Inteligentni sistem preporuke filmova kombinovanjem analize sentimenta i IMDB ocena

1. Naziv teme

Hibridni sistem preporuke filmova korišćenjem analize sentimenta korisničkih recenzija i IMDB metadata

2. Definicija problema

Problem koji se rešava je kreiranje naprednog sistema preporuke filmova koji kombinuje kvantitativne podatke (IMDB ocene, žanrove, godine izdanja) sa kvalitativnom analizom korisničkih recenzija kroz analizu sentimenta. Cilj je da se generiše personalizovana lista preporuka filmova za korisnika na osnovu njegovih prethodnih ocena i preferencija, uzimajući u obzir kako numeričke ocene tako i emotivni ton recenzija drugih korisnika sa sličnim ukusom.

Konkretno, sistem treba da: - Analizira sentiment korisničkih recenzija za filmove - Kombinuje sentiment analizu sa IMDB ocenama i metadata - Generiše personalizovane preporuke na osnovu korisničkih preferencija - Rangira filmove prema verovatnoći da će se korisniku svideti

3. Motivacija za problem

Postojeći sistemi preporuke često se oslanjaju samo na numeričke ocene ili jednostavne algoritme kolaborativnog filtriranja, zanemarujući bogato semantičko sadržaj korisničkih recenzija. Ovo može dovesti do preporuka koje su numerički tačne, ali ne odražavaju stvarne razloge zašto se film sviđa ili ne sviđa korisnicima.

Praktična primena ovakvih sistema je značajna za:

- **Streaming platforme** (Netflix, Amazon Prime, HBO Max) - poboljšanje korisničkog iskustva i zadržavanje korisnika
- **Kinematografska industrija** - bolje razumevanje publike i ciljano marketiranje
- **Film kritičare i blogere** - automatizovano generiranje preporuka na osnovu analize velikog broja recenzija
- **Edukacione platforme** - preporučivanje filmova za akademske i obrazovne potrebe

Sistem bi mogao da smanji vreme koje korisnici troše na pronalaženje filmova koji im odgovaraju, što direktno utiče na satisfakciju korisnika i poslovne rezultate platformi.

4. Skup podataka

Dataset: IMDB Movie Dataset sa Kaggle

- **Link**: https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset
- **Broj instanci**: 5000+ filmova sa detaljnim informacijama
- **Glavni atributi**:
  - Naziv filma, godina, žanr, IMDB ocena, broj glasova
  - Opis filma, režiser, glavni glumci
  - Meta-informacije (trajanje, sertifikat)

Ciljno obeležje: - Za klasifikaciju sentimenta: binarna klasifikacija (pozitivna/negativna recenzija) - Za sistem preporuke: kontinuirani skor verovatnoće preporuke (0-1) - Raspodela: približno balansirana između pozitivnih i negativnih recenzija

5. Način pretprocesiranja podataka

1. **Čišćenje tekstualnih podataka**:
   - Uklanjanje HTML tagova, specijalnih karaktera i interpunkcije
   - Konvertovanje u mala slova
   - Tokenizacija i uklanjanje stop reči
2. **Numeričke karakteristike**:
   - Normalizacija IMDB ocena na skalu 0-1
   - One-hot encoding za žanrove filmova
   - Binarna kodiranja za kategoričke atribute
3. **Podela podataka**: 70% train, 15% validation, 15% test

6. Metodologija

Hibridni pristup koji se sastoji od tri glavne komponente:

6.1 Sentiment Analysis modul

- **Tehnika**: Fine-tuning pretreniranog BERT modela
- **Ulaz**: Tekstualne recenzije filmova
- **Izlaz**: Sentiment skor (pozitivno/negativno + confidence)

6.2 Content-based Recommendation

- **Tehnika**: Cosine similarity na TF-IDF vektorima opisa filmova
- **Ulaz**: Filmski opisi i metadata
- **Izlaz**: Sličnost između filmova

6.3 Hibridni Recommendation sistem

- **Kombinovanje rezultata** sentiment analize i content-based filtriranja

- **Tehnika**: Weighted ensemble ili Neural Network za kombinovanje
- **Finalni izlaz**: Rangirani lista preporučenih filmova

6.4 Collaborative Filtering modul

- **Tehnika**: Matrix Factorization (SVD/NMF) + User-Item similarity
- **Ulaz**: User-movie rating matrica (sintetička ili realna)
- **Izlaz**: Predicted ratings za user-movie parove

Dijagram procesa:

```
Korisnik → [Profil korisnika] → [Content filtering]
                                                    → [Hibridni model] → Preporuke
Recenzije → [Sentiment analiza] → [Sentiment skorovi]
```

7. Način evaluacije

7.1 Sentiment Analysis

- **Metrike**: Accuracy, Precision, Recall, F1-score
- **Validacija**: Stratified K-fold cross-validation (k=5)

7.2 Recommendation sistem

- **Metrike**:
  - Precision@K i Recall@K (K = 5, 10)
  - Mean Average Precision (MAP)
  - Root Mean Square Error (RMSE) za numeričke predikcije
- **Evaluacija**: Leave-one-out cross-validation
- **Baseline**: Content-based filtering bez sentiment analize

7.3 End-to-end sistem

- **A/B testiranje simulacija**: Poređenje sa jednostavnijim pristupima
- **Qualitative evaluacija**: Manuelna provera kvaliteta top-10 preporuka za sample korisnike

8. Tehnologije

- **Programski jezik**: Python 3.9+
- **Machine Learning**: scikit-learn, TensorFlow/Keras
- **NLP**: NLTK, spaCy, Transformers (Hugging Face)
- **Data manipulation**: pandas, NumPy
- **Vizualizacija**: matplotlib, seaborn, plotly

9. Relevantna literatura

1. **Burke, R. (2002)**. "Hybrid recommender systems: Survey and experiments". *User modeling and user-adapted interaction*, 12(4), 331-370.

2. **Pang, B., et al. (2002)**. "Thumbs up? sentiment classification using machine learning techniques". *Proceedings of EMNLP*.

Očekivani rezultati: Hibridni sistem koji postiže accuracy $> 85\%$ na sentiment analizi i Precision@10 > 0.7 za film preporuke, nadmašujući baseline pristup za najmanje 10%.