UNIWERSYTET EKONOMICZNY W POZNANIU

Wydział Informatyki
i Gospodarki Elektronicznej

Wojciech Kowalczyk

Rekomendacja filmów z wykorzystaniem metod wielowymiarowych na przykładzie platformy Netflix

Movie recommendation with the use of multivariate methods based on Netflix data

Praca licencjacka

Promotor: dr Łukasz Wawrowski

Data przyjęcia:

Podpis promotora

Kierunek: Informatyka i ekonometria

Specjalizacja: Analityka gospodarcza

Poznań 2019

Spis treści

Wprowadzenie

Wstęp

Gwałtownie rozwijający się rynek filmowy i telewizyjny w Polsce i na świecie, a także wzrastająca dostępność do serwiów VOD (Video On Demand) stanowi główną motywację prowadzonych w pracy badań. Platformy dysponują coraz bogatszą ofertą filmową a tym samym liczba klientów nieustannie się powiększa. Zdecydowanym liderem światowym jest plaforma Netflix. Dziesiejszy rynek zależy w zdecydowanej większości od klienta, i to on narzuca zasady. Dlatego duże platformy chcą spełnić ich wymagania. Jednym ze sposobów jest personalizacja produktów oraz traktowanie klienta jako jednostkę. Dlatego serwisy VOD postanowiły rozwiązać ten problem poprzez systemy rekomendacji, które personalizują i rekomendują tytuły dla każdego użytkownika osobno. Wraz z rozwojem tych platform oraz wzrostem liczby użytkowników, aktualnie zauważana jest potrzeba poprawy działania już istniejących algorytmów ale także budowanie skuteczniejszych systemów rekomendacji.

Celem pracy jest analiza danych dotyczących filmów pochodząca z serwisu Netflix oraz ich grupowanie. W pracy podjęto próbę stworzenia algorytmu rekomendacji filmów z platformy Netflix bazującym na ocenach użytkowników oraz danych na temat tytułów pochodzących z serwisu Internet Movie Database.

Praca składa się z trzech rozdziałów, z których każdy podzielony jest na podrozdziały i ma charakter teretyczno-empiryczny. W rozdziale pierwszym omówiono aspekty związane z serwisami oferującymi materiały filmowe na żądanie oraz skupiono się na omówieniu i analizie platformy Neflix. Poruszony został również temat czynników wykorzystywanych w systemach rekomendacji.

Rozdział drugi to zaprezentowanie metod grupowania i klasyfikacji. Wyjaśniona została istota analizy skupień i jej zastosowanie. Dodatkowo opisane zostały metody

iteracyjno-optymalizacyjne oraz hierarchiczne służące do grupowania danych. Zaprezentowana została również metoda K Najbliższysch Sąsiadów.

Natomiast rozdział trzeci ma charakter empiryczny. Dokonano w nim krótkiej charakterystyki i opisu danych pochodzących z serwisu Netflix oraz Internet Movie Database. W tym celu przeprowadzono kompleksową analizę struktury, a wybrane zmienne przedstawiono na histogramach. Następnie dokonano grupowania danych, które wyodrębniło skupienia użyte w dalszej kolejności w algorytmie rekomendacji. Wyniki stworzonego systemu oceniono na przykładzie losowo wybranego użytkownika serwisu Netflix.

Rozdział 1

Netflix jako nowoczesna platforma i wynalazek XXI wieku

1.1 Rozwój rynku Video On Demand w XXI wieku

Video On Demand, a w języku polskim pojęcie to szerzej znane pod terminem wideo na życzenie, to szeroko pojęta usługa, która ma na celu umożliwienie użytkownikowi dostępu do programu telewizyjnego, filmu lub transmisji w innym terminie niż jest emisja(Lidia Drabik, 2006). Materiał na platformach jest najczęściej dostępny przez cztery główne kanały dystrybucji:

- Set-top-box — narzędzie kompatybilne z telewizorem służące do odtwarzania wideo i dźwięku oraz współpracuje wraz z dostawcą telewizji
- Telewizja hybrydowa — odbiór danych od nadawcy telewizyjnego oraz tych, które pochodzą z Internetu.
- Strony internetowe — platformy, które umożliwiają oglądanie materiałów w przeglądarce internetowej.
- Aplikacje mobilne — program jest wyświetlany w aplikacji dostępnej na telefonie.

Na rozkwit VOD składa się wiele czynników, które są spowodowane rozwojem technologii czy zmian społecznych. Po pierwsze, spowodowany jest szybkim rozwojem Internetu oraz telefonii trzeciej generacji. Nie zadziwia więc fakt, że jest to świetna alternatywa dla dobrze nam znanych kaset wideo oraz magnetowidów. To właśnie VOD pozwala na

dostęp do materiałów niezależnie od miejsca, a jedynym ograniczeniem jest subskrypcja dowolnego serwisu filmowego oraz mobilność sprzętu. Ponadto nie wymaga on miejsca do przechowywania kaset, ma wysoką jakość i jest o wiele wygodniejsze niż tradycyjna metoda. Ponadto serwisy VOD to legalne źródło filmów i seriali a w dzisiejszych czasach już część popkultury. Ze względu na rosnącą popularność i gwałtowny przyrost użytkowników powstało wiele rodzajów VOD w zależności od potrzeb użytkownika, dostępnej technologii i miejsca przechowywania treści.(Stępka, 2009).

Tabela 1.1 zawiera podział materiałów wideo ze względu na miejsce przechowywania treści:

Tabela 1.1. Podział materiałów wideo ze względu na miejsce przechowywania treści

Push VoD	Materiały przetrzymywane są na dysku użytkownika a nie u dostawcy usługi. Mają określony czas egzystencji oraz użytkownik ma pełną dowolność co do aktywności. Infrastruktura ta nie posiada kanału zwrotnego lub nie jest on używany.
Pull VoD	Materiały przetrzymywane są u dostawcy usługi, a nie użytkownika. Infrastruktura to aktywnie korzysta z kanału zwrotnego a materiały są dostępne cały czas.

Z kolei podział ze względu na model biznesowy zawarty jest w tabeli 1.2 :

Tabela 1.2. Podział ze względu na model biznesowy

Free on Demand (FOD)		Dostęp do materiałów jest darmowy dzięki wyświetlanym reklamom w czasie trwania materiału.
Rental VoD		
	Pay per Download	Dostęp do materiału, dostępny jedynie po pobraniu. Ograniczony czas na obejrzenie materiału.
	Subscribtion VoD	Dostęp do materiału, dostępny jest dzięki wykupieniu abonamentu i płaceniu opłaty subskrypcyjną.
	Packages	Klient posiada określoną gotówkę na koncie, za którą może wypożyczać materiały.
	Packs	Klient za każdym razem wybiera materiał, który chce mieć udostępniony za opłatą.
Download to own VoD		Dostęp do materiału jest dostępny po zapłaceniu a materiał może być przetrzymywany na dysku bez ograniczeń czasowych.

1.2 Platforma Netflix

Netflix to aktualnie jeden z najpopularniejszych i powszechnie dostępnych usług na rynku wykorzystujący przesyłanie strumieniowe czyli dostarczanie obrazu i dźwięku od dostawcy w bardzo krótkim czasie — na życzenie lub na żywo (Ziółkowska, 2011). Jest to platforma internetowa, która umożliwia oglądanie materiałów audiowizualnych niezależnie od miejsca czy czasu za pośrednictwem urządzeń, które mają dostęp do Internetu. Obowiązkowym jest zakup jednego z trzech dostępnych rodzajów subskrypcji (w Stanach Zjednoczonych czterech), które różnią się jedynie maksymalną liczbą dostępnych ekranów dla użytkowników.

Historia przedsiębiorstwa zaczęła się w 1997 kiedy to Reed Hastings oraz Marc Randolph postanowili otworzyć wypożyczalnię kaset wideo i płyt DVD , które były dostarczane za pomocą poczty. Za niewielką, miesięczną opłatę klient mógł wypożyczać filmy bez wychodzenia z domu. Przez pierwsze 10 lat udało dostarczyć się miliard płyt dla klientów a w 2008 roku wraz z rozwojem Internetu i urządzeń posiadających dostęp do niego postanowiono zacząć działalność oferującą dostęp do filmów bez wychodzenia z domu dla osób posiadających subskrypcję bez dodatkowych kosztów(Lusted, 2013). Niestety platforma aż do 2016 roku była dostępna w niewielu krajach (wcześniej w 2010 roku zadebiutowała w Kanadzie), lecz postanowiono znieść blokadę i rozpowszechnić Netflix na całym świecie, rozszerzając usługę o 130 krajów. Aktualnie jedynymi krajami, które nie posiadają dostępu do platformy jest Korea Północna, Syria, Chiny oraz Krym — część jednostki organizacyjnej Rosji(„Netflix launches in nearly every country but China", 2016). Materiały dostępne są w wielu wersjach językowych oraz posiadają szeroki wybór. Ważnym jest fakt, iż dostępność materiałów różni się od kraju, w którym zakupiona jest subskrypcja, co jest tematem wielu dyskusji.

Ale Netflix to nie tylko firma, która dostarcza filmy i seriale, lecz sama również je produkuje. Pierwszą oryginalną produkcją Netflix'a jest odtworzenie popularnego serialu "House of Cards", który zadebiutował w 2013 roku. Od tego momentu oryginalne produkcje Netflix'a cieszą się ogromną popularnością. Niezaprzeczalnie, powodem jest wysoka jakość materiałów, dobry marketing i szeroka tematyka, która znajduje swoich odbiorców. Netflix do tej pory wyprodukował ponad 700 oryginalnych seriali i filmów, z których wiele nagrodzono nagrodami. Ponadto w 2018 przeznaczył 12 miliardów dolarów na ich wykonanie. Szacuje się, że w 2019 ta kwota wyniesie 15 miliardów

dolarów(Paszkowski, 2019).

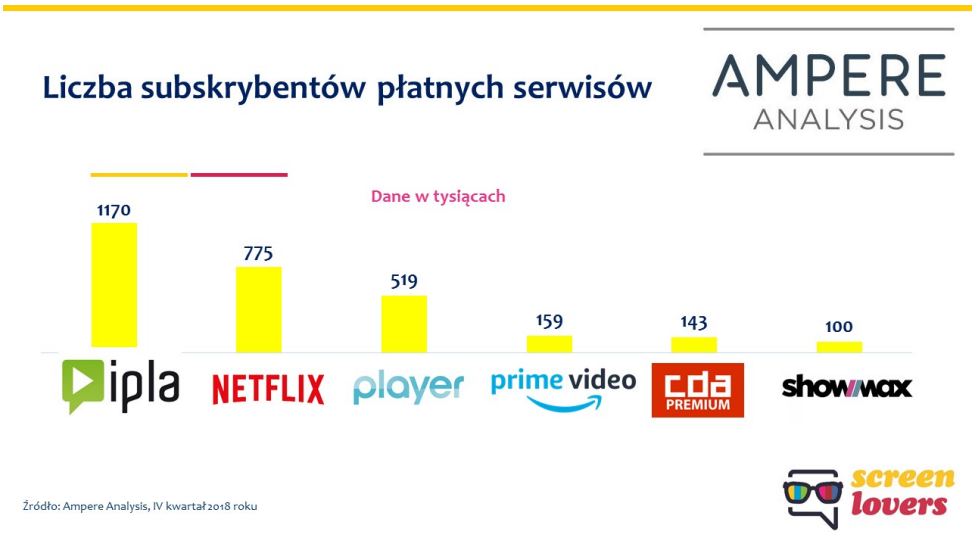
1.3 Rynek Netflix w Polsce i na świecie

Netflix to już znana i szeroko rozpoznawalna marka na całym świecie i obecnie nie ma żadnego konkurenta, który zagrażałby pozycji budowanej przez wiele lat. Mimo że rynek na całym świecie, tak i w Polsce oferuje wiele serwisów streamingowych (Hulu, Showmax, Amazon), to jednak Netflix przyciąga największą uwagę i usuwa konkurencję. Przykładem może być opuszczenie firmy Showmax w Polsce z dniem 31.01.2019 roku ze względu na zbyt wysoką konkurencję i mało odbiorców na polskim rynku(Grabiec, 2019). Dlaczego mówi się o Netflix'ie jako numer jeden i daję się go jako przykład globalizacji? Liczba subskrybentów Netflix w dniu 19 stycznia 2019 roku wynosiła 139 milionów osób na całym świecie, ale szacuje się, że nawet ponad 300 milionów osób korzysta z usług Netflix'a (subskrypcja jest rozdzielona na 4 osoby bo maksymalnie 4 urządzenia mogą być używane z jednego konta) (Fiegerman, 2019). W ostatnich trzech miesiącach liczba ta wzrosła o 9 milionów. Ponadto, ponad połowa subskrybentów to użytkownicy spoza Stanów Zjednoczonych, ale aż 57 milionów Amerykanów posiada aktywne konto Netflix.

W Polsce Netflix jest dostępny od 2016 roku, ale udało mu się pozyskać wielu fanów a liczba stale wzrasta. W grudniu 2018 roku liczba użytkowników Netflix'a w Polsce wynosiła prawie 3,5 miliona odbiorców, co na rynku polskim daje drugi wynik po vod.pl. Tuż za plasują się Player.pl oraz CDA(Gemius/PBI, 2019).

Rysunek 1.1 przedstawia liczbę subskrybentów serwisu VOD w Polsce. Jeżeli chodzi o ilość wykupionych subskrypcji, Netflix również znajduje się na drugim miejscu po serwisie Ipla z 775 tysiącami kont (IV kwartał 2018) (Okopień, 2019).Warto również podkreślić, że to właśnie Netflix posiada najwięcej odsłon oraz cieszy się niezmierną popularnością we wszystkich grupach wiekowych. Osiągniecie pozycji vice lidera w niecałe dwa lata to dosyć imponujący wynik — a liczba abonamentów będzie stale rosnąć.

