YouTube trending videos

Autorzy: **Bartosz Górka** INF127228 bartosz.gorka@student.put.poznan.pl
Kajetan Zimniak INF127229 kajetan.zimniak@student.put.poznan.pl

28 maja 2020

1 Wstęp

Celem projektu było przeprowadzenie procesu odkrywania wiedzy z rzeczywistych danych. W szczególności: pozyskania danych z różnych źródeł, przetworzenia ich do dogodnej reprezentacji, oceny jakości danych, oceny ważności atrybutów, poszukiwanie współzależności między nimi, odkrycia użytecznych i potencjalnie interesujących regularności, oraz dokonanie interpretacji znalezionych regularności.

Projekt przygotowany w ramach kroków, które zostaną krótko podsumowane poniżej.

2 Podsumowanie etapów

2.1 Atrybuty tekstowe

Pierwszy etap projektu polega na przygotowaniu wstępnych statystyk danych, wykorzystaniu metod wizualizacji, zapoznaniu się z danymi oraz ich jakością. Pod uwagę wzięto atrybuty tekstowe i liczbowe, z pominięciem obrazków, które będą analizowane w kolejnym etapie.

W pierwszej kolejności wprowadzono dodatkowe atrybuty, które miały służyć pewnego rodzaju normalizacji atrybutów liczbowych. Dodano stosunek liczby polubień, łapek w dół oraz komentarzy do liczby wyświetleń, oraz stosunek liczby polubień do łapek w dół.

Przedstawiono również statystyki dotyczące liczby wyświetleń w zależności od pory dnia i dnia tygodnia, sprawdzono czy istnieje różnica pomiędzy USA a Wielką Brytanią (nasze dane pochodzą z tych krajów), oraz zbadano tendencję aktywności użytkowników na przestrzeni badanych lat.

Podczas analizy tytułów i tagów zwrócono szczególnie uwagę na ich długość (zarówno liczbę słów jak i znaków), powtórzenia, liczbę użytych punktorów ('!', '?', '...' itd.), wielkość liter, częstotliwość występowania tzw. *stopwords*. Uznano, że ważnym elementem jest również dobór słów używanych w tytułach, tagach i opisach. Sporządzono ranking najczęstszych z nich.

Sprawdzono także, które kategorie cieszą się największą popularnością oraz ile filmów zezwala na komentarze i oceny.

Analizie poddano również korelację pomiędzy atrybutami.

2.2 Atrybuty wizualne

Na początku etapu dokonaliśmy poprawy zbioru danych. W poprzednim etapie zauważyliśmy obecność NAZWA? jako identyfikatora filmu, co było błędem w danych. Bazując na strukturze adresów URL jakie są generowane przez YouTube przygotowaliśmy poprawę identyfikatora uwzględniając atrybut thumbnail_link.

W naszym problemie używamy już częściowo danych z YouTube Data API, które udostępnia nam obrazy klipów video w wyższej rozdzielczości. Takie zachowanie jest podyktowane przeprowadzonym eksperymentem. W przypadku obrazów dostarczanych w oryginalnym zbiorze, mamy do czynienia z najmniejszą rozdzielczością. Powoduje to straty w odczycie tekstu czy wykrywaniu obiektów. Zgodnie z oczekiwaniami projektu, chcemy wykorzystać wszelką dostępną wiedzę i wyciągnąć jak najwięcej informacji z danych.

Dokonując oceny eksperymentalnej obserwujemy znacznie lepsze rozpoznawanie obiektów i tekstu przy możliwości wykorzystania większych rozdzielczości.

Dokonaliśmy analizy wykrytego tekstu oraz jak przekłada się obecność opisu na zainteresowanie filmem. Z naszych obserwacji wynika, że zarówno tekst jak i obrazek powinny nawiązywać do tego samego i wzbudzać zainteresowanie.

Zainteresowaliśmy się również obiektami oraz emocjami na obrazie. Mimo kosztownego obliczeniowo i bardzo czasochłonnego przetwarzania, udało się przeanalizować całą kolekcję w celu wykrycia obiektów oraz rozpoznania ludzkich emocji na obrazie. Na miniaturkach filmów warto umieszczać obiekty które nawiązują do tematyki, skłaniają odbiorcę do zapoznania się z materiałem. Jeżeli chodzi o rozpoznane obiekty to dominują ludzie, czego się spodziewaliśmy. Do tego dochodzą części ubrania i elementy życia codziennego - kubek, książka, TV, pies, laptop.

2.3 Ocena ważności atrybutów i ich ewentualna redukcja

W tym etapie skupiliśmy się na ocenie już posiadanych atrybutów. Przedstawiliśmy listę atrybutów, ich znaczenie oraz wskazaliśmy które można wyeliminować gdyż nie są przydatne w naszym problemie. Każde odrzucenie poparliśmy wyjaśnieniem, naszym komentarzem dlaczego właśnie tak czynimy.

Dokonana została analiza korelacji pomiędzy atrybutami aby wyeliminować te, które są ze sobą ściśle powiązane.

Sprawdziliśmy również zmienność atrybutów, aby na tej podstawie móc sugerować zachowanie klientowi. Chcieliśmy zastosować redukcję wymiarowości, jednakże obawialiśmy się jak interpretować takie modele. Stosując chociażby drzewa decyzyjne nie jesteśmy w stanie powiedzieć jasno jaka jest reguła, aby film był oznaczony jako trending kiedy zastosowano chociażby PCA.

