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 dogodnej reprezentacji, oceny jakości danych, oceny ważności atrybutów, poszukiwanie współzależności między nimi, odkrycia użytecznych i potencjalnie interesujących regularności, ew. skonstruowania modelu klasyfikacyjnego. Należy także dokonać interpretacji i oceny znalezionych regularności. Z metodologicznego punktu widzenia sugeruje się wykorzystywanie poznanych metod analizy danych zarówno statystycznych jak i wywodzących się ze sztucznej inteligencji, w tym 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 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, 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
        - 5% atrybuty tekstowe (semantyka)
        - 5% atrybuty tekstowe (meta)
- **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**
    - Ocena atrybutów
    - Znalezienie korelacji między atrybutami, poszukiwania atrybutów niewnoszących informacji, usunięcie atrybutów które mogą być niepotrzebne
    - Selekcja
    - OCENA (15%)
        - 6% analiza korelacji/innych miar między atrybutami, atrybutów nieprzydatnych itp.
        - 6% selekcja atrybutów (z wyjaśnieniem)
        - 3% dyskusja odnośnie atrybutów przydatnych do zadania etykietowania vs. atrybutów przydatnych do ostatecznego zadania (dostarczenie klientowi konkretnej wiedzy)
- **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 stosują taką metodę a nie inną
    - Porównanie metod, wybór jednej z nich
    - OCENA (15%)
        - 8% pierwsza metoda (z wytłumaczeniem)
        - 4% druga metoda (z wytłumaczeniem)
        - 3% porównanie metod, wybór jednej
- **Etap 5 - YouTube API - zbieranie danych i weryfikacja wyników**
    - Pozyskanie ground truth z YouTube API (spójnego z ustaloną charakterystyką danych)
    - Weryfikacja wyników w oparciu o porównanie uzyskanych wyników z tzw. ground truth
    - 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%)
        - 5% ground truth + weryfikacja wyników z semi-supervised learning
        - 6% zgromadzenie odpowiednich danych kontrastujących z trending
        - 4% 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 / zwłaszcza binarne - mogą prowadzić do wielu miar, także gdy dane są niezbalansowanych - wtedy dobrać właściwe)
    - 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 (15%)
        - 4% pierwszy klasyfikator (z wyjaśnieniem)
        - 4% drugi klasyfikator (z wyjaśnieniem) + wybór
        - 7% opracowanie wiedzy dla klienta
- **Finał - Raport końcowy**
    - Ostatnie poprawki
    - Pełen raport łączący wszystkie etapy w całość
    - Opis zmian wprowadzonych w porównaniu do wcześniejszych prac prezentowanych w ramach checkpointów
    - Nie tylko wyniki, ale również kod, który został użyty w trakcie (np. w formie notebooka Jupyter), z pełnym opisem jak uzyskano wyniki, przetworzono dane, strojono algorytmy itp.
    - Podsumowanie
    - Przygotowane jak dla klienta - być może z krótka prezentacją - mozna też wykorzystać wykład do pokazania najciekawszych dla całości grupy?
    - OCENA (10%)
        - 5% pełen opis ostatecznego procesu odkrywania wiedzy (z uwzględnieniem etapów właściwego pozyskania i przetwarzania danych)
        - 5% użyty kod, z wyjaśnieniami

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