Oczywiście olbrzymie zainteresowanie i bardzo duża ilość wykupionych subskrypcji przynoszą firmie olbrzymi zysk. Wyniki finansowe firmy są bardzo imponujące. Tylko w czwartym kwartale 2018 roku przychód wynosił 4,19 miliarda dolarów amerykań-



Rysunek 1.1. Liczba subskrybentów serwisów VOD w Polsce
Źródło: Ampere Analysis

skich. To zysk o miliard dolarów w porównaniu z ostatnim kwartale 2017 roku. Roczny dochód pochodzący z 2018 wyniósł 16 miliardów dolarów jeśli chodzi o ogólny zysk a zysk netto 1,2 miliarda(Business Insider Polska, 2019).

1.4 Interakcja klient—platforma

Aby móc korzystać z platformy Netflix obowiązkowym jest posiadania ważnej subskrypcji i konta na portalu. Pozwala ona oglądać materiały niezależnie od obecnego miejsca za pomocą wcześniej opisanych sieci strumieniowych. Na samym początku należy wybrać profil użytkownika, z którego się korzysta, a następnie przechodzi się do wyboru samego materiału, które na stronie podzielone są na kategorie oraz materiały ostatnio odtwarzane — zapisane w momencie ostatniego zatrzymania filmu co bardzo ułatwia komfortowy powrót do oglądanego materiału. Interesujące jest to, że Netflix dobiera miniatury filmów bazując na preferencjach użytkownika. Jeżeli osoba ogląda wiele materiałów komediowych, materiały zawierać będą śmieszne sytuacje lub bohaterów, a jeśli miłosne — zakochane pary(Oomen, 1970).Ma to na celu przyciąganie jeszcze większej uwagi do poszczególnych tytułów.

Ale w jaki materiały są dostępne i co musi się wydarzyć żeby materiały były widoczne na urządzeniu? Podstawowym narzędziem wykorzystywanym przez platformę Netflix jest Internet, i jest on wymagany aby odtworzyć programy (wyjątkiem jest pobranie

materiału na urządzenie, który jest potem dostępny tylko przez 48 godzin). Dzięki wykorzystaniu sieci strumieniowej firma jest w stanie szybko przesłać materiały — obraz i dźwięk z serwerów na ekran klienta. Ale zanim się to stanie, musi zostać wykonanych parę innych operacji.

Gdy tylko materiał, który ma być transmitowany zostanie wybrany, następnym krokiem jest wyszukanie najbliższego serwera, który posiada dany materiał. Odległość ta ma znaczenie jeśli chodzi o wysoką jakość materiału oraz prędkość ładowania. Następnie wybierana jest najkorzystniejsza droga przesyłania danych do dostawcy internetowego, posiadanego przez subskrybenta. To właśnie operator internetowy jest odpowiedzialny za ostateczne wrażenia klienta i jakość. Zdarza się, że przepustowość jest za niska, wtedy materiał nie jest prawidłowo wyświetlany lub następuje błąd i można zgłosić się do centrum pomocy. W innym wypadku, seans się zaczyna.

Jednakże Netflix z dnia na dzień chce podnosić jakość swoich usług i stara się podejmować decyzje biznesowe, które jeszcze bardziej polepszą standard usług. Kluczowymi działaniami oraz trendami, które napędzają Netflix do zmian i są szczególnie brane pod uwagę w bieżącym modelu biznesowym to (Oomen, 1970):

- Technologia — rozwój platform oraz jak najkorzystniejsze i bezproblemowe dostarczanie materiału dla klienta.
- Komfort — dostarczenie produktu szybko i komfortowo.
- Obsługa żądań — mobilność usługi oraz pomoc — Help Desk.
- Rozwój ilości subskrypcji — jak najwięcej aktywnych użytkowników przy jak najmniejszej opłacie.
- Zarządzane danymi — Skupienie się na systemie rekomendacji orz dbanie o przechowywanie danych i ciągnięcia z nich wniosków.

1.5 Wady i zalety platformy Netflix

Wszystko na świecie ma swój porządek, jest zdefiniowane, posiada wiele opinii oraz punktów widzenia klientów. Dla niektórych usługa jest nienaganna ale dla innych będzie ona nieintuicyjna i pełna błędów. Podobnie jest z platformą Netlfix, o której można

bardzo długo dyskutować, a ogólne korzyści i wady platformy Netflix można wymieniać w nieskończoność.

Zdecydowanie największą zaletą Netflix'a jest odtwarzanie materiałów bez względu na aktualne miejsce pobytu, o ile urządzenie posiada stały dostęp do Internetu. Usługa nie wymaga podpisywania żadnej umowy a klient nie czuje się zobowiązany do wykupienia abonamentu na określony, często długi okres czasu. Jest dostępna od razu po założeniu konta więc nie trzeba wychodzić z domu aby cieszyć się z korzystania z serwisu. Ponadto posiada olbrzymią bazę materiałów dostępnych dla użytkownika a sama firma tworzy rocznie wiele produkcji samodzielnie. Niezaprzeczalna jest również jakość materiałów (szczególnie gdy wykupi się opcję z ekranem 4K) oraz jeśli jest się w posiadaniu odpowiedniego sprzętu. Wtedy seans może przynosić wiele satysfakcji. Dodatkowo koszty można rozłożyć na czterech użytkowników, a platforma na to przystaje co jest kolejną zaletą serwisu.

W zasadzie każdy punkt wymieniony powyżej dla innych użytkowników może być kwestią negatywną, przez które rezygnują z Netflix'a. Lecz większość klientów, zgodzi się że największą wadą i niesprawiedliwością jest fakt, iż biblioteka Netflix'a, mimo że szeroko różni się wielkością, tytułami czy nawet sezonami pomiędzy krajami. Polska ma inny — mniejszy materiał niż kraje Europy Zachodniej, a Kraje Europy Zachodniej mniejsze niż Stany Zjednoczone. Warto też wspomnieć, że produkcje często pojawiają się z opóźnieniem. Kolejnym negatywnym aspektem może być fakt, że Netflix działa tylko i wyłącznie gdy posiada połącznie z Internetem. W innych sytuacjach, poza ściągnięciem materiału na urządzenie, nie będzie można odtworzyć wideo.

Jednakże zestawienie wszystkich plusów i minusów platformy, dla większości osób wypada zdecydowanie na korzyść Netflix'a. Użytkownik musi zdawać sobie sprawę o pewnych ograniczeniach i być świadomy wszystkich funkcjonalności. W końcu to legalna usługa, która oferuje bardzo szeroką bibliotekę programów.

1.6 Opis systemu rekomendacji materiałów

Platforma Netflix ostatnimi czasy pracowała nad systemem rekomendacji materiałów, dla każdego użytkownika osobno. Algorytm ten ma na celu polecanie i odkrywanie seriali i filmów, które mogą się spodobać i zaciekawić. Szacowane jest to w procentach i

oparte są o dane związane z poprzednimi oglądanymi materiałami, czasie spędzonym na oglądaniu danego gatunku czy nawet porównuje seriale z użytkownikami o podobnych preferencjach. Ponadto każdy materiał posiada krótki opis oglądanego materiału. To wszystko sprawia, że użytkownik może czuć się usatysfakcjonowany a zarazem pragnie oglądać więcej. Jednakże proces ten jest o wiele bardziej skomplikowany niż by się mogło wydawać (Krawczyński, 2018).

Użytkownik, który korzysta z serwisu zawsze korzysta z algorytmu i systemu rekomendacji, ponieważ widzi ekran główny, który sugeruje tytuły przybliżone do jego upodobań. Każdy z użytkowników logując się na platformę widzi inny widok strony. Jakie czynniki wpływają na ostateczny wygląd i polecane programy?

- Wszystkie interakcje z serwisem — obejrzane programy, oceny materiałów.
- Sugerowanie się użytkownikami o podobnych gustach.
- Informacje o obejrzanych materiałach — gatunek, kategoria, aktorzy.
- Godziny, w których programy są oglądane najczęściej.
- Średnia ilość czasu poświęcana na platformie (jeden materiał dziennie lub binge-watching czyli zjawisko kompulsywnego oglądania kilku odcinków serialu z rzędu. (Tomaszewski, 2016).
- Urządzenia, na których materiały są wyświetlane.
- Kultura, wydarzenia i zjawiska charakterystyczne dla kraju, w który posiadane jest konto.

Co ciekawe wszystkie dane związane z demografią takie jak płeć czy wiek użytkownika nie są brane pod uwagę. Podkreśla to, że dla platformy każdy użytkownik jest tak samo ważny a najważniejsze jest zadowolenie klienta i dopasowanie do jego potrzeb. Zebrane dane trafiają do algorytmów stworzone przez firmę. Należy pamiętać, że jeżeli tytuł, który nie został wyświetlony na naszej stronie głównej, a nas interesuje to zawsze jest dostępny do obejrzenia w bibliotece.

Zbieranie danych potrzebnych do stworzenia całego systemu rekomendacji zaczyna się od samego początku istnienia konta, kiedy to użytkownik musi wybrać tytuły, które już widział oraz te które mu się podobały. Na tej podstawie, już na samym początku

korzystania z serwisu system działa, a im więcej filmów i seriali się ogląda, tym system staje się dokładniejszy. Mimo, że każdy z nas widzi inny widok główny, to jednak na całym świecie znajdą się osoby o podobnym guście. Dlatego algorytmy nie ograniczają się tylko do jednego kraju, ale są globalne. Doskonałym przykładem, który uwypukla jak ważną rolę pełni zbieranie danych globalnych mogą być osoby, które prowadzą zdrowy tryb życia i interesuje ich tematyka odżywiania oraz całej branży restauracyjnej. Tylko w jednym kraju znajdzie się mało osób, które rzeczywiście interesują się tym tematem i jest potrzeba posiadania danych globalnych. Wynikiem jest to, że wszyscy z nich będą mieć w poleconych programach programy, które są związane z tym tematem (Gomez-Uribe, 2016).

System rekomendacji materiałów nie tylko wybiera odpowiednie programy, ale także pozycjonuje je w rzędy tematyczne na odpowiednim miejscu w widoku strony internetowej. Skomplikowane i tajne algorytmy i mechanizmy mają jeszcze bardziej usatysfakcjonować użytkownika. W widoku strony wyróżnia się wiele rzędów oraz trzy warstwy personalizacji.

- Rzędy które mówią o ich rodzajach — ”Oglądaj dalej”, ”Komedie romantyczne” czy ”Popularne teraz”.
- Tytuły filmów i programów wraz z miniaturką.
- Kolejność wyświetlania.

Oczywiście im rząd jest wyżej, tym bardziej jest zgodny z naszymi dotychczasowymi preferencjami a tytuł bliżej lewej strony są z danej kategorii dla użytkownika najbardziej atrakcyjne.

1.7 Problem rekomendacji

Netflix chwali się, że ich algorytm, który zbiera dane od wielu lat doskonale przyporządkowuje i proponuje materiały do obejrzenia używając wielu technologii takich jak uczenie maszynowe czy rankingi. Mimo wszystko platforma zmierza się nadal z wieloma kłopotami i próbuje rozwiązywać problemy napotkane podczas rozwoju a szczególnie w czasie wprowadzania usługi na nowe rynki. To właśnie w krajach poza Stanami

Zjednoczonym czy Kanadą stoi największe wyzwanie — jak zmaksymalizować i zbudować model poleceń nie mając dotychczas danych odnośnie użytkownika, jak to jest w przypadku krajów Ameryki Północnej gdzie dane z Netflix'a i Amazon'a zbierane są od wielu lat chociażby sugerując się danymi historycznymi dotyczące zakupionych książek i filmów z Amazon'a.

Niestety wraz ze wzrostem liczby użytkowników oraz ilością materiału algorytm Netfixa jest coraz mniej trafny, a użytkownicy skarżą się, że jeden obejrzany film może wpływać na polecenia zmieniając całkowicie preferencje oraz że wymyślony dawno temat algorytm nie jest już aktualny z obowiązującą w dzisiejszych czasach technologią. W między czasie serwis przeznacza pieniądze na sam wygląd, atrakcyjność, optymalizację oraz przede wszystkim zachęcenie i utrzymanie użytkowników. Fakt ten jest tłumaczony tym, że okres 90 sekund to czas, w którym klient decyduje czy zainwestować, wykupić pakiet lub zrezygnować. Lecz jest wiele pomysłów by jednocześnie poprawić system rekomendacji, a w to wszystko zaangażować użytkownika — dzięki temu jednocześnie go zatrzymujemy i zachęcamy a z drugiej strony poprawiamy działalność algorytmu.

Bardzo dobrym pomysłem jest pozwolenie użytkownikom na wystawianie ocen, samodzielnego grupowania ulubionych filmów i wystawiania opinii. Zachęciło by to użytkowników do większej interakcji, dzielenia się opiniami, a przede wszystkim dało by to więcej danych dla serwisu i algorytmu. Kolejno, zwiększenie przepływów danych i i ich zbieranie umożliwiłoby obserwowanie siebie nawzajem przez użytkowników, tworzenie samodzielnych list i katalogów, z możliwością wybrania tylko tych danych, które są dla niego interesujące. Dobrym pomysłem byłoby również możliwość dzielenia się oglądanymi treściami na platformach social media. W dzisiejszych czasach nie liczy się tylko ładny wygląd i rozpoznawalność marki, ale także wykorzystywana technologia i zadowolenie użytkowników. Jeśli Netflix nie zacznie poprawiać algorytmu i rozszerzać wykorzystywanej technologii, to możliwe, że użytkownicy będą odchodzić na rzecz innych rozwiązań.

