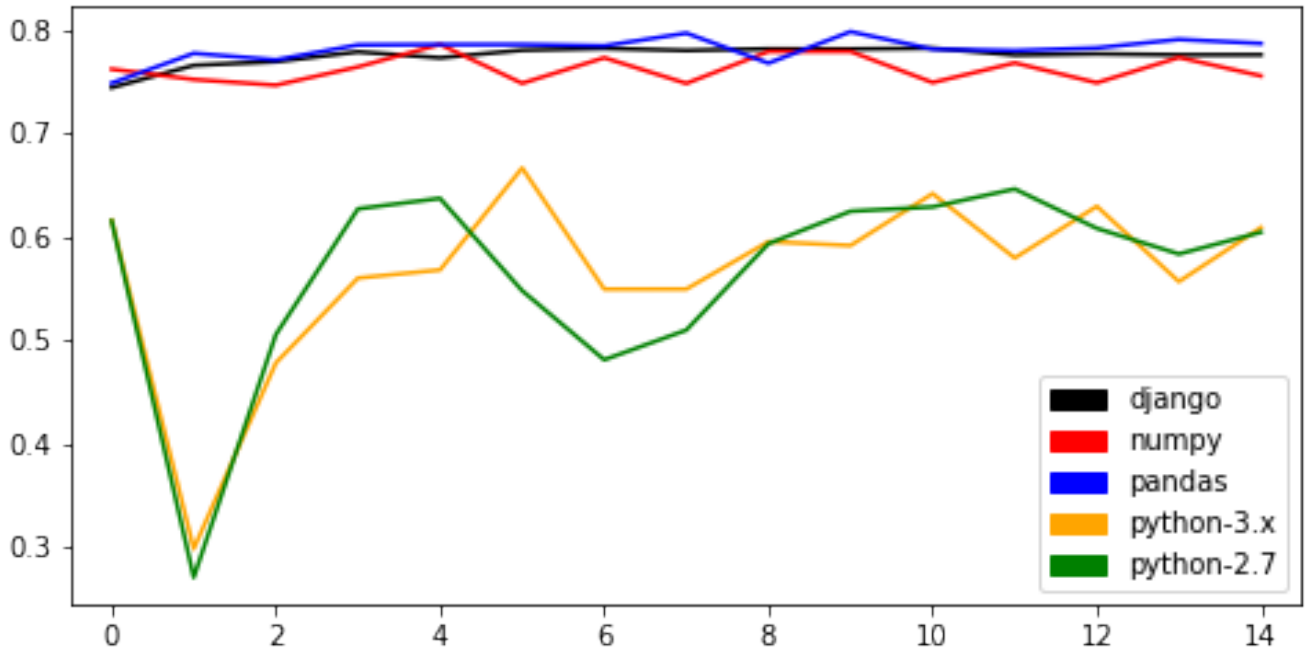
A modell az egyes címkéken 15 epochon keresztül volt tanítva.



Példa a veszteségfüggvény alakulására egy nehezen és egy könnyen tanulható címke esetén.

A modell értékelése

A teljesítményt F_1-scoreban mértem. Az ábra a validációs halmazon mért F_1-score alakulását mutatja a különböző címkékre.



Kitekintés: adathalmaz

Érdemes több adatot gyűjteni.

A nagyon rövid kódrészletekből nem tud tanulni a modell, de könnyen szűrhetőek hossz alapján.

A modell figyelembe vehetné a kódrészletek listájának összes elemét.

Az adat szűrhető pontszám alapján is.

What you're looking for is `setattr` I believe. Use this to set an attribute on an object.

11

```
>>> def printme(s): print repr(s)
>>> class A: pass
>>> setattr(A,'printme',printme)
>>> a = A()
>>> a.printme() # s becomes the implicit 'self' variable
< __ main __ . A instance at 0xABCDEFG>
```

Share Follow

answered Aug 7, 2008 at 11:30

HS.

14.7k • 8 • 38 • 47

Kitekintés: modell

Célom kiterjeszteni a feladatot többcímkés klasszifikációra. Ekkor az egyes címkékre nagyon ferde lesz az adathalmaz.

Érdemes lehet saját tokenizációt tervezni.

A RoBERTán kívül más (előtanított) modellt is érdemes kipróbálni.

A várható jobb beágyazás mellett érdemes összetettebb fejet tanítani.

Köszönöm a figyelmet!