Mendelova univerzita v Brně
Provozně ekonomická fakulta

Detekce duplicit v geoprostorových datech

Diplomová práce

Vedoucí práce:
Ing. Pavel Turčínek, Ph.D.

Bc. Adam Prchal

Brno 2024

Poděkování

Velké poděkování patří vedoucímu diplomové práce Ing. Pavlovi Turčínkovi, Ph.D. za užitečné rady, vedení a ochotu konzultovat v jakoukoliv hodinu. V neposlední řadě patří poděkování také všem, kteří se jakkoliv podíleli na zlepšení kvality této práce.

Čestné prohlášení

Prohlašuji, že jsem práci **Detekce duplicit v geoprostorových datech** vypracoval samostatně a veškeré použité prameny a informace uvádím v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a v souladu s platnou Směrnicí o zveřejňování závěrečných prací.

Jsem si vědom, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

Brno 2024 .....

podpis

Abstract

.

.

Abstrakt

.

.

Obsah

Todo list

Seznam obrázků

1 Úvod a cíl

1.1 Úvod

h the sen-
.

Dobrý den, tohle je moje nové TODO. Pěkný, no ne? (Christen, 2012)

1.2 Cíl

Cílem práce je TODO

```
1   import numpy as np
2
3   def incmatrix(genl1,genl2):
4       m = len(genl1)
5       n = len(genl2)
6       M = None #to become the incidence matrix
7       VT = np.zeros((n*m,1), int)  #dummy variable
8       a = "Asdasdasasdasdas"
9       #compute the bitwise xor matrix
10      M1 = bitxormatrix(genl1)
11      M2 = np.triu(bitxormatrix(genl2),1)
12
13      for i in range(m-1):
14          for j in range(i+1, m):
15              [r,c] = np.where(M2 == M1[i,j])
16              for k in range(len(r)):
17                  VT[(i)*n + r[k]] = 1;
18                  VT[(i)*n + c[k]] = 1;
19                  VT[(j)*n + r[k]] = 1;
20                  VT[(j)*n + c[k]] = 1;
21
22                  if M is None:
23                      M = np.copy(VT)
24                  else:
25                      M = np.concatenate((M, VT), 1)
26
27                  VT = np.zeros((n*m,1), int)
28
29      return M
```

2 Literatura

CHRISTEN, PETER *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin Heidelberg, 2012. 270 s. ISBN 978-3-642-43001-5 978-3-642-31163-5..

HUYEN, CHIP *Designing machine learning systems: an iterative process for production-ready applications*. Beijing Boston Farnham Sebastopol Tokyo, 2022. 367 s. ISBN 978-1-09-810796-3..

MCCLAIN, BONNY P. *Python for geospatial data analysis: theory, tools, and practice for location intelligence*. Beijing Boston Farnham Sebastopol Tokyo, 2023. 262 s. ISBN 978-1-09-810479-5..

MCGREGOR, SUSAN E. *Practical Python data wrangling and data quality*. Sebastopol, CA, 2022. 395 s. ISBN 978-1-4920-9150-9..

NAUMAN, FELIX; HERSCHEL, MELANIE *An Introduction to Duplicate Detection*. , 2022. 84 s. ISBN 978-3-031-01835-0..

WITTEN, IAN H.; FRANK, EIBE; HALL, MARK A.; PAL, CHRISTOPHER J. *Data mining: practical machine learning tools and techniques*. Amsterdam Boston Heidelberg London New York Oxford Paris San Diego San Francisco Singapore Sydney Tokyo, 2017. 621 s. ISBN 978-0-12-804291-5..

SCIKIT-LEARN DEVELOPERS *scikit-learn: machine learning in Python — scikit-learn 1.4.2 documentation* [online]. 2024 [cit. 2024-04-30]. Dostupné z `https://scikit-learn.org/stable/index.html`.