Rozdział 2

Wielowymiarowe metody grupowania i klasyfikacji

2.1 Analiza skupień

2.1.1 Rola analizy skupień

Analiza skupień to pojęcie i metoda podziału dużej ilości danych obejmująca kilka różnych algorytmów klasyfikacji oraz narzędzie do eksploracyjnej analizy danych i uczenia maszynowego. Aczkolwiek grupowanie obiektów jest dla każdego człowieka zdolnością, którą nabywa wraz z rozwojem intelektualnym oraz doświadcza jej na co dzień siedząc przy osobnych stolikach w restauracji lub uczęszczając na zajęcia w różnych grupach dziekańskich. Co więcej grupowanie to jest możliwe do zaobserwowania nawet w przyrodzie. Dlatego również temat ten nie mógł zostać pominięty w nauce i statystyce.

Celem analizy jest wyróżnienie dowolnej ilości grup ze zbioru danych, tak aby obiekty znajdujące się w jednym klastrze były ze sobą jak najbardziej podobne, ale też jak najmniej powiązane z innymi wyodrębnionymi obserwacjami. Podobieństwo może być określane też jako odległość między obserwacjami. (Jiawei Han, 2012) W ten sposób powstają grupy, które są podobne do obserwacji w danym, poszczególnym klastrze niż spoza niej. (Bruce A.Maxwell, 2002) W analizie skupień, obserwacje w poszczególnych zbiorach danych informują nas o strukturze, ale nie tłumaczą ich pochodzenia i dokonanej klasyfikacji. Jednakże w sytuacji, gdy hipotezy badawcze nie są znane, jest to

dobra metoda do wyselekcjonowania mniejszych grup i dalszej użytecznej klasyfikacji. Wyniki analizy można podzielić na dwa różne zastosowania.(Balicki, 2013)

Pierwszym z nich jest wykrycie homogenicznych grup obiektów dla każdej obserwacji. Wyliczane są one dla wszystkich cech obecnych w zbiorze, które mogą przyczynić się do podziału obiektów na dobrze powiązane i skorelowane grupy. Metoda ta opiera się na tworzeniu klasyfikacji bez wprowadzanych zmian i dodatkowych cech statystycznych, a skupieniu się na naturalnych powiązaniach grupowych.

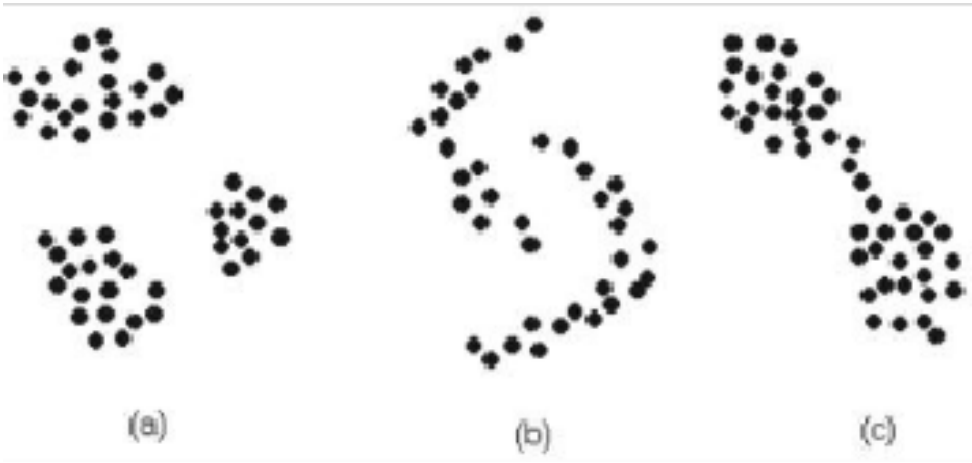
Drugim rodzajem jest ocena wymiarowości złożonego zjawiska. W metodzie tej obowiązkowym jest wybranie tylko tych wartości, które rzeczywiście są istotne dla wykonywanej analizy i bardzo dobrze odzwierciedlają jej własności. Dzięki redukcji cech powstają nadal grupy, a z czasem nawet bardziej dokładnie obliczone i skorelowane ze względu na mniejszą ilość cech. Technika ta, przez zmniejszanie zbioru danych określana jest jako efekt uboczny analizy skupień ze względu na stosowane techniki.

Niezależnie od tego, jaka metoda zostanie wybrana w celu analizy skupień, uwaga osoby przeprowadzającej analizę skupiona będzie na powstałych w wyniku jej grupach. Podzielenie zbiorowości na homogeniczne podzbiory to priorytet, ale żeby dobrze wykonać grupowanie, warto zapoznać się czym jest podzbiór. Grupa lub też inaczej klaster i skupienie to obiekty należące do tego samego zbioru łączące obiekty o podobnych cechach. Grupy powinny być między sobą odległe, aczkolwiek obserwacje wewnątrz siebie bliskie. Tak powstałe zbiory można opisywać za pomocą dwóch pojęć — wewnętrznej spójności oraz zewnętrznej izolacji. (Perner, 2007)

Na podstawie poniższej ilustracji można to doskonale zaobserwować. Rysunek 2.1 przedstawia powiązania. Kolejno (a) przedstawia trzy grupy, które są jednocześnie spójne i nie mają ze sobą powiązań, (b) — jest bardzo dobrze izolowany, aczkolwiek nie jest spójny ze względu na dużą odległość między krańcowymi obserwacjami. Rysunek (c) doskonale ukazuje spójność obiektów ale występuje brak izolacji ze względu na obserwacje pośredniczące.

2.1.2 Zalety grupowania

Podstawową czynnością, która pozwoli na dalsze badanie i analizę jest uogólnienie i porządkowanie materiału statystycznego. Musimy zdecydować, które informacje są przydatne, a które mogą przeszkadzać w utworzeniu grupowania. Dzięki temu wyod-



Rysunek 2.1. Rysunek przedstawiający rodzaje skupień
Źródło: Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne, Andrzej Balicki

rębnienie jednorodnych i podobnych do siebie grup będzie możliwe.(Piłatowska, 2006)

Grupowanie posiada również bardzo wiele korzyści. Pozwala na przyporządkowanie obserwacji tylko do jednej grupy, dzięki zaawansowanym technikom obliczeniowym. Gdyby osoba chciała podzielić zbiór danych na klastry, byłoby to wykonalne tylko dla bardzo niewielkiej liczby cech. Wraz z ich wzrostem, obserwacje należałyby do więcej niż jednej grupy. A dzięki analizie skupień można brać pod uwagę bardzo dużo atrybutów, a obserwacja będzie należeć tylko do jednego klastra.

Co więcej, jeśli przetwarzamy bardzo duże ilości danych, analiza skupień umożliwia odkrycie nieznanych i niewidocznych wcześniej powiązań między obiektami, które potem może być przydatne w modelach deskryptywnych służących do podejmowania decyzji. Dodatkowo jest pomocne w kategoryzowaniu, przetwarzaniu i interpretacji informacji co skutkuje w skutecznym przewidywaniu na podstawie powiązań zmiennych i przynależności do danego klastra.(Szeliga, 2007)

2.1.3 Algorytmy grupowania

Grupowanie, można podzielić na bardzo wiele algorytmów ale dla nich wszystkich trzeba podjąć te same kroki i pamiętać o wielu kluczowych zasadach w celu uzyskania prawidłowego grupowania. Cały proces możemy podzielić na etapy analizy i grupowania.(W.Milligan, 1996)

1. Zdefiniowanie obserwacji i zmiennych.

2. Normalizacja obserwacji i wybranie metody do obliczania odległości.

3. Klasyfikacja zmiennych.

4. Ustalenie liczby klastrów.

5. Ocena grupowania.

6. Profilowanie klas i ich interpretacja.

2.1.3.1 Zdefiniowanie obserwacji i zmiennych

Po pierwsze należy wybrać reprezentacje obiektów w postaci najczęściej obecnych cech. Należy też zdecydować czy badaniu będzie podlegać cała obserwacja i zmienne klasyfikacyjne, czy też jej próbkę pozyskaną dzięki losowaniu (podejście stochastyczne) czy też dane są nielosowe i pochodzą ze sprawozdawczości statystycznej (podejście opisowe). Selekcja zmiennych nie jest rzeczą prostą, ponieważ od niej zależeć będzie wiarygodność wyników grupowania. Branie pod uwagę cech, które nie są liczne i nie mają zdolności dyskryminacyjnych, zaburza strukturę i może być powodem do nieprawidłowej struktury klas.

2.1.3.2 Wybranie metody do obliczania odległości

Następnie należy obliczyć ich wartości podobieństwa dla wybranej grupy obiektów. Jest ono obliczane za pomocą odległości wielowymiarowej przestrzeni euklidesowej. Możemy wyróżnić trzy rodzaje obliczania odległości — Manhattan, Euklidesowa bądź Minkowskiego. Gdy żadna z powyższych metod nie zostanie wybrana, należy wtedy określić specjalną miarę odległości.

Gdy wszystkie obiekty w zbiorze danych są wartościami liczbowymi, wtedy obliczamy niepodobieństwo obiektów za pomocą zdefiniowania odległości euklidesowej pomiędzy dwoma obserwacjami — x_i oraz x_j wraz z wartością k oznaczającą wartość obserwacji x_i. Jest to podstawowy krok niezależnie od wybranej metody.

$$d_E(i,j) = \sqrt{\varepsilon_{n=1}^{k}(x_i \bar{} x_j)^2} \tag{2.1}$$

Możemy też użyć innej odległości. Przykładam może być wspomniana wcześniej odległość Manhattan(dystans jednowymiarowy). Zapisana jest ona za pomocą poniższego

wzoru:

$$d_{Mi}(i,j) = \varepsilon_{n=1}^{k} |x_i - x_j| \tag{2.2}$$

Jeszcze innym sposobem jest odległość Minkowskiego:

$$d_{Mi}(i,j) = (\varepsilon_{n=1}^{k} (|x_i - x_j|)^q)^{1/m} \tag{2.3}$$

Przedstawione powyżej metody pozwalają określić odległości między wszystkimi obserwacjami. Dodatkowo należy pamiętać o normalizacji wszystkich zmiennych przed grupowaniem by żadna z wartości nie została zdominowana lub pominięta.

Najpopularniejszymi algorytmami grupowania, które są stosowane w Data Science to metody hierarchiczne oraz iteracyjno-optymalizacyjne — ze względu na ich skuteczność oraz stosunkową prostotę. W dalszej części pracy zostaną one dokładnie opisane.

2.1.3.3 Klasyfikacja zmiennych

Dalszym krokiem jest wybór stosowanego algorytmu, a każdy z nich należy do jednej z pięciu grup. Algorytmy grupowania możemy podzielić na:

1. Algorytmy iteracyjno-optymalizacyjne — służą do tworzenia prostej i jednowymiarowej struktury danych.

2. Algorytmy hierarchiczne — służą do zbudowania hierarchii klastrów, dzieląc obserwacje na grupy bazując na podobieństwach między nimi.

3. Algorytmy gęstościowe — służą do grupowania przestrzennego zmiennych, pozwalając na wykrycie obszarów o dowolnym kształcie.

4. Algorytmy gridowe — służą do podzielenia obserwacji na znaną liczbę komórek i przydzielenia ich do nich.

5. Algorytmy modelowe — służą do przypisania obserwacji do klastra dzięki zmierzonemu prawdopodobieństwu.

Najpopularniejszymi algorytmami są metody hierarchiczne oraz optymalizujące wstępny podział zbiorów obiektów.

2.1.3.4 Ustalenie liczby klastrów

Determinowanie i zdecydowanie się na prawidłową a jednocześnie optymalną liczbę grup, na które dzielimy obserwacje jest podstawowym krokiem do jakichkolwiek dalszych obliczeń i badań. Jest to całkiem inne podejście niż w regresji czy klasyfikacji, gdyż polega na podzieleniu zbioru danych bez żadnej wiedzy na ich temat. Jest to więc technika uczenia nienadzorowanego. Dlatego też niektóre algorytmy, tak jak metoda k—średnich już na samym początku wymaga wiedzy na temat liczby skupień. Zdefiniowanie liczby grup nie jest zadaniem trywialnym oraz jednoznacznym. Wszystko to zależy od metod użytych do określania miar, odległości obserwacji a liczba grup może nie być nadal najlepsza. Dlatego też powstało wiele metod obliczeniowych, które pozwalają na określenie najbardziej prawdopodobnej liczby klastrów dla badanych obserwacji.

2.1.3.5 Ocena grupowania

Gdy grupowanie zostało przeprowadzone , ostatnim elementem jest przeprowadzenie testów dla powstałych rezultatów. Racjonalnym podejściem było by testowanie braku struktury klas na początku pracy w czasie klasyfikacji, a nie na końcu. Taka kolejność jest powodem małej użyteczności testów. Zakładamy hipotezę zerową mówiącej o braku struktury klas w zbiorze.(Marek Walesiak, 2012) Dlatego też można korzystać z trzech modeli:

1. Rozkład Poissona — przyjmuje, że odległości między wszystkimi obserwacjami w zbiorze jest taka sama odległość.

2. Rozkład jednomodalny — przyjmuje, że wszystkie obserwacje pochodzą z rozkładu częstości.

3. Rozkład losowej macierzy odległości — przyjmuje, że odległości między obserwacjami są losowe.

Dalszym sprawdzeniem jest analiza replikacji czyli przeprowadzanie grupowania jeszcze raz w celu ocenienia czy została dokonana prawidłowo. Wykonuje się ją na dwóch losowych próbach ze zbioru danych a następnie ocenia ich trafność. Schemat działania jest bardzo prosty — jeszcze raz wykonuje się klasyfikację na obu zbiorach

i oblicza odległość od centroidów. Czynność można stosować dowolną ilość razy. Następnie wylicza się miarę Randa oraz średnią zgodność, które służą do oceny jakości klasyfikacji. W zależności jaka wartość zostanie obliczona, za pomocą miernika można zinterpretować klasyfikację.

2.1.3.6 Profilowanie klas i ich interpretacja

Ostatnim etapem analizy jest interpretacja wyników. Gdy wszystkie klastry są już wyodrębnione należy je opisać — pokazać czynniki wspólne oraz wyjaśnić jakie są różnice między obserwacjami — jeśli istnieją. Dzięki temu potwierdzamy zgodność powstałej klasyfikacji. Warto też wykonać profilowanie, czyli udowodnić poprawność powstałych klastrów dzięki wskazaniu rozbieżności między grupami.

