Ciekawostki (wyjaśnione i niewyjaśnione) powstałe w trakcie projektu i laboratorium z Eksploracji danych

Eksploracja danych użytych w projekcie „Analiza danych MyAnimeList" przyniosła parę zaskakujących i nieoczywistych spostrzeżeń. Pierwszy z nich to ilość niepoprawnych danych wszędzie tam gdzie użytkownik mógł wprowadzić dowolną wartość. Wiele atrybutów dotyczących, np. położenia geograficznego czy daty urodzenia. Kolejnym aspektem, który należało zweryfikować była deklarowana aktywność użytkowników, gdzie w niektórych przypadkach można było bezwzględnie stwierdzić, że zgłaszana ilość obejrzanych anime była fizycznie niemożliwa do zrealizowania.

Jednym z ciekawych wyników był ten odnośnie wieku użytkowników oraz jego rozkład. Średnia w okolicach wartości 27 lat była zaskakująca, biorąc pod uwagę, że anime są zwykle kierowane do młodszych audiencji. Wiele gatunków jest kierowanych głównie do nastolatków. Ten wynik mógł być spowodowany faktem, że strona została założona w 2004 roku a dane pochodziły z 2016, jak również może być spowodowane chęcią użytkowników podawania się za dorosłych by mieć dostęp do wszystkich materiałów strony. Co więcej rozkład wieku miał kształt rozkładu normalnego (jeden pik).

Warto zwrócić uwagę, że zbiór użytkowników w projekcie był zdominowany (65%) przez użytkowników podających się za mężczyzn. Pokrywa się to z osobistymi obserwacjami autorów widzianymi w Internecie, aczkolwiek bez solidnych dowodów. W rzeczywistości proporcje mogą być inne pomiędzy widzami anime.

Klasyfikacja danych na zadane etykiety okazała się być niebywale dokładna. Odgadywanie np. typu anime (czyli czy coś jest serialem czy filmem itp.) miało wysoką dokładność nawet biorąc pod uwagę, że do odgadnięcia było 6 różnych klas. Co ciekawe odgadywanie ograniczenia wiekowego (4 różne klasy) również było relatywnie dokładne. Wzbogacenie odgadywania opartego o dane numeryczne o takie oparte o dane kategoryczne czy tekstowe mogło by przynieść bardzo dokładne wyniki. Natomiast odgadywanie płci użytkowników na podstawie jedynie danych numerycznych okazało się raczej losowe i jedynie najbardziej reprezentowana grupa mężczyzn była poprawnie odgadywana.

Podczas analizy zbioru opisującego kraje, zachorowania i śmierci na COVID-19 można było zauważyć pewnie zależności między metrykami, tj. PKB per capita i ilość łóżek na ilość mieszkańców. Zgodnie z intuicją, bogatsze kraje były lepiej przygotowane do starcia z pandemią, jednakże nie wpłynęło to na ich skuteczność w walce z wirusem.

Prosta metoda filtracji często używanych słów (TF-ID) znacząco poprawiła działanie nawet prostej metody klasyfikacji jaką jest metoda K-sąsiadów w ćwiczeniu laboratoryjnym dotyczącym eksploracji tekstów. Poprawa sięgała 20% różnicy w dokładności co jest znaczące, gdyż zbiór posiadał 20 różnych etykiet.

W ćwiczeniu o systemach rekomendacyjnych (zawierającym zbiór opisujący filmy) ciężko było ocenić jakość rekomendacji, gdyż zbiór nie posiadał wystarczającej liczności. Aczkolwiek algorytmy były w stanie rekomendować podobne filmy bazując na preferencjach do nich przekazanych.

Dominik Dziuba, Miłosz Filus