Projekt: YouTube trending videos

Celem projektu jest przeprowadzenie procesu odkrywania wiedzy z rzeczywistych złożonych danych. W takim podejściu należy dokonać właściwego pozyskania danych z różnych źródeł, przetworzenia ich do właściwej reprezentacji, oceny jakości danych, oceny ważności atrybutów, poszukiwanie współzależności między nimi, odkrycia użytecznych i potencjalnie interesujących regularności, i później ew. skonstruowania modelu klasyfikacyjnego. Należy także dokonać interpretacji i oceny znalezionych regularności - co jest powiązane z tzw. interpretowalnością wyników eksploracji danych. Z metodologicznego punktu widzenia sugeruje się wykorzystywanie poznanych metod analizy danych zarówno statystycznych jak i wywodzących się ze sztucznej inteligencji, w tym w szczególności uczenia maszynowego.

Wybrane dane dotyczą filmów z serwisu YouTube, które były w przeszłości proponowane użytkownikom w zakładce Trending. W oparciu o udostępnione dane, należy opracować strategię dla nowego youtubera (nazwijmy go roboczo Franek). Co powinien zrobić Franek, aby jego filmy miały większą szansę trafić do zakładki Trending? Oczywiście należy przyjąć realistyczne oczekiwania: nie możemy się spodziewać, że Franek założy zespół pop, bądź że wystartuje w telewizji ze swoim autorskim wieczornym programem rozrywkowym.
Niestety, trzeba zauważyć, że dane są surowe i niekompletne, w szczególności brakuje w nich przydziału do kategorii dla wielu z filmów. Franek nie udostępnił Ci również danych dotyczących popularnych filmów spoza kategorii trending - będziesz musiał/a uzyskać je na własną rękę, korzystając z YouTube API.

- **Etap 1 - Atrybuty tekstowe**
    - Wstępne statystyki danych, wykorzystanie metod wizualizacji, zapoznanie się z danymi oraz ich jakością; identyfikacja braków
    - Zmiana reprezentacji danych: atrybuty oparte na opisie, tytule i ew. innych atrybutach (bez obrazków)
        - Występowanie słów (jakie słowa są szczególnie informatywne dla naszego problemu?)
        - Atrybuty oparte na tytułach i opisach: długość, interpunkcja, wielkie litery, obecność linków itp.
        - Czas uploadu do youtube
        - Jakie atrybuty da się wykorzystać? Jakich nie? Dlaczego?
    - OCENA (15%)
        - 5% podsumowanie danych, wstępne statystyki
        - 6% atrybuty tekstowe
        - 4% inne atrybuty
- **Etap 2 - Atrybuty wizualne**
    - Cechy z mikro obrazów ang. thumbnaili - czy jakieś charakterystyki szczególnie często występują?
        - Atrybuty: Hand-crafted? Nauczone sieciami? Schematy kolorystyczne?

        - Rozważyć wykorzystanie wytrenowanych sieci imagenet, wektorów Fishera lub innych (do rozpoznawania emocji na twarzach, odczytywania napisów itp.)
    - OCENA (15%)
        - 7% ręcznie zaprojektowane atrybuty
        - 8% elementy występujące na thumbnailu (wykrywanie)
- **Etap 3 - Ocena ważności atrybutów i ich ewentualna redukcja**
    - Pozyskanie tzw. ground truth z YouTube API (spójnego z ustaloną charakterystyką danych)
    - Ocena atrybutów
    - Znalezienie korelacji między atrybutami, poszukiwania atrybutów niewnoszących informacji, usunięcie atrybutów które mogą być niepotrzebne
    - Selekcja
    - OCENA (15%)
        - 3% pozyskanie ground truth z YouTube API
        - 7% analiza korelacji/innych miar między atrybutami, atrybutów nieprzydatnych itp.
        - 5% selekcja atrybutów (z wyjaśnieniem)
- **Etap 4 - Wykorzystanie uczenia pół-nadzorowanego, uzupełnienie kategorii**
    - Uczenie pół-nadzorowane do celów uzupełnienia info o kategoriach
    - Więcej niż tylko jedna metoda, ale z dobrym wytłumaczeniem dlaczego taka metoda a nie inna
    - Weryfikacja wyników w oparciu o porównanie uzyskanych wyników z tzw. ground truth
    - Porównanie metod, wybór jednej z nich
    - OCENA (20%)
        - 8% pierwsza metoda (z wytłumaczeniem)
        - 7% druga metoda (z wytłumaczeniem)
        - 5% porównanie metod, wybór jednej
- **Etap 5 - YouTube API - zbieranie danych nie-trending**
    - Skorzystanie z YouTube API aby pozyskać dane nie-trending
    - Dane powinny być kompatybilne z naszymi trending - generalnie ten sam okres, równie popularne, zgodność atrybutów, itd.
    - Oczyszczenie i zintegrowanie danych non-trending
    - Odfiltrowanie nieinteresujących nas kategorii
    - Ewentualne zweryfikowanie ważności atrybutów wybranych we wcześniejszym etapie
    - OCENA (15%)
        - 10% zgromadzenie odpowiednich danych kontrastujących z trending
        - 5% przygotowanie danych non-trending, odfiltrowanie niektórych kategorii
- **Etap 6 - Klasyfikator, reguły, profil charakterystyczny i wiedza dla youtubera**
    - Wybranie miar oceny klasyfikatora (nie tylko trafność predykcji / uwzględnić niezbalansowane dane!)

    - Opracowanie klasyfikatora dla wersji trending / non-trending (interpretowalnego! - można zastosować różne podejścia - np. specjalne wizualizacje dla black boxes)
    - Stworzenie profilu charakterystycznych wartości atrybutów dla klasy trending / otwarty problem jak to zrobić i jakie podejścia do oceny wybrać (mogą być np. specjalne miary oceny reguł)
    - Opracowanie wiedzy dla klienta - co powinien robić, jakie sztuczki stosować, czego się wystrzegać, jeśli chce żeby jego filmy trafiły do klasy trending
    - OCENA (20%)
        - 6% pierwszy klasyfikator (z wyjaśnieniem)
        - 6% drugi klasyfikator (z wyjaśnieniem) + wybór
        - 8% opracowanie wiedzy dla klienta
- **Finał**
    - Ostatnie poprawki
    - Opis zmian wprowadzonych w porównaniu do wcześniejszych prac prezentowanych w ramach checkpointów

Przedstawiony schemat oceniania stanowi podstawę punktacji, w szczególnych przypadkach punktacja może zostać obniżona.