2.2 Metody iteracyjno-optymalizacyjne

Algorytmy iteracyjno-optymalizacyjne to dziś jedne z podstawowych metod bazujących na podziale obserwacji na klastry, jednocześnie nie tłumaczące czemu obserwacja należy do danej grupy. Obiekty są najpierw przypisywane do klastra, następnie następuje przenoszenie pomiędzy grupami.(Szeliga, 2007) Ponieważ liczba możliwych grupowań jest bardzo duża, trudno znaleźć jednocześnie optymalną i prawidłową liczbę klastrów, gdy bazujemy tylko na odległościach i macierzy danych. Powyższe algorytmy stosują podejście iteracyjne i zakładają, że ilość grup jest znana a za cel stawiają ulepszenie podziału. Jednym z algorytmów jest metoda k—średnich, która zostanie dokładnie opisana oraz metoda k-medoidów, gdzie każdy klaster posiada tylko pojedynczego przedstawiciela — obiekt.

2.2.1 Algorytm K-średnich

Algorytm K-średnich to metoda bardzo prosta do zrozumienia i popularnie używana służąca do nienadzorowanego uczenia maszynowego. Ma ona na celu podzielenie wszystkich obserwacji przyjmując inne założenia niż klasyczne grupowania — już na samym początku wiadoma jest liczba klastrów.(Woźniak, 2015) Algorytm K—średnich tworzony jest za pomocą poniższego schematu:

1. Na samym początku należy zdefiniować początkowy zbiór, który został już określony a następnie następuje podzielenie go na ustaloną liczbę grup. Dla wszystkich klastrów powinny być obliczone środki — centroidy — oraz odległości dla każdej obserwacji od niego. Obiekty w każdym klastrze są przyporządkowane losowo.

2. Obiekty, które zostały już przyporządkowane do grupy, nie zawsze są przypisane do niej prawidłowo. Dlatego też, następuje zmiana przynależności obserwacji do klastrów, gdzie odległość do centroida jest najmniejsza.

3. Gdy każdy obiekt należy już do prawidłowej grupy ponownie należy wyliczyć środek ciężkości wszystkich powstałych klastrów.

4. Powyższe kroki powtarzane są tak długo, dopóki wszystkie obserwacje będą należeć do klastrów o najbliższym punkcie ciężkości.

Algorytm K—średnich odróżnia się od innych modeli, które wykorzystują Data Science. Przede wszystkim, jest to algorytm, który nie jest nadzorowany. Dlatego w tym modelu należy wskazać wszystkie obserwacje i dokładny zbiór danych, które mają podlegać grupowaniu, a nie zmienną wyjściową. Ciekawym jest fakt, że algorytm ten posiada dwa zbiory wyjściowe — pierwszy to przyswojony model, drugim są wszystkie iteracje, po których nastąpiło ponowne przemieszczenie ze względu na niedopasowanie obserwacji do klastrów.

Dzięki powyższej metodzie grupowania można wyznaczyć:

- Maksymalną ilość powstałych grup — dzięki centroidom
- Początkowe położenie klastrów
- Podobieństwo i różnice między obiektami
- Ostateczny podział na klastry

Żeby określić liczbę klastrów powstało wiele metod ich wyliczenia. Dzięki nim możemy przypuszczać ile klastrów powinno powstać dla jak najlepszego grupowania i już na samym początku zainicjować ich prawidłową ilość w celu uzyskania rezultatu za pomocą K—średnich.

2.2.1.1 Metoda profilu — Silhouette wyznaczania optymalnej liczby klastrów

Pierwsza metoda służąca do wyznaczania optymalnej liczby klastrów to algorytm Silhouette. Analiza Silhouette może być użyta w celu kształtowania i wnioskowania liczby klastrów bazując na odległościach między obserwacjami. Generowany wynik i wykres przedstawia miarę jak blisko każdego punktu w jednym klastrze, znajduje się punkt w sąsiednich klastrach. Za pomocą odległości i wizualizacji możemy określić odpowiednią ilość grup. Schemat wykonywania algorytmu prezentuje się następująco:

1. Obliczenie średniej odległości dla wszystkich obiektów do obserwacji obecnych w tym samym klastrze a^i
2. Obliczenie średniej odległości dla wszystkich obiektów do obserwacji obecnych w najbliższym klastrze b^i
3. Obliczenie współczynnika

Współczynnik, który oceni poprawność powstałych klastrów oblicza się za pomocą wzoru:

$$S(u) = \frac{1}{n} n \sum_{i=1}^{n} \frac{b(i)^{-}a(i)}{max[a(i);b(i)]} \tag{2.4}$$

Współczynnik Silhouette może przyjmować wartości od <—1,1>. Interpretuje się je w następujący sposób:

- Dla wartości bliskich 1 — nastąpiło słabe przyporządkowanie do klastrów i widoczna jest nieprawidłowość.
- Dla wartości bliskich 0 — klaster jest blisko sąsiadującej grupy.
- Dla wartości bliskich 1 — klaster jest położony w znaczącej odległości od innych klastrów.

Celem każdej analizy k—średnich jest dobry podział na klastry, a gdy wartości są równe 1, grupowanie zostało przeprowadzone prawidłowo. Wtedy też, za pomocą metody Silhouette można prawidłowo wyznaczyć odpowiednią ilość grup.

2.2.1.2 Metoda łokciowa wyznaczania optymalnej liczby klastrów

Druga metoda, która służy do wyznaczania prawidłowej ilości klastrów to metoda łokciowa (ang. elbow method), określana także mianem wykresu osypiska. Jest to algorytm interpretujący i badający spójności wewnątrz każdego klastra. Stosuje się ją przy grupowaniu k—średnich, gdzie wewnętrzna wariancja w każdym klastrze jest minimalizowana. Miara wewnętrznej wariancji mierzącej spójność dla każdego klastra dąży do tego by była jak najmniejsza.

Dokładniej, metoda łokciowa analizuje procent wariancji wyjaśniony jako funkcję liczby klastrów. Na początku należy wybrać liczbę klastrów, aby dodanie kolejnego klastra nie dawało lepszego modelowania danych. W ten sposób, dzięki wykreśleniom powstaje wykres przedstawiający procent wariancji w stosunku do klastrów. Optymalna liczba klastrów wybierana jest poprzez ”kryterium łokcia”, który nie zawsze może być łatwy do zdefiniowania. Kolejność wykonywania algorytmu:

1. Pogrupowanie danych w dowolną ilość klastrów — najczęściej od 1 do 10.
2. Obliczenie wewnątrz-klastrową sumę wariancji dla każdego klastra.
3. Przedstawienie wykresu wszystkich sum wariancji dla klastrów.
4. Określenie ”punktu łokciowego” i odczytanie optymalnej liczby klastrów.

Metoda łokciowa jest często niejednoznaczna i istnieje wiele alternatywnych sposobów określenia prawidłowej liczby klastrów. Zaletą tej metody jest jednak fakt obliczenia spójności oraz wizualizacja na wykresie. (David J.Ketchen, 1996)

2.3 Metody hierarchiczne

Kolejną podstawową metodą służącą do klasyfikowania danych są metody hierarchiczne — aglomeracyjne i deglomeracyjne jak i metody określające już na samym początku wstępne grupowanie obiektów o znanej liczbie klastrów. Jak nazwa wskazuje efektem grupowania jest określenie hierarchii powstałych klastrów, bazując na podobieństwach między nimi. To co wyróżnia grupowanie hierarchiczne to fakt, że określenie liczby tworzonych klastrów na wejściu nie jest konieczne.

2.3.1 Hierarchiczne metody aglomeracyjne

Hierarchiczne metody aglomeracyjne to najbardziej popularne metody grupowania hierarchicznego. Algorytm ten jest bardzo przejrzysty i zawsze działający według takiego samego schematu, wyniki grupowań są przedstawiane klarownie i równocześnie, a co więcej można je zobrazować i przedstawić jako wykres w formie dendrogramu czyli drzewa połączeń. Drzewa hierarchiczne przedstawiają poprawne rozmieszczenie klastrów i obserwacji, które są w nich zawarte.

Algorytm hierarchiczny jako punkt wyjścia przyjmuje każdą obserwację obecną w zbiorze $A_i = (1, ...n)$ a na początku razem tworzą jedną grupę $P_(i)$. Kolejne kroki następujące prezentują się następująco:

1. Znalezienie w macierzy odległości klas o największym podobieństwu — czyli takich gdzie występuje najmniejsza odległość.
2. Zredukowanie liczby klastrów o jeden.
3. Przekształcenie odległości pomiędzy klasami.
4. Powyższe kroki powtarzane są tak długo aż wszystkie obserwacje będą obecne w jednym klastrze.

Warto zaznaczyć, że procedury mogą się miedzy sobą nieznacznie różnić. Wynika to z faktu, że odległości między obserwacji mogą być definiowane za pomocą innych miar do obliczania odległości w przestrzeni dwuwymiarowej. Mogą być to:

- Połączenia pojedyncze — czyli połączenia pomiędzy obserwacjami gdzie odległość między nimi jest jak najmniejsza spośród obiektów wszystkich klastrów. Często obserwacje łączą się w łańcuch co może powodować tworzenie klastrów o małej mierze podobieństwa, lecz zauważalne jest to tylko w celu dołączania obiektu do klastrów, a rzadziej do tworzenia nowego grupowania. Jednorodność zbioru jest bardzo ważna, więc klaster powstały za pomocą może być wadą.
- Połączenia kompletne — czyli połączenia pomiędzy obserwacjami gdzie odległość między nimi jest jak największa spośród obiektów wszystkich klastrów.
- Połączenia za pomocą średniej klasowej — czyli przyjęcie wartości pośrednich dla ustalania odległości między-klasowej.

2.3.2 Hierarchiczne metody deglomeracyjne

Kolejną metodą jest hierarchiczna metoda deglomeracyjna czyli metoda klasyfikująca przez podział. Są one przeciwieństwem do metod aglomeracyjnych — gdyż podział następuje od gotowych klastrów na mniejsze części. Na samym początku wszystkie obserwacje obecne w zbiorze — $A_i, ...A_n$. W każdym etapie grupowania liczba klastrów jest większa o jedną grupę — nie jest tworzona nowa, ale istniejąca podzielona na dwie. Po n^-1 iteracjach wynikiem jest liczba klastrów równa liczbie obserwacji w zbiorze — każdy obiekt jest oddzielnym klastrem. Jest to metoda, która znajduje swoje zastosowanie w identyfikowaniu dużych klastrów. Przebieg algorytmu przebiega w następujący sposób:

1. Należy wyodrębnić najbardziej odległe obiekty dla każdego klastra oddzielnie. Dzielony jest tylko ten klaster, gdzie odległość między obserwacjami w grupie jest największa.

2. Dla obserwacji wybranego klastra P_s liczona jest średnia odległość do innych obserwacji w grupie.

3. Obserwacja, dla której średnia odległość jest największa, stanowi początek nowego klastra A. Pozostałe obiekty tworzą klasę tymczasową — B.

4. Dla każdej obserwacji należacej do B obliczana jest średnia odległość od wszystkich obiektów w niej pozostałych oraz od obiektów w klastrze A. Powstają wtedy obserwacje: $\bar{d_{B_i}}, d^i_{A_i}$.

5. Obiekty z klasy B zostają przenoszone do klastra A dla tych których otrzymuje się $max|\bar{d_{B_i}}^-\bar{d_{A_i}}| > 0$.

6. Dla pozostałych obiektów w klasie B należy powtórzyć krok 4 i 5. Algorytm podziału klastra wybranego na początku kończy się w momencie gdy $max|\bar{d_{B_i}}^-\bar{d_{A_i}}| <= 0$.

7. Cały algorytm jest powtarzany n^-1 razy aż do momentu otrzymania liczby klastrów równej liczbie wszystkich obserwacji w badanym zbiorze aż do momentu gdy każda obserwacja ze zbioru tworzy samodzielnie oddzielny klaster.

2.3.3 Dendrogram

Efektem hierarchicznych grupowań jest najczęściej dendrogram czyli diagram ukazujący hierarchiczną relację między obiektami. Przedstawia on o wiele większą skalę niż zwykłe grupowanie, opisując miarę podobieństwa, która została wykryta w zbiorze danych. Żeby cała analiza była użyteczna i można z niej było wnioskować, podstawą jest kompleksowa wizualizacja danych i ich prawidłowe przedstawienie — do czego służą wspomniane wcześniej dendrogramy — graficzna reprezentacja grupowań.

Dendrogram to schematycznie przedstawione drzewo posiadające główny pionowy trzon wraz z odchodzącymi od niego węzłami, które przedstawiają powstałe klastry odpowiednio podzielone na sekcje. Dla klastrowania hierarchicznego i aglomeracyjnego, gdzie dendrogamy zostały oryginalnie utworzone odległości między węzłami są proporcjonalne do miary niepodobieństwa pomiędzy obserwacjami. Dla najbardziej sprzężonych typów, jest wymagane by klastry powstałe później na wykresie dendrogramu znajdowały się wyżej gdyż są monotoniczne. Z kolei dla klastrowania spornego i tworzącego podziały, gdzie normalnie obliczanie wewnątrz-klastrowej miary niepodobieństwa nie jest wymagane, odległości mogą odpowiadać w odwrotnej kolejności niż tworzenie podziałów na klastry. Dendrogramy są tym czytelniejsze im mniej węzłów i odgałęzień posiada. Złym podejściem jest również wprowadzanie wiele poziomów do dendrogramu.(Cichosz, 2015)

2.4 Klasyfikacja oparta na odległości — Metoda K Najbliższych Sąsiadów

Algorytm K najbliższych sąsiadów(ang. K Nearest Neighbours) to nieparametryczna metoda regresji i klasyfikacja powszechnie używana w statystyce do prognozowania wartości zmiennej losowej oraz do klasyfikacji obserwacji. Jest to leniwa metoda nadzorowanego uczenia maszynowego wykorzystywana w data-mining, która wykorzystuje w nich dosyć skomplikowane metody klasyfikacyjne. K najbliższych sąsiadów to algorytm łatwy do zaimplementowania — nie posiada jednoznacznej fazy wstępnej oraz nie ma żadnych założeń do wejściowych danych. Dzięki temu brane są pod uwagę wszystkie dane wejściowe — nawet te, które nie są prawidłowe ze względu na rozkład

jednorodny czy separację liniową.

