Predikcia popularity článkov

Objavovanie znalostí

Martin Číž, Peter Uherek

FIIT STU, LS 2015

Dataset

- 15 dní návštev článkov na sme.sk
- Články:
  - Lematizovaný nadpis
  - Lematizovaný obsah
  - Počet unikátnych návštev
  - Dátum publikovania

Selekcia dát

- Články z rozmedzia 2 týždňov.
- Pre každý článok sme si vypočítali návštevnosť 1 deň od publikovania.

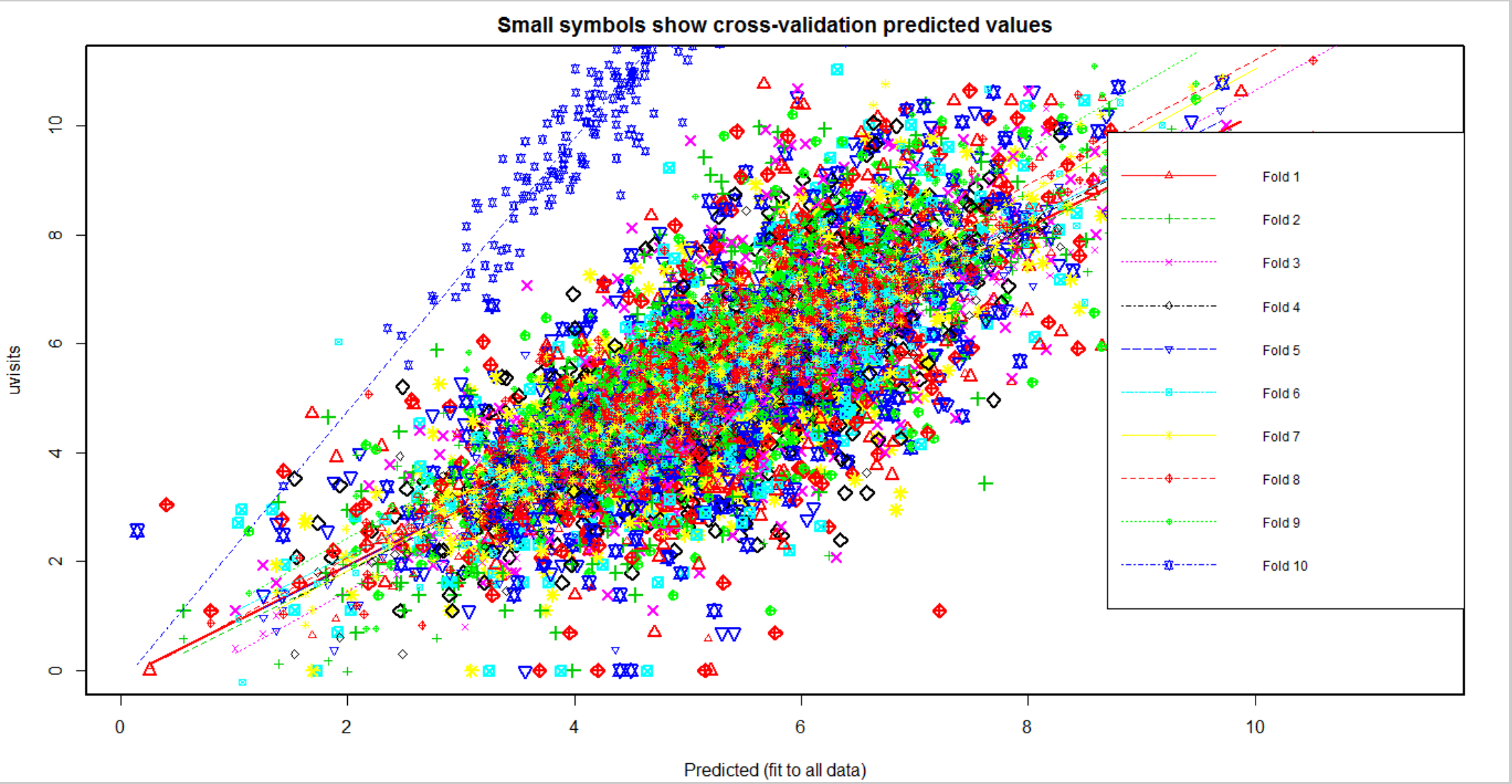
Kroky

- Lematizácia
- http://text.fiit.stuba.sk/lemmatizer/#fast
- Odstránenie stop slov
- Transformácia na vektor metódou TD-IDF
- Prevod na logaritmický tvar
- Lineárna regresia
- Krížová validácia s 10 vrstvami (folds)

