Multimodalna klasifikacija filmskih žanrova korišćenjem teksta i vizuelnih podataka

Definicija problema

Cilj projekta je izgradnja sistema koji automatski određuje filmski žanr na osnovu dostupnih deskriptivnih podataka i vizuelnih informacija. Ovo je višeklasna klasifikacija, jer jedan film može pripadati većem broju žanrova istovremeno (npr. akcija i avantura). Ulaz u sistem biće tekstualni opis filma i poster, dok će izlaz predstavljati predikcija jednog ili više pripadajućih žanrova.

Motivacija

Klasifikacija žanrova filmova ima značajnu praktičnu primenu u filmskoj industriji, digitalnim platformama i sistemima preporuka. Rešavanje ovog problema doprinosi poboljšanju korisničkog iskustva, kao što su personalizovane preporuke korisnicima na striming platformama. Pored toga, poboljšanje organizacije i pretraživanja velikih baza podataka o filmovima čini korišćenje ovih baza podataka efikasnijim.

Skup podataka

Za projekat će biti korišćen „The Movies Dataset“, dostupan na Kaggle.com: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Datoteke u ovom skupu podataka sadrže metapodatke za 45,000 filmova. Skup podataka se sastoji od filmova objavljenih do jula 2017. godine. Najvažniji podaci uključuju pregled filma (tekstualni opis), postere (relativna putanja do filmskog postera, koja se može rekonstruisati u URL) i lista žanrova, koja je ujedno i ciljno obeležje. Moguće je da film pripada u više od jedne klase. Ukupno postoji oko 20 različitih žanrova (akcija, drama, komedija, horor, ...).

Potrebno je dopunsko preuzimanje postera preko TMDB API-ja ili direktnim linkovanjem preko ’https://image.tmdb.org/t/p/w500/’.

Način pretprocesiranja podataka

Pošto ima i teksta (opis filmova) i slike (filmske postere), pretprocesiranje ću vršiti odvojeno za svaku vrstu unosa.

- Pretprocesiranje teksta
  Prvo ću koristiti tokenizaciju i lematizaciju, a zatim ću odbaciti stop reči. Koristiću NLP tehnike, kao što je TF-IDF, da konvertujem reči u numeričke vrednosti. Što se tiče klasifikacije, izlaz je obično multi-hot vektor koji predstavlja žanrove koji se odnose na film.

- Pretprocesiranje slike
  Pošto su slike različitih veličina i rezolucija, promena veličine i normalizacija su deo pretprocesiranja. Pored toga, koristiću augmentacije slika poput rotacije i podešavanja osvetljenosti koje povećavaju raznolikost skupova podataka i pomažu u izbegavanju overfitting-a.

Metodologija

Proces rešenja problema biće podeljen u nekoliko faza:

1. Učitavanje i priprema podataka
   - Preuzimanje i obrada tekstualnih i vizuelnih atributa
   - Tekst i slike biće obrađeni odvojeno kako bi se dobile odgovarajuće reprezentacije
   - Sinhronizacija po ID-u filma
   - Skup podataka će biti podeljen na trening (70%), validacioni (15%) i test skup (15%)
2. Kreiranje tekstualnog modela
   - Iz pretprocesiranih tekstualnih podataka kreiraće se TF-IDF reprezentacije
   - Na njima će se trenirati Naive Bayes klasifikator za predikciju žanrova
3. Modelovanje slika
   - Iz pretprocesiranih postera izdvojiće se vizuelne karakteristike u poput kolor histograma
   - Dobijeni vektori karakteristika će zatim biti klasifikovani korišćenjem algoritma KNN (K-Nearest Neighbours), koji dodeljuje žanrovsku oznaku filmu
4. Fuzija rezultata
   - Rezultati oba modela biće kombinovani u pristupu kasne fuzije
   - Predviđanja iz naivnog Bajesovog tekstualnog modela i KNN modela slike će biti agregirana da bi se dobio konačni skup predviđenih žanrova za svaki film
   - Izlaz će biti vektor binarnih vrednosti dužine ~20 (koliko ima žanrova u skupu podataka)

Način evaluacije

Podela podataka: train (70%) / validation (15%) / test (15%).

Za evaluaciju ću koristiti sledeće metrike:

- **F1-score** za merenje balansa između preciznosti i odziva po žanrovima, nezavisno od njihove učestalosti
- **Haming loss** za procenu procenta pogrešno klasifikovanih žanrova po instanci
- **Accuracy per label** za praćenje tačnosti klasifikacije za svaki pojedinačni žanr

Posebno ću analizirati tačnost predikcija tekstualnog modela, vizuelnog modela i objedinjeni rezultati nakon kasne fuzije.

Tehnologije

- Koristiću programski jezik Python
- Biblioteke: scikit-learn, NLTK, OpenCV, numpy, matplotlib, pandas

Relevantna literatura

- Skup podataka i dokumentacija:
  - https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
  - https://developer.themoviedb.org/docs/getting-started
- Radovi:
  - A multimodal approach for multi-label movie genre classification – Rafael B. Mangolin, Rodolfo M. Pereira, Alceu S. Britto Jr., Carlos N. Silla Jr., Valeria D. Feltrim, Diego Bertolini, Yandre M. G. Costa https://www.din.uem.br/yandre/MTAP_2020.pdf
  - Multi Label Text Classification with Scikit-Learn – Susan Li (Medium posts) https://medium.com/data-science/multi-label-text-classification-with-scikit-learn-30714b7819c5
  - From Theory to Practice: Implementing Multi-class Classification with KNN SKLearn – Mihirsaurkar (Medium posts) https://medium.com/@mihirsaurkar/from-theory-to-practice-implementing-multi-class-classification-with-knn-sklearn-2d85f9adc5f7