Podstawową ideą algorytmu KNN jest obliczanie odległości pomiędzy wybranym punktem a wszystkimi innymi obserwacjami zawartymi w zbiorze danych. Ich odległość obliczana jest za pomocą miar jak odległości Euklidesowe czy Manhattan, które zostały opisane już w podrozdziale 2.1.3.2. Z nich wszystkich zostają wybrane najbliższe obserwacje — K, gdzie ich liczba nie jest ustalona i może przyjąć dowolną liczbę całkowitą. Na tej podstawie przypisywane są zmienne do jednego klastra, a w niej zawarte są również najbliższe obserwacje — a co za tym idzie najbardziej do siebie podobne. Wszystko to opiera się na przewidywaniu wartości wcześniej niepowiązanych i nieznanych do tych, które są już poznane. Algorytm K najbliższych sąsiadów używamy ze względu na dwie podstawowe użyteczności:

- używany do klasyfikacji — wynikiem klasyfikacji KNN jest przynależność do klastra. Pojedyncza obserwacja przypisywana jest do tej grupy, do której należą również jej najbliżsi k sąsiedzi. Liczba K to cyfra całkowita, najczęściej mała. Jeśli $k = 1$ to obiekt przyporządkowywany jest do klastra tego najbliższego sąsiada.

- używany do regresji — W regresji KNN wynikiem jest średnia wartość k najbliższych sąsiadów.

KNN używany do klasyfikacji czy regresji może mieć nadane różne wagi. Dzięki temu bliżsi sąsiedzi mogą wnosić więcej do średniej od tych dalszych. Należy pamiętać, że sąsiedzi są zawsze wybierani ze zbioru obiektów, dla którego znana jest klasyfikacja(dla klasyfikacji KNN) oraz wartość obiektu(dla regresji KNN). Charakterystycznym jest fakt, że jest on wrażliwy na lokalną strukturę i wygląd danych. Dlatego też należy dbać o poprawność danych a wynik analizy KNN będzie zadowalający.

Aby algorytm KNN był poprawnie przeprowadzony musi spełniać poniższe założenia:

- Zbiór, który będzie podlegał algorytmowi musi posiadać skończoną liczbę obserwacji, a każda z nich powinna posiadać wektor zmiennych objaśniających $X_1 ... X_N$ oraz minimum jedną zmienną objaśnianą Y.

- Zbiór posiada obserwację C z obecnym wektorem zmiennych $X_1 ... X_N$ dla którego będzie prognozowana wartość zmiennej objaśnianej Y.

- Obiekty, które są do siebie podobne w zbiorze mają tą samą decyzję, dlatego też cechy warunkujące je powinny być dobrze dobrane do wybranego problemu.

2.4.1 Algorytm K Najbliższych Sąsiadów

Klasyfikacja i dopasowanie obserwacji do odpowiednich grup w algorytmie KNN odbywa się przez podobieństwo istniejące z innymi obecnymi danymi. Podobieństwo to wynika z faktu wzajemnego porównywania do siebie obserwacji(podejście bezpośrednie) lub też poprzez uczenie się obserwacji danego zbioru(podejście pośrednie). Dzięki temu można prawidłowo klasyfikować zmienne. Algorytm K Najbliższych Sąsiadów opiera się na odległościach pomiędzy obiektami w zbiorze, których Bliskość wyliczana jest na podstawie wybranych cech zawartych w zbiorze. Odległość ta może być liczona na bardzo wiele sposobów ale podstawą jest opisanie wektorami każdej zmienną:

$$x =< x_1, ..., x_N > \tag{2.5}$$

Podobieństwo wylicza się na podstawie metryki R^m. Odległości możemy wyliczyć za pomocą metryki Euklidesowej, metryki miejskiej(Manhattan) lub metryki Minkowskiego. Metody te zostały opisane w podrozdziale 2.1.3.2 prezentującym wybrane metody do obliczania odległości. W problemie klasyfikacji wyszukujemy najbardziej pewnej decyzji dla obserwacji dla całej przestrzeni. Dlatego też wybór odpowiedniej liczebności K wybierany jest w następujący sposób:

1. Ustalana jest przestrzeń i okolica punktu.
2. Konstruowany jest histogram decyzji.
3. Wybierana jest największa wartość z histogramu.

Obiekt jest przyporządkowywany do tej grupy, gdzie jest największa liczba obserwacji z grup. Algorytm K Najbliższych Sąsiadów jest następujący:

1. Dokonywana jest standaryzacja i normalizacja danych obecnych w zbiorze danych.
2. Obliczenie odległości pomiędzy wybraną obserwacją a wszystkimi innymi wektorami obecnymi w zbiorze.

3. Posortowanie odległości od największej do najmniejszej.
4. Wyznaczenie liczby k najbliższych wektorów i zrobienie histogramu liczby k najbliższych obserwacji w zbiorze.
5. Przypisanie obserwacji do grupy najbardziej licznej w przestrzeni.
6. W przypadku sytuacji, gdy grupy są tak samo liczne to wybieramy ją losowo.

2.4.2 Zalety i wady metody KNN oraz jej zastosowanie

Algorytm KNN jest bardzo często używany w nauce i biznesie gdyż posiada wiele zalet świadczących o wielu korzyściach. Przede wszystkim jest on dosyć prosty do zaimplementowania i obsługi a przy tym daje zadowalające wyniki. Dodatkowym plusem jest szybkość wykonywania obliczeń — nie ma potrzeby do wykonywania skomplikowanych obliczeń, budowania modeli czy tworzenia hipotez i przypuszczeń. KNN jest również uniwersalny — może być zarówno użyty do klasyfikacji, regresji czy wyszukiwania. Jednakże wraz ze wzrostem obserwacji algorytm KNN może stawać się wolniejszy a więc tam gdzie wynik ma być od razu dla bardzo wiele obserwacji, nie jest to najlepsza metoda. (Drew Conway, 2012)

Algorytmy KNN są używane tam gdzie odbywa się to za pomocą wyszukiwania podobieństwa obiektów. Doskonale sprawdza się dla systemów rekomendacji, gdzie wyszukiwane są polecane produkty dla obserwacji lub ich zbioru. Jest to uzyskane za pomocą wyszukiwania podobieństwa na podstawie odległości.

Rozdział 3

System rekomendacji na przykładzie platformy Netflix

3.1 Źródła danych

W 2006 roku platforma Netflix postanowiła zaangażować informatyków, przedsiębiorstwa zajmujące się oprogramowaniem oraz studentów na całym świecie, którzy w żaden sposób nie byli związani z serwisem i ogłosiła ogólnoświatowy konkurs na najlepszy algorytm bazujący na metodzie collaborative filtering. Regułą było całkowite poleganie i sugerowanie się tylko i wyłącznie danymi dostarczonymi przez platformę czyli numerze identyfikującym, numerze filmu, ocenie nadanej przez użytkownika i dacie. Zakazanym było poleganie na innych informacjach na temat filmu oraz danymi użytkowników. Nagrodą w konkursie było wprowadzenie algorytmu na platformę oraz nagroda pieniężna w wysokości 1 000 000 dolarów. Warto podkreślić, że konkurs na algorytm trwał 5 lat aż do 2011 roku i był podzielony na kilka etapów. Przez ten okres wyłoniono kilku zwycięzców. Dlatego też zmieniał się algorytm polecający filmy, a gdy stworzono lepszy i skuteczniejszy, nowy zastępował obecny.

Zbiór danych można uznać za obszerny, mimo niewielkiej liczby cech statystycznych. Składał się on z 100 480 507 ocen w skali od 1 do 5 nadanych przez ponad 450 000 użytkowników. Zidentyfikowano w nim zbiór 17 700 filmów. Co ciekawe, liczba ocen nadanych przez użytkowników była różna — niektórzy ocenili kilka filmów ale byli też tacy, którzy oddali tysiące głosów.

W celu pozyskania większej ilości przydatnych danych na temat filmów, pobrane zo-

stały dane z serwisu zbierającego dane na temat filmów. Zostały one pozyskane dzięki OMDB API (The Open Movie Database), które stworzone jest w celu wyszukiwania informacji z bazy Internet Movie Database, czyli międzynarodowej stronie o filmach i serialach. Napisany skrypt pozwolił na pobranie wielu cennych informacji na temat filmów oraz połączenie ich z filmami, które zostały udostępnione przez platformę Netflix.

Ze względu na ograniczony dostęp do danych z platformy Netflix, ograniczono się do danych z okresu 10 lat od 1994 do 2004 roku. Konkurs organizowany był już w 2006 roku opierając się na danych z serwisu Netflix. Z powodu tego, że od tamtego czasu nie zostały ujawnione żadne dodatkowe informacje na temat filmów utrzymywanych w serwisie czy oceny użytkowników, dane w pracy nie pochodzą z ostatnich lat. W ten sposób, ograniczając się do 10 lat i biorąc pod uwagę równoczesną dostępność filmów na platformie IMDB oraz zbiorze danych z konkursu, zbiór danych ograniczył się do 816 filmów oraz 8 981 335 ocen filmów.

Dzięki interfejsowi aplikacji, dane na temat filmów zostały rozszerzone o dodatkowe informacje, które mogą okazać się przydatne do analizy, grupowań i systemu rekomendacji. Możemy podzielić je na cechy ilościowe czyli czas trwania filmu, rok wydania, ocena w serwisie IMDB, ocena na Netflix czy liczba głosów. Cechy jakościowe to głównie gatunek filmu czy reżyserowie i aktorzy.

Cechy jakościowe w pracy były szczególnie interesujące do uporządkowania oraz przetwarzania. Doskonale wiadomo, że jeden film posiada wiele gatunków filmów a były one przetrzymywane jako wartość w jednej kolumnie. Dlatego też, w celu jak najszybszego i skutecznego korzystania z danych rodzajowych filmów, postanowiono stworzyć dla każdego gatunku filmu osobną kolumnę gdzie nadałem wartości 1 — jeśli akurat dany film posiadał tę cechę a wartość 0 — gdy nie. W ten sposób dane są bardziej czytelne i łatwiejsze do przetwarzania w dalszej analizie.

Bazując na powyższych danych a dodatkowo ocenach użytkowników platformy Netflix, pogrupowano dane ze względu na podobieństwo cech kolejno na klastry przy użyciu różnych metod grupowania oraz stworzono autorski algorytm rekomendacji, który może sugerować użytkownikowi jakie filmy mogę być dla niego ciekawe. Praca oparta jest głównie na programie R, i to w nim wszystkie poniższe obliczenia, wykresy i grupowania zostały wykonane.

3.2 Analiza struktury danych wejściowych

Na samym początku pracy ze zbiorem danych warto zapoznać się z rozkładem najważniejszych cech czy podstawowymi obliczeniami statystycznymi. Dlatego też przedstawione zostaną rezultaty i wyniki związane z ocenami i filmami, posługując się przy tym różnymi rodzajami wykresów oraz metod.

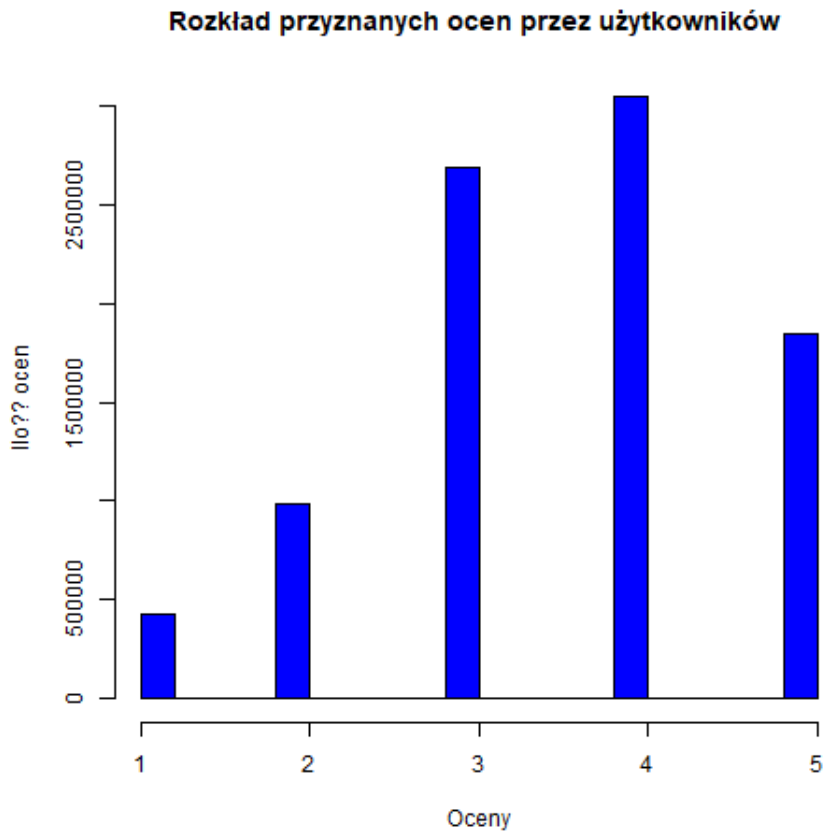
Pierwszym obszarem zainteresowania były oceny nadawane przez użytkowników poszczególnym filmom. Jest to informacja na tyle ważna, iż to oceny w dużej mierze wpływają na rekomendacje kolejnych tytułów.