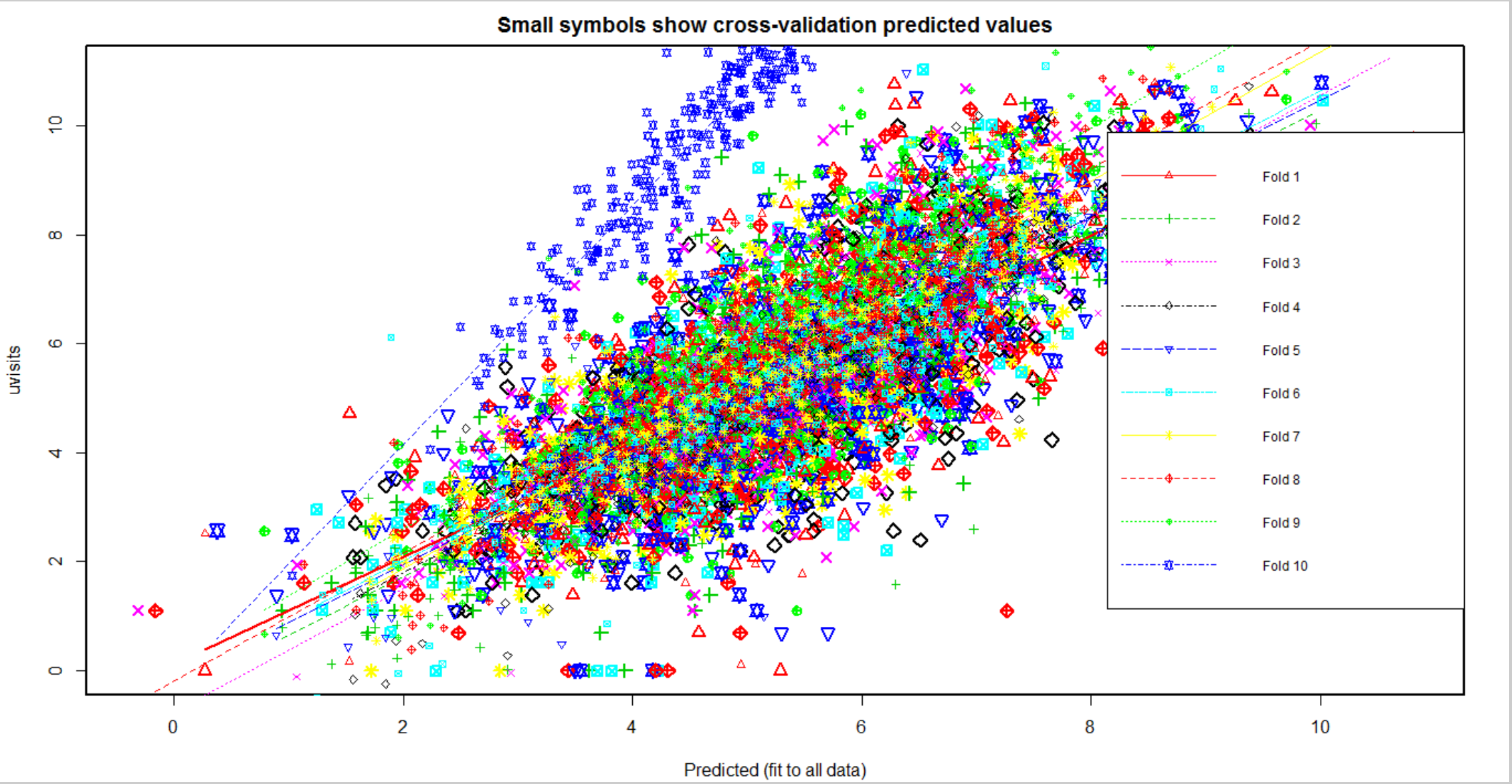
Výsledky

- Prvý pokus
  - Počet článkov 4125
  - Najčastejšie sa vyskytujúce slová (106 / 56 000)
  - CV (Cross-validation error estimate) = 74
  - Priemerná chyba = 1297
  - Medián = 157
- Ďalšie pokusy s viac slovami:
  - 404, 1079, 1357, **2649**, **4056, 5006**

