Metody Odkrywania Wiedzy

Projekt 2018 Z

Predykcja ocen książek – badania

Igor Markiewicz
Aleksander Droszcz

Prowadzący – dr inż. Paweł Cichosz

Spis treści

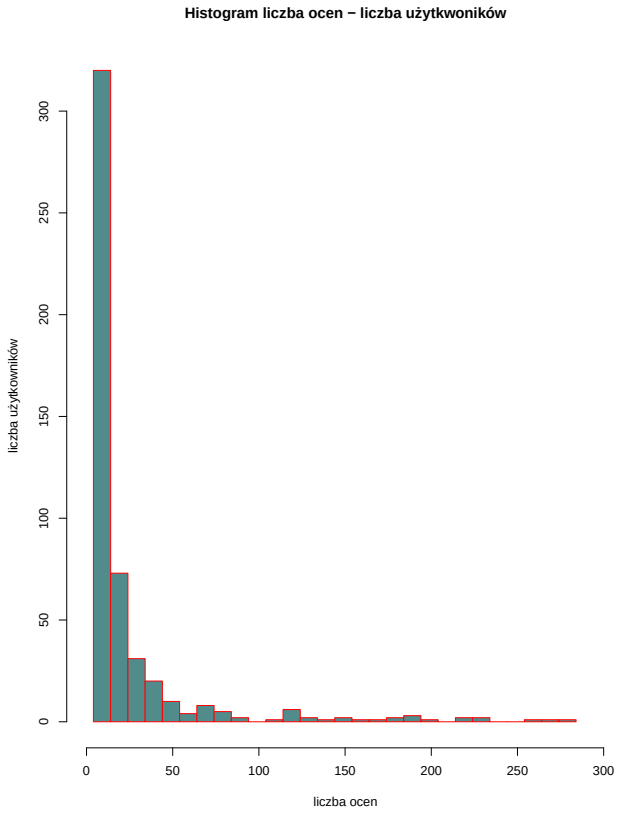
1 Założenia

Po wstępnych testach zostały poczynione następujące założenia:

- zrezygnowano z użycia metody Item Based Collaborative Filtering – dla mniejszych danych zwracała wartości nieokreślone, dla większych jej działanie trwało bardzo długi oraz było bardzo zasobożerne pamięciowo
- na wstępie zostało ustawione ziarno standardowego generatora liczb pseudolosowych na wartość 1648
- w każdym teście zastosowano 5. krotną walidację krzyżową
- do zbioru ucząco – testowego zostało wybranych losowo 500. użytkowników wraz z ich wszystkimi ocenami, ale tylko takich których liczba ocen jest równa co najmniej 5
- procedura testowa polega tym że jeśli istnieje n ocen, a użytkownik ma ich $m \leq n$, to wybieramy losowo $k \leq m$ ocen przeznaczonych dla algorytmu do odtworzenia $n-k$ ocen. Po odtworzeniu przeprowadzamy testy na nieużytych $m-k$ ocenach przeznaczonych do validacji. W przypadku testów zdecydowano się na testowanie pojedyńczej oceny (parametr *given = -1*)

2 Badania

2.1 Wstępna analiza danych



Rys. 1: Histogram liczba ocen – liczba użytkowników

2.2 Optymalizacja algorytmu User Based Collaborative Filtering

2.2.1 Testy rodzaju normalizacji danych dla UBCF

2.2.2 Testy metryk podobieństwa dla UBCF

2.2.3 Testy dla różnej ilości najbliższych sąsiadów

2.3 Optymalizacja algorytmu Funk SVD

2.3.1 Testy rodzaju normalizacji danych dla SVDF

2.3.2 Testy dla różnych ilości składowych utajonych SVDF

2.3.3 Testy dla różnych współczynników szybkości uczenia SVDF

2.4 Porównanie najlepszych modeli

Wykonano test t – Studenta dla dwóch populacji:

$$\begin{cases} H_0 : \mu_{UBCF} - \mu_{SVDF} = 0 \\ H_1 : \mu_{UBCF} - \mu_{SVDF} \neq 0 \end{cases}$$

gdzie μ_{UBCF}, μ_{SVDF} oznaczają średnie arytmetyczne RMSE/MAE z pięciu prób dla najlepszych modeli. W efekcie otrzymano

- dla RMSE: p–value = 0.5464
- dla MAE: p–value = 0.4937

W obu przypadkach z racji na duże wartości p–value nie mamy podstaw do odrzucenia hipotezy zerowej o równości średnich w obu populacjach.

2.5 Binaryzacja danych

2.6 Proste statystyki

3 Uwagi i wnioski

4 Bibliografia

[1] https://grouplens.org/datasets/book-crossing/

[2] https://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf

[3] https://cran.r-project.org/web/packages/recommenderlab/recommenderlab.pdf

[4] https://en.wikipedia.org/wiki/Collaborative_filtering

[5] https://cran.r-project.org/web/packages/rrecsys/rrecsys.pdf

[6] https://en.wikipedia.org/wiki/Matrix_factorization_(recommender_systems)#Funk_SVD

[7] http://nicolas-hug.com/blog/matrix_facto_1

[8] http://nicolas-hug.com/blog/matrix_facto_2

[9] http://nicolas-hug.com/blog/matrix_facto_3

[10] https://www.slideshare.net/DKALab/collaborativefilteringfactorization