Jak można zauważyć na rysunku 3.1, najczęściej nadawaną oceną przez użytkowników jest liczba 4. Może to oznaczać, że dla większości film ten był przyjemny do oglądania, ale nie zrobił bardzo dobrego wrażenia jak filmy z oceną 5, których liczba plasuje się na trzecim miejscu za oceną 3. Ponadto filmy z ocenami od 1 do 3 są dla użytkowników średnią rozrywką i pewnie nie chcieli by powtórzyć seansu. Dlatego też, w dalszej części pracy czyli stworzeniu systemu rekomendacji, zaprezentowana zostanie metoda, dzięki której użytkownicy unikną nietrafionych tytułów. Histogram rozkładu ocen filmów prezentuje się następująco:

Kolejnym elementem, na który warto zwrócić uwagę jest gatunek filmów w zbiorze danych. Jak wiadomo, jeden film może posiadać wiele cech, dlatego rozdzielono rodzaje ze zbioru danych pobranych dzięki API IMDB i podzielone zostały na oddzielne komórki. Filmy dramatyczne stanowią największą część obserwacji — ponad połowa filmów posiada taką cechę. Popularne są również filmy akcji, komedie i thrillery. Można wnioskować, że te, które posiadają takie cechy są najpopularniejsze i najczęściej spotykane, bo właśnie na nie jest największy popyt ze względu na szerokie grono odbiorców, a producenci zarabiają najwięcej pieniędzy. Rysunek 3.2 przedstawia liczbę obserwacji filmów.

3.3 Grupowanie filmów

Jeśli mówimy o dużych zbiorach danych, warto również zobaczyć czy da podzielić się obserwowane dane na podobnie strukturalnie i klasyfikacyjnie zbiory. Do tego może posłużyć nam analiza skupień. Dzięki niej obiekty, które są ze sobą mocno powiązane należą do tej samej grupy, przez co w zbiorze danych może powstać wiele klastrów. Jed-

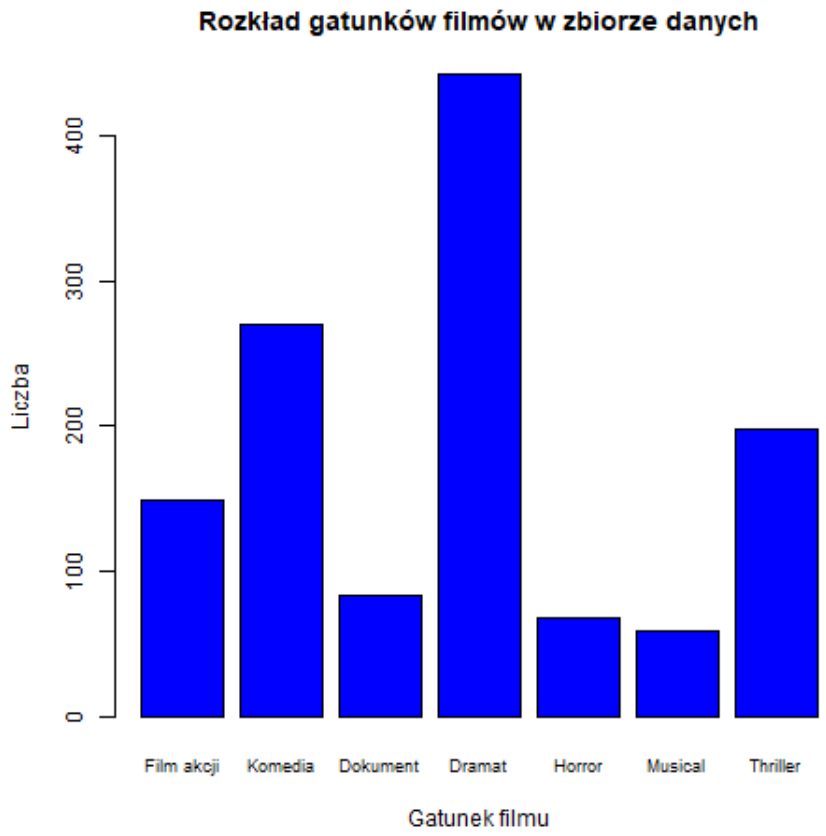


Rysunek 3.1. Rozkład przyznanych ocen przez użytkowników
Opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

nakże powyższa analiza pozwala nam tylko na odkrywanie zależności strukturalnych ale nie wyjaśnia nam czemu one występują. W pracy do przyporządkowania danych, postanowiono użyć algorytmu k—średnich i na jego podstawie podzielono je na odpowiednie grupy. Jednakże na początku należy obliczyć optymalną liczbę klastrów, które często mogą przynosić inne rezultaty. Dlatego do wybrania odpowiedniej liczby klastrów należy zawsze podejść z dystansem i brać pod uwagę nie tylko metody ale własną wiedzę, intuicję i postawiony cel grupowania. Do obliczenia optymalnej liczby klastrów oraz grupowania skorzystano głównie z pakietu ClusterR.

3.3.0.1 Zastosowanie metody Sillhoutte

Pierwszą metodą, która została użyta w celu wybrania optymalnej liczby klastrów jest algorytm Sillhoutte. Bierze on pod uwagę odległości obserwacji od innych obiektów, oraz ich średnie. Na podstawie ich wyznacza się współczynnik miary, które przyjmuje wartości od —1 do 1. W programie R istnieje pakiet oraz metoda obliczająca powyższe



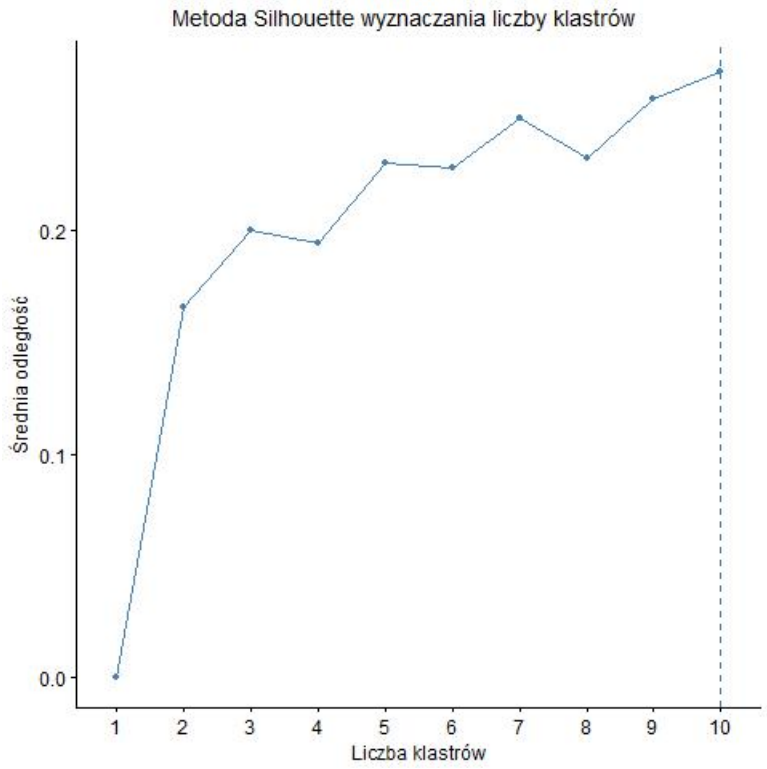
Rysunek 3.2. Rozkład obserwacji pod względem gatunków filmów
Opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

wartości i generująca wykres przedstawiającą średnią odległość. Im większa odległość klastra od innych grup tym lepiej — ponieważ wtedy obserwacje są do siebie najmniej zbliżone. Według metody Sillhoutte optymalną liczbą klastrów jest podzielenie obserwacji na 10 grup co widać na załączonym poniżej obrazku.

W tym przypadku nie przyjmować 10 grup do klastrowania metodą k—średnich. Rozdzielenie filmów na 10 grup jest zbyt szczegółowe i trywialne ponieważ w każdym klastrze znajdowałby się tylko jeden gatunek filmów, a celem jest rozdzielenie ich oraz odkrycie podobieństw do innych.

3.3.0.2 Zastosowanie metody łokciowej

Kolejną metodą, która pomaga w określeniu poprawnej liczby klastrów jest metoda łokciowa. Bada ona wewnętrzną spójność każdego klastra i analizuje procent wariancji, który dla każdego klastra jest minimalizowany. W ten sposób powstaje wykres, który przedstawia liczbę klastrów oraz stosunek wariancji do nich. Na każdym wykre-

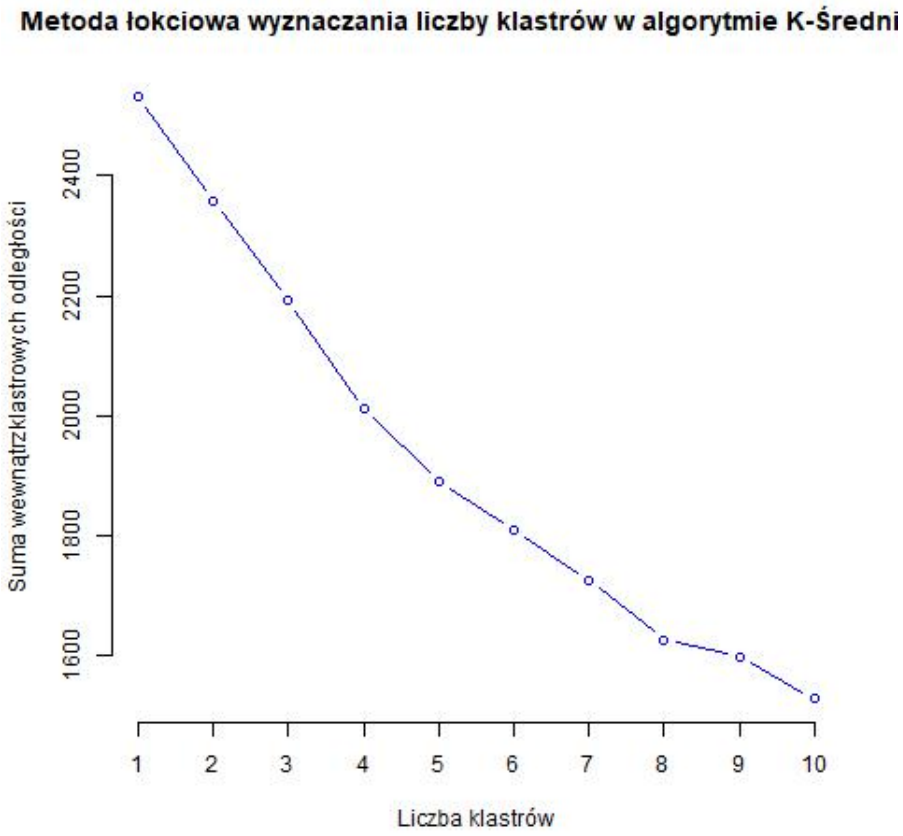


Rysunek 3.3. Metoda Sillhoute służąca do wyznaczenia optymalnej liczby klastrów
Opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

sie można wyróżnić punkt łokcia, czyli miejsce zgięcia całego wykresu. Jest to punkt kulminacyjny ponieważ stanowi granicę pomiędzy obserwacjami, w których widoczna jest zależność. W tym przypadku, przyjęta została liczba 8 jako optymalna liczba grup. Metoda łokciowa jest jedną z najskuteczniejszych oraz najczęściej wybieranych metod. Dlatego w pracy postanowiono podzielić zbiór danych na 8 klastrów.

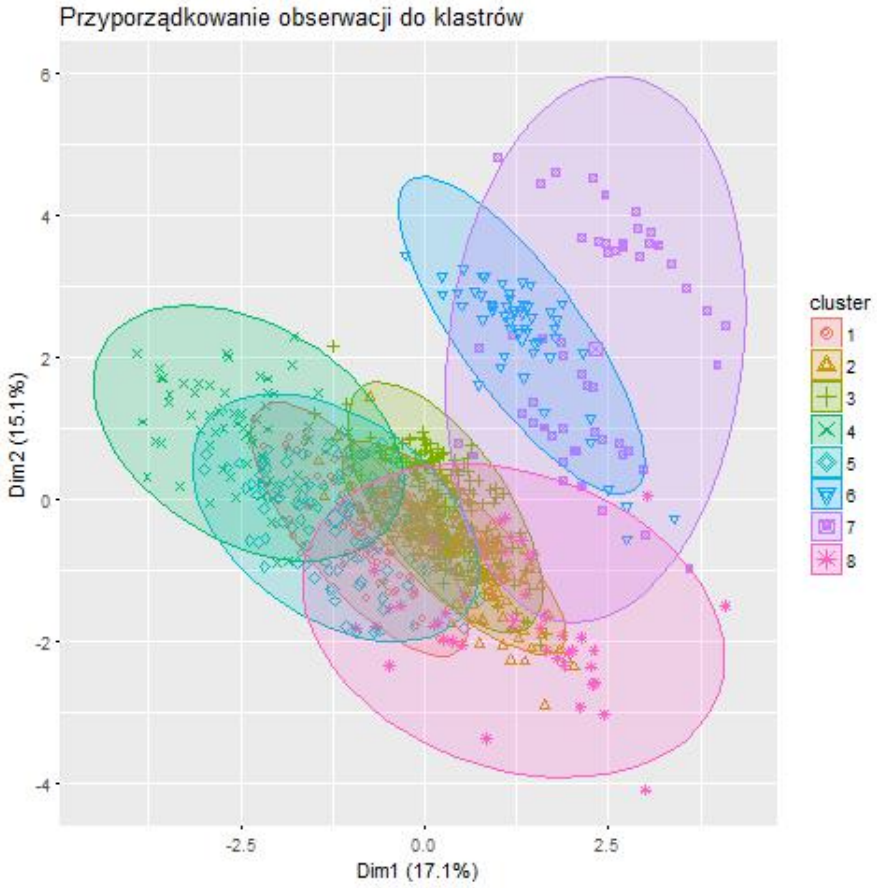
W ramach grupowania k—średnich postanowiono podzielić wszystkie filmy na 8 klastrów. Klastry te zawierają w sobie obserwacje filmów, które są do siebie podobne ze względu na cechy je opisujące. Grupowanie k—średnich w zbiorze danych Netflix Prize Data przedstawia się następująco:

Jak doskonale widać, obserwacje zostały przyporządkowane do 8 różnych grup. Każda z nich charakteryzuje się innymi cechami. Jak można zauważyć istnieją takie zbiory, które są zawarte w innym klastrze, lecz do niego nie należą. Spowodowane jest to dużą liczbą obserwacji i małą liczbą miejsca, lecz grupowanie jest przeprowadzone jest zgodne. Świadczą o tym odległości obserwacji do centroidu klastra. Gdyby zmniejszyć liczbę klastrów, grupowanie byłoby mniej dokładne, a klastry połączyły by się. Największe znaczenie w grupowaniu miały rodzaje filmów.