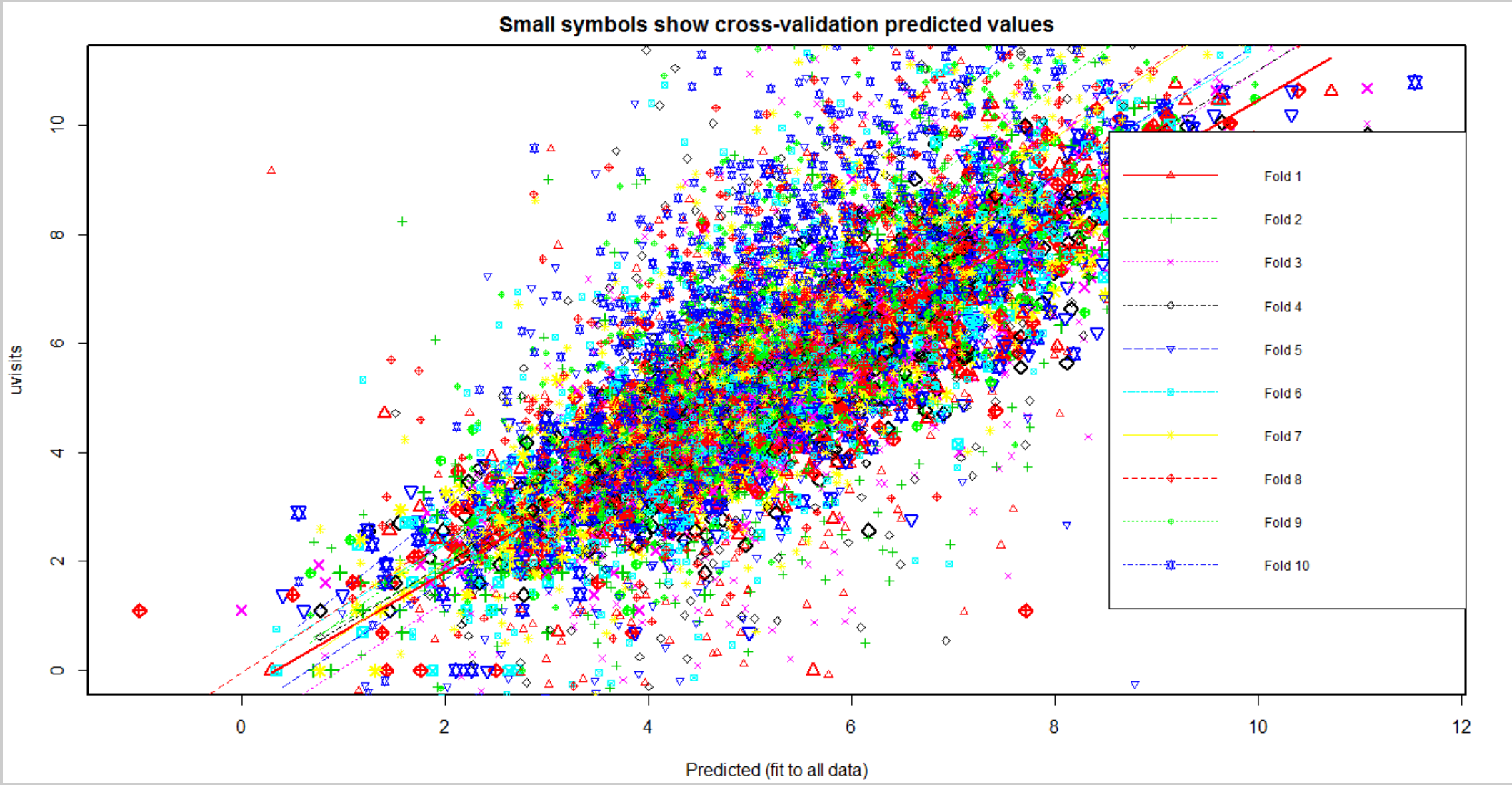
Výsledky


Small symbols show cross-validation predicted values
uvisits
Predicted (fit to all data)
0
2
4
6
8
10
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10