2.4 Wykorzystanie uczenia pół-nadzorowanego, uzupełnienie kategorii

Zbliżając się do końca projektu, uzupełniliśmy dane dotyczące kategorii przypisanych do filmów. W naszym podejściu zastosowaliśmy metodę opartą o podział na k-grup bazując na liczbie zaetykietowanych unikalnych kategorii filmów. Zastosowana została metoda cluster and label, która przypisuje obiekty do jednej z k-grup, aby następnie przypisać etykietę klasy.

Kolejnym krokiem było zastosowanie pseudo-labeling czyli procesu dodawania pewnych przewidywanych danych testowych do danych treningowych. Na samym końcu zastosowano gotowe metody z pakieto sklearn tj. label-spreading.

2.5 YouTube API - zbieranie danych i weryfikacja wyników

W tym etapie połowa zadań została zrealizowana w etapach wcześniejszych. Nasze filmy już miały wcześniej przypisane oczekiwane kategorie. Również w poprzednim etapie dokonaliśmy analizy trafności naszego klasyfikatora w oparciu o faktyczne, rzeczywiste dane, pochodzące z YouTube Data API.

Pozyskaliśmy dane non-trending w oparciu o YouTube Data API, gdzie zdecydowaliśmy się na wykorzystanie parametru relevantTo. Opisaliśmy znane nam problemy z ograniczeniem zbioru danych, wskazaliśmy potencjalne inne sposoby podejścia do problemu wczytania danych. Mimo naszych obaw o brak danych spoza kanału twórcy, okazało się że około połowa propozycji nie jest od tego samego autora. Szczegóły wskazujemy w naszej pracy.

Pozyskane dane uspójniliśmy tj. dokonaliśmy takich samych operacji przypisania nowych atrybutów, jak w przypadku danych trending. Mimo jeszcze bardziej czasochłonnego procesu, udało się nam całość przygotować jako nowy zbiór danych.

Pozbyliśmy się także części kategorii - tych, które nie były zbyt liczne (jak chociażby przykład z kategorią zawierającą mniej niż 5 przykładów).

2.6 Klasyfikator, reguły, profil charakterystyczny i wiedza dla YouTubera

Ostatnim już etapem było przygotowanie klasyfikatora i reguł dla YouTubera. W naszym podejściu zdecydowaliśmy się na wykorzystanie dwóch prostych, ale dobrych metod rozpoznania trending / non-trending video.

Pierwsza metoda to drzewo decyzyjne. Ta prosta metoda pozwala na bardzo łatwą interpretowalność, która jest kluczowa w naszym projekcie. Ograniczając rozmiar drzewa oraz je wizualizując, można bardzo łatwo wyprowadzić zestaw reguł przypisujących filmom kategorie trending / non-trending. Zadanie wskazania reguł pozostawiamy czytelnikowi z uwagi jego trywialność.

Drugim z modeli był NaiveBayes. Model opisujący dane poprzez rozkłady, nie okazał się zbyt dobry w praktyce. W niewielkiej tylko części dobrze przypisywał etykiety klasie mniejszościowej, znacząco promując klasę większościową.

Porównaliśmy modele pod względem kilku wybranych miar oceny jakości. Ostatecznie wybraliśmy drzewo decyzyjne z uwagi na jego wysoką interpretowalność.

3 Wskazówki dla klienta

W naszej analizie zaobserwowaliśmy duże zainteresowanie oficjalnymi teledyskami i zwiastunami filmów. Niestety nie możemy wymagać tego od klienta, aby był w stanie takie materiały tworzyć.

Jednakże bazując na tej obserwacji, a także na obserwacji dotyczącej zainteresowaniu pewnego rodzaju sezonowością zainteresowań odbiorców możemy zasugerować, aby nagrania klienta poruszały tematy popularne w danym czasie. Jako przykład możemy wskazać ogromne zainteresowanie Star Wars w okresie świątecznym, kiedy właśnie odbywała się premiera.

W przypadku tytułów - powinny one zainteresować odbiorcę. Unikamy w nich zbędnych zwrotów, stosując proste słownictwo i prostą formę przekazu.

Dominującą kategorią jest ogólnie przyjęta rozrywka. Dotyczy to zarówno samej kategorii rozrywka, jak i muzyka czy sport.

W opisie filmów zalecane jest wskazanie odwołań do innych materiałów, opisanie o czym jest dany film. Ponadto referencje do portali społecznościowych, aby budować swoją markę i przywiązanie odbiorców do naszej twórczości. Również tagi powinny określać o czym jest dany film. Stosowanie wielu tagów pozwala na odnalezienie naszej produkcji w sytuacji, kiedy odbiorca stosuje filtry.

Polubienia naszego materiału mają istotny wpływ na jego zasięg oraz liczbę komentarzy. Niezbyt lubiane materiały nie mają zbyt wielu szans na zaistnienie na liście trending.

Również miniatura filmu powinna być ciekawa dla potencjalnego odbiorcy. Jako obiekty dominują ludzkie twarze - wskazanie autora danego filmu, twarzy osoby która tworzy film jest zbliżeniem się do odbiorcy. Oglądający może się "wczuć" w autora, utożsamiać się z nim. Obrazy nie powinny być jednolite, warto użyć tzw. cutscene czyli wycięcia klatki filmowej z zawartości, dodać napis - to pozwala wskazać, co będzie się działo w filmie.

4 Informacje dodatkowe

W wyniku pracy nad projektem, udało się przygotować kod dostępny w ramach repozytorium GitHub pod adresem `https://github.com/bartoszgorka/youtube-trending-videos`.
Dodatkowo kod dostępny jest w ramach usługi Google Colab pod adresem
`https://colab.research.google.com/drive/1ZJ4jhiWV6wQhGSoc6De9rhAHPNl8aGn-`