Rysunek 3.4. Metoda łokciowa służąca do wyznaczenia optymalnej liczby klastrów
Opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

1. Klaster pierwszy skupia filmy dramatyczne i thrillery.
2. W klastrze drugim widocznie przeważają tylko filmy dramatyczne.
3. Klaster trzeci to w większości komedie.
4. Horrory i thrillery składają się na klaster czwarty.
5. Klaster piąty to w większości filmy akcji wraz z Thrillerami.
6. W klastrze szóstym wyróżnia się głównie filmy dokumentalne.
7. Musicale i dramaty to główne gatunki filmów w klastrze siódmym.
8. Klaster ósmy nie posiada głównego gatunku — składają się na nie wszystkie rodzaje. Dodatkowo, w tym klastrze znajdują się filmy o bardzo wysokich ocenach filmu.



Rysunek 3.5. Metoda łokciowa służąca do wyznaczenia optymalnej liczby klastrów
Opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

Rozkład liczby filmów w klastrach bardzo się różni. Jest to związane z tym, że niektóre gatunki filmów są popularniejsze i chętniej oglądane od drugich. Można wywnioskować, że filmy dramatyczne, thrillery i filmy dokumentalne w tamtych latach były najchętniej oglądanymi rodzajami filmów. Rozkład liczby filmów w klastrach przedstawiony jest w tabeli 3.1.

Wszystkie grupy bardzo się od siebie różnią — niektóre skupiają głównie jeden rodzaj filmów a inne wszystkie, ze względu że obserwacje mogą posiadać wiele cech gatunkowych. Grupowanie to doskonale przydaje się w rekomendacji czy wyszukiwania podobnych obserwacji.

3.3.1 Dendrogram

Wynikiem grupowania hierarchicznego jest dendrogram czyli diagram w kształcie drzewa, który przedstawia powiązania między obserwacjami i wybranymi cechami. Interpretując dendrogram należy skupić się na wysokościach, które wskazują kolejność

Tabela 3.1. Rozkład liczby filmów w klastrach

Klaster	Liczba obserwacji
1	311
2	60
3	41
4	72
5	144
6	215
7	46
8	14

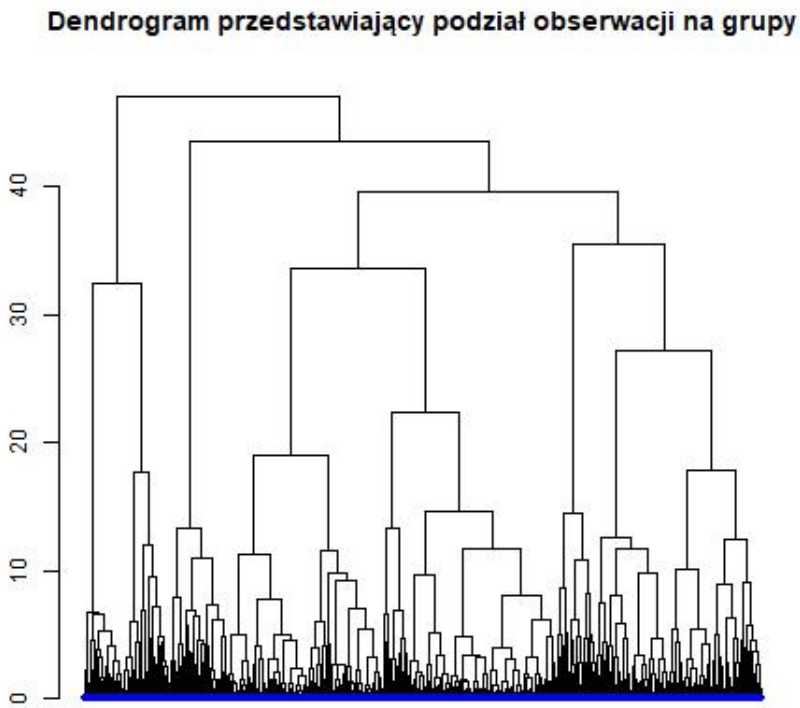
Źródło: opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

łączenia obserwacji oraz odległościach między nimi. Im znajdują się bliżej siebie, tym bardziej są do siebie podobne, a poszczególne odgałęzienia często tworzą razem klastry. Jednakże powstały dendrogram i interpretacja wyników metody grupowania hierarchicznego nie zawsze są skuteczne. W sytuacji gdy zbiór danych posiada zbyt wiele obserwacji, powstały diagram nie będzie czytelny, gdyż obserwacje nakładają się na siebie. Ponadto powstaje bardzo wiele odgałęzień, trudnych do zinterpretowania oraz analizowania powiązań. Dlatego też z wielu diagramów nie można wyciągnąć cennych interpretacji, tak jak przedstawionym na rysunku 3.6, powstałym ze zbioru danych Netflix'a. Pokazuje on rozdzielenie obserwacji na wiele grup, lecz odczytanie informacji jest utrudnione.

3.4 System rekomendacji filmów

Systemy rekomendacji są używane w bardzo wielu różnych obszarach — nie tylko dla rekomendacji filmów ale też produktów. Internet pozwala na masową personalizację, a systemy rekomendacji są jego kluczową częścią. Systemy rekomendacji budują modele dotyczące preferencji użytkowników w celu personalizacji doświadczenia użytkownika. Aktualnie są one podstawą dla serwisów społecznościowych, sklepów internetowych oraz również dla Netflix'a. W dzisiejszej erze cyfrowej firmy często mają setki tysięcy przedmiotów do zaoferowania swoim klientom — a systemy rekomendacji mogą pomóc w rozwoju biznesu — jeśli system jest dobry — lub też przeszkodzić gdy nie jest skuteczny. Algorytmy klastrowania i obliczania odległości stanowią podstawę wielu z nich.

W niniejszej pracy postawiono stworzyć własny system rekomendacji bazując na



Rysunek 3.6. Rozkład filmów przedstawiony za pomocą grupowania hierarchicznego

Opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

ocenach użytkowników oraz rozszerzonych informacjach dotyczących filmów. Systemy rekomendacji mogą być oparte na użytkownikach lub elementach na platformie. Podejście oparte na obserwacjach jest zazwyczaj preferowane w stosunku do opartego na użytkowniku. Podejście oparte na użytkownikach jest często trudniejsze do skalowania ze względu na dynamiczną naturę użytkowników, podczas gdy przedmioty zwykle nie zmieniają się zbytnio, a podejście oparte na przedmiotach często można obliczyć offline i podawać bez ciągłego ponownego szkolenia. W poniższej pracy postanowiono zaimplementować grupowanie oparte na obserwacjach, do czego posłużył algorytm najbliższego sąsiada. Jest to bardzo prosta i nieparametryczna metoda uczenia maszynowego. KNN nie przyjmuje żadnych założeń dotyczących rozkładu, ale opiera się na podobieństwie elementów. Oblicza on odległości od filmów a następnie szereguje odległości i zwraca najbliższe jako najbardziej podobne. W pracy użyto pakietu FNN służącego do obliczenia odległości między najbliższymi sąsiadami oraz stworzenia algorytmu rekomendacji.

3.4.1 Studium przypadku wybranego użytkownika

Zbiorem danych wejściowych, na podstawie których tworzony jest system rekomendacji jest uporządkowany i przetworzony zbiór ponad 800 filmów i blisko 9 milionów ocen tych filmów. Algorytm składa się z kilku kroków i bazuje na wspomnianej wcześniej metodzie KNN. Schemat algorytmu jest następujący:

1. Wylosowanie użytkownika

Każdorazowo losowo wybierany jest dowolny użytkownik ze zbioru danych. W celu pokazania działania systemu rekomendacji wylosowany został użytkownik o numerze 1318071.

2. Podział zbioru

Aby skutecznie odczytać rekomendowane tytuły i obliczyć odległości między filmami, koniecznym jest podzielenie filmów na już obejrzane i ocenione. Lista obejrzanych filmów użytkownika przedstawiona jest w tabeli 3.2.

To właśnie na podstawie wyżej wymienionych filmów powstanie rekomendacja dla użytkownika z pozostałych — jeszcze nieobejrzanych tytułów. Jednakże już z powyższych danych możemy przewidzieć jakie cechy obserwacji będą najprawdopodobniej również obecne w zbiorze filmów rekomendowanych. Można to wywnioskować dzięki interpretacji cech wcześniej obejrzanych produckji. Powyższe tytuły to głównie komedie oraz horrory — stanowią zdecydowaną większość tytułów. Mediana oceny na platformie IMDB wśród tych danych to 7/10

3. Obliczenie odległości

Kolejnym krokiem jest obliczenie odległości filmów obejrzanych od wszystkich innych.

4. Wybranie rekomendowanych filmów

Ostatnim elementem algorytmu jest wybranie najbliższych i najbardziej podobnych filmów do tych, które zostały już obejrzane przez użytkownika. Zbiór tych filmów została zamieszczona w tabeli 3.3.

Wszystkie rekomendowane filmy dzięki stworzonemu systemowi rekomendacji bazującym na algorytmie K Najbliższych Sąsiadów są adekwatne do filmów już przez niego obejrzanych. Aż 12 filmów posiada cechę filmów typu Horror a 4

tytuły są ponadto komedią. Inne gatunki filmów nie występują. Mediana filmów również wynosi 7. Na tej podstawie możemy wywnioskować, że rekomendacja przy użyciu KNN jest skuteczna i prawidłowa.

Tabela 3.2. Lista obejrzanych filmów przez użytkownika

	ID	Title
1	191.00	X2: X—Men United
2	197.00	Taking Lives
3	457.00	Kill Bill: Vol. 2
4	459.00	Basquiat
5	571.00	American Beauty
6	607.00	Speed
7	886.00	Ray
8	896.00	Dangerous Minds
9	937.00	Fallen
10	985.00	The Mummy
11	1180.00	A Beautiful Mind
12	1220.00	Man on Fire
13	1470.00	Bend It Like Beckham
14	1602.00	Dungeons & Dragons
15	1604.00	Tae Guk Gi: The Brotherhood of War
16	1632.00	Mean Creek
17	1637.00	Trapped
18	1810.00	U.S. Marshals
19	1905.00	Pirates of the Caribbean: The Curse of the Black Pearl
20	1975.00	Hollow Man
21	1983.00	Final Fantasy: The Spirits Within
22	2095.00	Liar Liar
23	2112.00	Identity
24	2152.00	What Women Want
25	2192.00	The Hurricane
26	2342.00	Super Size Me
27	2372.00	The Bourne Supremacy
28	2699.00	The Missing
29	2780.00	Dark City
30	2782.00	Braveheart
31	3015.00	The Final Cut

Źródło: opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

Tabela 3.3. Lista rekomendowanych filmów dla użytkownika

	ID	Title
1	57.00	Richard III
2	652.00	Marvin's Room
3	768.00	The Star Maker
4	1039.00	Lawn Dogs
5	1179.00	The Education of Little Tree
6	1307.00	S.W.A.T.
7	1428.00	The Recruit
8	1460.00	Soul of the Game
9	1503.00	The Boxer
10	1682.00	Absolute Power
11	1706.00	Strings
12	2331.00	Bent
13	2518.00	Things to Do in Denver When You're Dead
14	2577.00	A Walk in the Clouds
15	2931.00	Cutaway

Źródło: opracowanie własne na podstawie danych z Netflix Prize Data oraz IMDB

Podsumowanie

Systemy rekomendacji oraz grupowanie danych to przyszłość informatyki i działania każdego przedsiębiorstwa. Dlatego też warto bazować na danych oraz prawidłowo nimi zarządzać. Warto do tego wykorzystać uczenie maszynowe. Dane są przyszłością gospodarki oraz użyteczne dla przedsiębiorców by dostosowywać produkty dla klientów.

Celem niniejszej pracy było przedstawienie teoretycznych i praktycznych aspektów grupowania danych w oparciu o oceny użytkowników platformy Netflix oraz informacje z serwisu Internet Movie Database. Został także stworzony algorytm rekomendacji filmów dla użytkownika. W pracy bazowano na danych z lat 1994-2004 pochodzących z serwisu Netflix oraz Internet Movie Database.

W pracy dokonano podziału danych dotyczących filmów na 8 grup wykorzystując w tym celu algorytm k-średnich. Liczba klastrów została określona na podstawie metody łokciowej. Klastry podzielone są zgodnie gatunkami filmów oraz ocenami filmów, gdyż istnieje klaster o wyróżniających się tytułach. Gatunki filmów w klastrach są do siebie zbliżone. Ponadto długość filmu czy rok wydania też miał wpływ na przypisanie obserwacji do grupy.

Dzięki metodzie K Najbliższych Sąsiadów stworzony został algorytm rekomendacji filmów dla losowo wybranego użytkownika. Algorytm ten poleca filmy bazując na ocenach, gatunkach filmów czy ich cechach ilościowych. Ze zbioru danych wybierane są filmy, które zostały już wcześniej obejrzane przez użytkownika, a następnie wyszukiwane są ich najbliżsi sąsiedzi. Wyniki algorytmu są satysfakcjonujące, gdyż rekomendowane są tytuły powiązane i bliskie już obejrzanym tytułom - głównie pod względem gatunkowym.