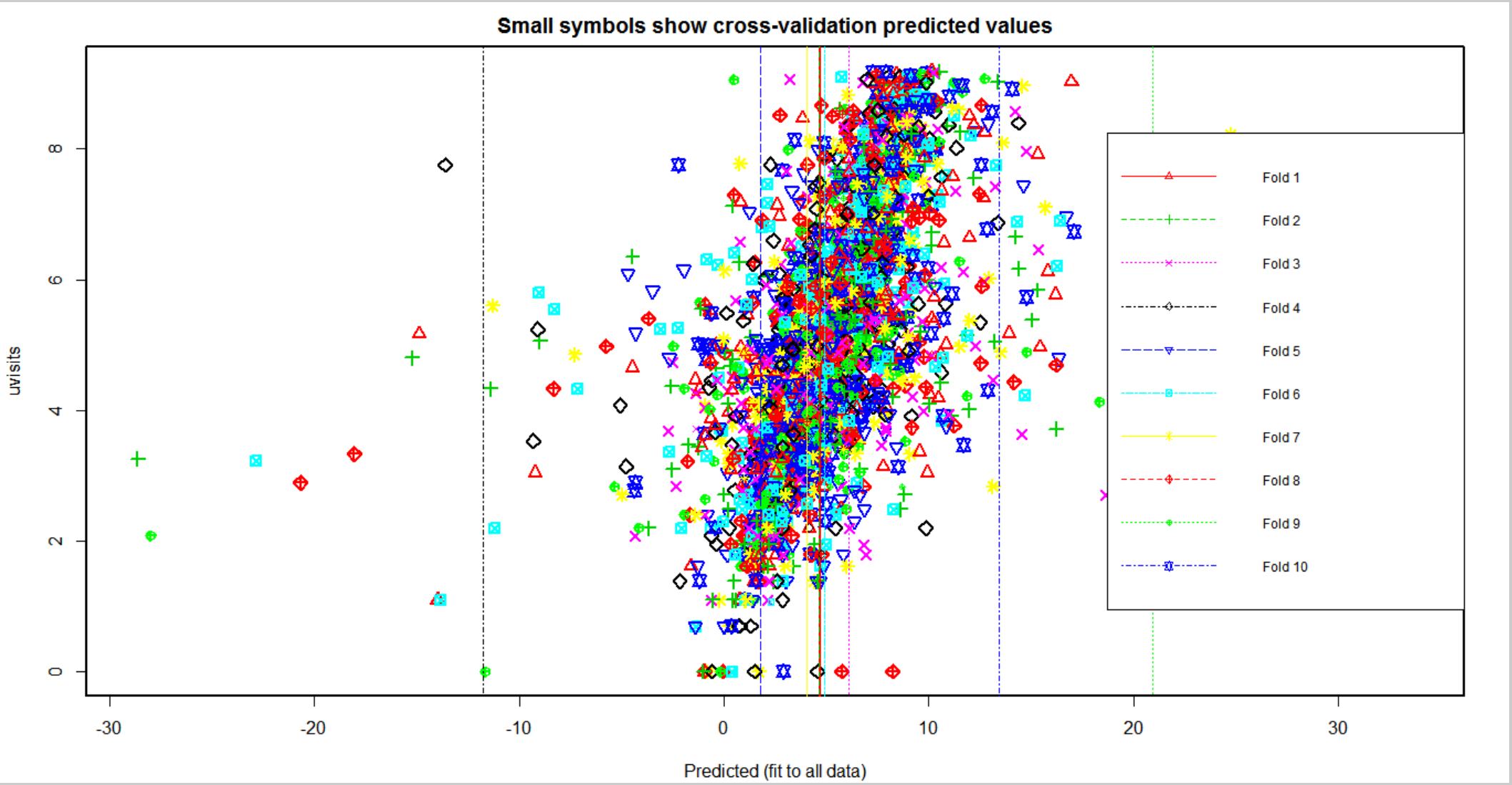
Výsledky


Small symbols show cross-validation predicted values
uvisits
Predicted (fit to all data)
0
2
4
6
8
10
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10

Výsledky


Small symbols show cross-validation predicted values
uvisits
Predicted (fit to all data)
0
2
4
6
8
10
12
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10

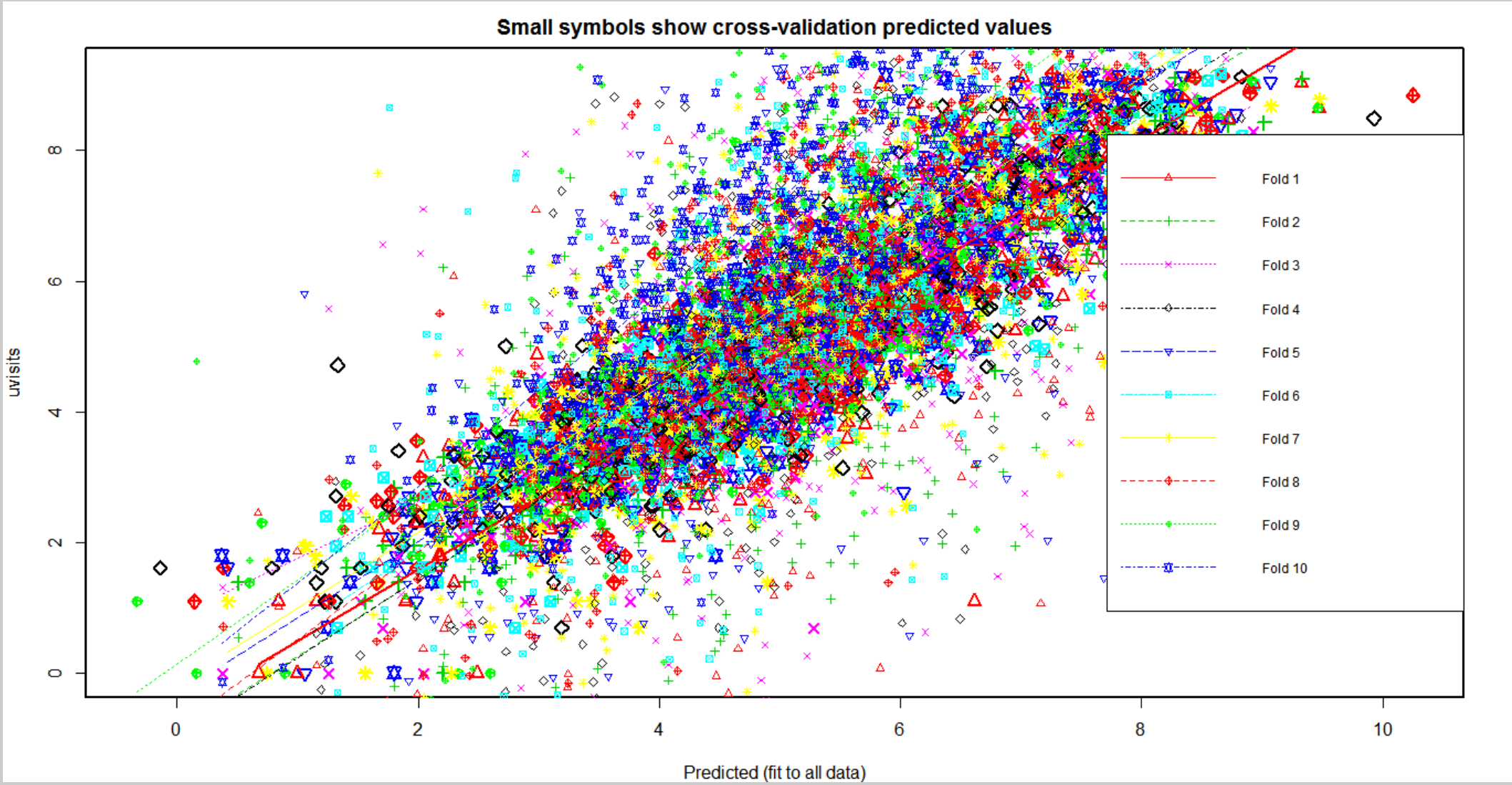
Slepá ulička



Riešenie

- Odstránili sme články s čítanosťou nad 10 000
  - 141 článkov
- Zlepšenie výsledkov
  - CV predtým = 8,73
  - CV potom = 7,7
- Pridanie nových atribútov do vektoru
  - Počet slov (CV = 7.41)
  - Kategória (CV = 7,1)
  - Sekcia, SME kategória (CV = 6,98)
  - Pridanie času (CV = 7,12)
- Najlepší výsledok
  - Priemerná chyba = 421
  - Medián chyby = 68,7

Najlepší výsledok



Čo ďalej

- Pridanie nových atribútov
  - Je uvedený na titulnej strane alebo nie je?
  - Analýza fotografie
  - Analýza nadpisu
  - Iná metrika popularity
  - Presnejšia lematizácia
  - Podobnosť slov, synonýma
- Vyskúšať iné modely
  - Iné regresné modely
  - Nerúnové siete