Dane i systemy rekomendacji to przyszłość biznesu i technologii uczenia maszynowego. Szybkie grupowanie i zarządzanie danymi a na ich podstawie personalizacja produktów dla klienta to konieczny krok w celu dalszego rozwoju.

Bibliografia

Balicki, A. (2013). *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*. Wydawnictwo Uniwersytetu Gdańskiego.

Bruce A.Maxwell, C. S., Frederic L. Pryor. (2002). poLCA: Cluster analysis in cross-cultural research. *Journal of Statistical Software*.

Business Insider Polska, P. (2019). Zysk Netfliksa przebił oczekiwania analityków, przychody wręcz przeciwnie.

Cichosz, P. (2015). *Data Mining Algorithms: Explained using R*. Wiley.

David J.Ketchen, C. L. (1996). The application of cluster analysis in Strategic Managament Research: An analysis and critique.

Drew Conway, J. M. W. (2012). *Machine Learning for Hacking*. O'Reilly Media,

Fiegerman, S. (2019). Netflix adds 9 million paying subscribers, but stock falls.

Gemius/PBI. (2019). Netflix wyprzedził Premium CDA.pl i Playera, a HBO GO Iplę. W dół Showmax (top platform VoD).

Gomez-Uribe, C. (2016). *A Global Approach to Recommendations*. https://media.netflix.com/en/company-blog/a-global-approach-to-recommendations.

Grabiec, P. (2019). *To już koniec Showmax w Polsce. Serwis opuszcza rynek, na którym brylują Netflix i HBO GO*. https://www.spidersweb.pl/2018/12/showmax-polska-koniec.html.

Jiawei Han, J. P., Michelina Kamber. (2012). *Data Mining - Concept and Techniques*. Morgan Kauffman.

Krawczyński, J. (2018). *Jak działa algorytm rekomendacji Netflix?* https://hdtvpolska.com/algorytm-rekomendacje-netflix/.

Lidia Drabik, E. S., Aleksandra Kubiak-Sokół. (2006). *Słownik języka polskiego PWN*. Wydawnictwo Naukowe PWN.

Lusted, M. A. (2013). *Netflix, The company and its founders*. Abdo Publishing Company.

Marek Walesiak, E. G. (2012). *Statystyczna analiza danych z wykorzystaniem programu R*. Wydawnictwo Naukowe PWN.

Netflix launches in nearly every country but China. (2016).

Okopień, P. (2019). Ponad 700 tysięcy abonentów Netflixa w Polsce – szczegółowe dane o stanie VOD.

Oomen, M. (1970). *Netflix: How a DVD rental company changed the way we spend our free time*. https://www.businessmodelsinc.com/exponential-business-model/netflix.

Paszkowski, M. (2019). Netflix planuje pobić kolejny rekord w tym roku. Jeszcze więcej miliardów na własne produkcje!

Perner, P. (2007). *Machine Learning and Data Mining in Pattern Recognition*. Springer.

Piłatowska, M. (2006). *Repetytorium ze statystyki*. Wydawnictwo Naukowe PWN SA.

Stępka, P. (2009). Rynek wideo na żądanie (VoD) w Polsce.

Szeliga, M. (2007). *Data Science i uczenie maszynowe*. Springer.

Tomaszewski, M. (2016). *Binge-watching – jasne i ciemne strony szału oglądania*. https://lekturaobowiazkowa.pl/na-ekranie/binge-watching-jasne-i-ciemne-strony-szalu-ogladania/.

W.Milligan, G. (1996). Clustering validation: Results and implications for applied analyses.

Woźniak, K. (2015). *Narzędzia analityczne w naukach ekoomicznych*. Mfiles.pl.

Ziółkowska, W. S. M. (2011). *Efektywność mediów strumieniowych*. Biuletyn Instytutu Automatyki i Robotyki.

Spis tabel

Spis rysunków

Kody języka R

```
library(tidyverse)
library(zoo)
library(readr)
library(dplyr)

#combined_data1
cd1 <- read.table("combined_data_1.txt")
cd1_s <- cd1 %>%
  separate(V1, c("custid", "rating", "date"), sep = ",")
cd1_s <- cd1_s%>%
  mutate(custid=as.numeric(gsub(":", "", custid)),
         movieid=ifelse(is.na(rating) & is.na(date), custid, NA),
         movieid=na.locf(movieid),
         rating=as.numeric(rating))

#combined_data2
cd2 <- read.table("combined_data_2.txt")
cd2_s <- cd2 %>%
  separate(V1, c("custid", "rating", "date"), sep = ",")
cd2_s <- cd2_s %>%
  mutate(custid=as.numeric(gsub(":", "", custid)),
         movieid=ifelse(is.na(rating) & is.na(date), custid, NA),
         movieid=na.locf(movieid),
         rating=as.numeric(rating))

#combined_data3
cd3 <- read.table("combined_data_3.txt")
cd3_s <- cd3 %>%
  separate(V1, c("custid", "rating", "date"), sep = ",")
cd3_s <- cd3_s%>%
  mutate(custid=as.numeric(gsub(":", "", custid)),
         movieid=ifelse(is.na(rating) & is.na(date), custid, NA),
         movieid=na.locf(movieid),
         rating=as.numeric(rating))

#combined_data4
cd4 <- read.table("combined_data_2.txt")
cd4_s <- cd2 %>%
  separate(V1, c("custid", "rating", "date"), sep = ",")
cd4_s <- cd4_s %>%
  mutate(custid=as.numeric(gsub(":", "", custid)),
         movieid=ifelse(is.na(rating) & is.na(date), custid, NA),
         movieid=na.locf(movieid),
         rating=as.numeric(rating))

final1 <- cd1_s %>%
filter(movieid %in% c(database_movies$ID))
final2 <- cd2_s %>%
filter(movieid %in% c(database_movies$ID))
final3 <- cd3_s %>%
filter(movieid %in% c(database_movies$ID))
final4 <- cd4_s %>%
filter(movieid %in% c(database_movies$ID))

write.table(final1,"final1.txt",sep=",",row.names=FALSE)
write.table(final2,"final2.txt",sep=",",row.names=FALSE)
write.table(final3,"final3.txt",sep=",",row.names=FALSE)
write.table(final4,"final4.txt",sep=",",row.names=FALSE)

RatingsNew2 <- read.table("RatingsNew2.txt",header=TRUE,sep = ",")
```

Listing 3.1. Kod programu napisanego w języku R, służącego do załadowania danych z serwisu Netflix do programu

```r
library(tidyverse)
library(imdbapi)
library(sqldf)

movies <- read.csv("movie_titles.csv", header = F)
names(movies) <- c("no", "year", "name")

movies_10 <- movies %>%
  mutate_all(as.character) %>%
  mutate(year=as.numeric(year)) %>%
  filter(year %in% c(1994:2004))

omdb_api_key(force = T)

movies_imdb <- data.frame()

for(i in 1:10000000){
  tryCatch({
    m <- find_by_title(title = movies_10$name[i], year_of_release = movies_
       10$year[i], api_key = omdb_api_key())
    if(nrow(m) > 0){
      movies_imdb <- union_all(movies_imdb, m)

    }
    m <- find_by_title(title = movies_10$name[i], year_of_release = movies_
       10$year[i], api_key = omdb_api_key())
  }, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})

}
```

Listing 3.2. Kod programu napisanego w języku R, służącego do pobrania dodatkowych danych dzięki API IMDB

```r
library(VIM)
library(readr)
library(cluster)
library(factoextra)
library(flexclust)
library(DT)
library(ape)
library(dplyr)
library(factoextra)
library(NbClust)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(ClusterR)

Ratings <- read.table("Ratings2.txt", header = TRUE, sep = ",")
save(Ratings, file = "ratings.RData")
load("ratings.RData")

mean <- Ratings %>%
  group_by(movieid) %>%
  summarise(netflixVotes=n(),
            netflix_sr=mean(rating))
save(mean, file = "filename.RData")
APIandNETFLIXcombined2 <- inner_join(APIandNETFLIXcombined, mean, by=c("ID"
    ="movieid"))
imdb_netflix <- APIandNETFLIXcombined2 %>%
  filter(!is.na(Runtime))

mydata <- imdb_netflix
mydata <- mydata[!duplicated(mydata$Title),]

mydata <- imdb_netflix[!duplicated(imdb_netflix$Title),]

mydata3 <- mydata %>%
  select(ID:imdbVotes, netflixMean=netflix_sr, netflixVotes, Action, Comedy
      , Documentary, Drama, Horror, Music, Thriller) %>%
  group_by(Title) %>%
  mutate(genre_sum=sum(Action, Comedy, Documentary, Drama, Horror, Music,
      Thriller)) %>%
  filter(genre_sum!=0) %>%
  ungroup() %>%
  select(-genre_sum)

# standaryzacja
cechy_z <- mydata3 %>%
  select(Year, Runtime, netflixMean:Thriller) %>%
  scale()
Optimal_Clusters_KMeans(cechy_z, max_clusters = 20, criterion = "WCSSE")

grupowanie <- KMeans_rcpp(cechy_z, clusters = 8)

mydata3$grupy <- grupowanie$clusters

mydata3 %>%
  count(grupy)

mydata3 %>%
  select(Year, Runtime, netflixMean:Thriller, grupy) %>%
  group_by(grupy) %>%
  summarise_all("mean")

mydata3 %>%
  filter(grupy == 1)

#obliczenie liczby klastr w
number <- NbClust(cechy_z, distance = "euclidean",
                  min.nc = 2, max.nc = 9,
```

```
                  method = "complete", index ="all")

#metoda lokciowa
jpeg('lokciowa.jpg')
plot(stepFlexclust(cechy_z, nrep=6, 2:10),type="l",col="blue",xlab="Liczba
    klastr w",ylab="Suma wewn trzklastrowych odleg o ci" ,main="Metoda
     okciowa  wyznaczania liczby klastr w w algorytmie K- rednich ")
dev.off()

#metoda Silhouette
jpeg('sillohoute.jpg')
fviz_nbclust(cechy_z, kmeans, method = "silhouette") +
  ggtitle("              Metoda Silhouette wyznaczania liczby klastr w")
    +
  labs(x = "Liczba klastr w", y = " rednia  odleg o  ")
dev.off()

#Grupowanie k- rednich
head(skalaObserwacji, n = 3)
set.seed(123)
km.res <- kmeans(cechy_z, 5, nstart = 25)
print(km.res)
head(km.res$cluster, 3)
km.res$size
fviz_cluster(km.res,data=cechy_z, geom = "point", frame.type = "norm")+
  ggtitle("Przyporz dkowanie obserwacji do klastr w")


h <- data.frame(cechy_z)


#wstepne odleg osci
odleglosci<- dist(t(cechy_z))
plot(odleglosci,main="Odleg o ci w zbiorze danych mi dzy obserwacjami",
    col="blue",
       xlab="",ylab="")

#Dendrogram
jpeg('dendrogram.jpg')
dd <- dist(scale(cechy_z), method = "euclidean")
hc <- hclust(dd, method = "ward.D2")
plot(hc, hang = -1, cex = 0.6)
hcd <- as.dendrogram(hc)
nodePar <- list(lab.cex = 0.5, pch = c(NA, 19),
                cex = 0.8, col = "blue")
plot(hcd, nodePar = nodePar, leaflab = "none", main="Dendrogram
    przedstawiaj cy podzia  obserwacji na grupy")
dev.off()

set.seed(100)
model <- kmeans(h, centers = 3, nstart = 20)
table(model$cluster)
fviz_cluster(model, data=odleglosci)
```

Listing 3.3. Kod programu napisanego w języku R, służącego do obliczania odległości, powiązań, algorytmów i grupowania za pomocą k-średnich

```r
library(FNN)
library(dplyr)
library(tidyr)

knn10 <- get.knn(cechy_z, k = 10)

knn10_movieid <- as.data.frame(knn10$nn.index)
names(knn10_movieid) <- paste0("sasiad",1:10,"id")

mydata3_knn <- cbind(mydata3, knn10_movieid)

knn10_moviedist <- as.data.frame(knn10$nn.dist)
names(knn10_moviedist) <- paste0("sasiad",1:10,"dist")

mydata3_knn <- cbind(mydata3_knn, knn10_moviedist)

mydata3$index <- 1:nrow(mydata3)

netflix_rat <- Ratings %>%
  filter(movieid %in% mydata3$ID)

uid <- sample(netflix_rat$custid, 1)

filmy_user <- netflix_rat %>%
  filter(custid == uid)

filmy_obejrzane <- mydata3_knn %>%
  filter(ID %in% filmy_user$movieid) %>%
  select(ID, Title, sasiad1id:sasiad10dist)

filmy_obejrzane_id <- filmy_obejrzane %>%
  select(ID, Title, sasiad1id:sasiad10id) %>%
  gather(sasiad, sasiad_id, -ID, -Title) %>%
  select(-sasiad)

filmy_obejrzane_dist <- filmy_obejrzane %>%
  select(ID, Title, sasiad1dist:sasiad10dist) %>%
  gather(sasiad, sasiad_dist, -ID, -Title) %>%
  select(-sasiad)

filmy_obejrzane_id_dist <- filmy_obejrzane_id
filmy_obejrzane_id_dist$sasiad_dist <- filmy_obejrzane_dist$sasiad_dist

id_reco <- filmy_obejrzane_id_dist %>%
  arrange(sasiad_dist) %>%
  slice(1:15) %>%
  .$sasiad_id

filmy_reco <- mydata3 %>%
  filter(index %in% id_reco)
```

Listing 3.4. Kod programu napisanego w języku R, służącego do stworzenia systemu rekomendacji filmów dla losowo wybranego użytkownika