

PREDIKCIJA ŽANRA FILMA NA OSNOVU OPISA

IDEJA

Cilj je stvoriti tačan i precizan algoritam koji može automatski dodeliti žanr filmu na osnovu njegovog opisa

Korisnici ne moraju unapred znati tačan žanr filma koji traže



PODACI

Podaci se sastoje od 54200 instanci,
gde su atributi koji su korišćeni

- Naslov (Title)
- Opis filma (Description)
- Žanr filma (Genre)

Podaci se nalaze u data-set.csv



ALGORITMI

- K-NEAREST NEIGHBOURS (KNN)
- NAIVE BAYES
- RECURRENT NEURAL NETWORK
(MANY - ONE ARCHITECTURE)



KNN

Obrada podataka

```
data_set = pd.read_csv('data-set.csv')
data_set['Title_Description'] = data_set['Title'] + " " + data_set['Description']
grouped_data_set = data_set.groupby('Id').agg({'Title_Description': ' '.join, 'Genre': 'first'}).reset_index()
grouped_data_set.drop(['Id'], axis=1, inplace=True)
X_train, X_test, y_train, y_test = train_test_split(grouped_data_set[['Title_Description']], grouped_data_set['Genre'], test_size=0.2, random_state=42)
```

Formiranje n-grama i njihovih frekvencija

- Prolazak jednom kroz tekst
- Eksponencijalna slozenost formiranja svih n-grama
- Ne uzima kontekst ni znacenje teksta
- Lako za implementaciju

FORMIRANJE TABELE

Na osnovu k najfrekventnijih n-grama, njihovih frekvencija i zanrova, formira se tabela

Iz te tabele se izračunavaju rastojanja na osnovu Euklidske distance, te najmanja vrednost te distance postaje odabrani žanr

Rezultati dobijeni za $n=3$ i $k=2000$, dobiju se F mere:

- Macro F measure: 50.07%
- Micro F measure: 54.22%

NAIVE BAYES

Obrada podataka

- Podela teksta na tokene sa rečima (bez tz. belih reči, kao i reči koje nemaju neko određeno značenje poput the, a i sličnih)
- Bag of words reprezentacija reči (čuvaju se određene reči kao i njihove frekfencije)
- Podela skupa na trening i test podatke

Formula: $P(\text{genre}|\text{text}) \sim P(\text{genre}) * P(\text{text}|\text{genre})$

- $P(\text{genre})$ - trivijalno izračunavanje
- $P(\text{text}|\text{genre})$ - moramo uzeti da su reči nezavisne, pa to pretvoriti u proizvod nezavisnih reči
- Može se desiti da neka vrednost $P(\text{word}|\text{genre})$ bude 0, te će onda ceo izraz biti 0, te smo za to uveli alternativu Laplace Add Alpha smoothing - gde se parametar Alpha podešava
- Kako proizvod može biti suviše mali, bolje je ceo izraz logaritmovati i proizvod pretvoriti u sumu

Rezultati dobijeni za ovaj data set i za parametar alpha=1,
tj Laplace Add One Smoothing su:

- Macro F measure: 74.40%
- Micro F measure: 82.08%

RNN

Obrada podataka

```
data_set = pd.read_csv('data-set.csv')
data_set['Title_Description'] = data_set['Title'] + " " + data_set['Description']
grouped_data_set = data_set.groupby('Id').agg({'Title_Description': ' '.join, 'Genre': 'first', 'Text_cleaning': 'first'}).reset_index()
grouped_data_set.drop(['Id'], axis=1, inplace=True)

genres = sorted(grouped_data_set['Genre'].unique())

word_processing=WordProcessing(genres)

grouped_data_set['Title_Description'] = grouped_data_set['Title_Description'].apply(lambda n: word_processing.unicode_to_ascii(n))

X_train, y_train, X_valid, y_valid, X_test, y_test = word_processing.train_valid_test_split(grouped_data_set, 0.15, 0.2)
```

- Podela skupa na trening, test i validacioni skup podataka

- Korišćenje One-hot encoding vektora, gde će veličina vektora biti veličina vokabulara
- Za kreiranje rekuretnе neuronske mreže potrebne su 3 matrice, ulazna matrica U, skrivena matrica W i izlazna matrica V
- Aktivaciona funkcija je korišćena LogSoftmax
- Broj epoha - 5
- Optimizer - ADAM sa learning rate 0.05
- Loss funkcija - Negative Log Likelihood Loss

Rezultati dobijeni za ovaj data set i za ove parametre su:

- Macro F measure: 64.72%
- Micro F measure: 72.18%

MATRICE KONFUZIJE

- Za svaki žanr će se napraviti posebna matrica, koja će sadržati True Positive, False Positive, False Negative i True Negative vrednosti
- Macro F measure - za svaku matricu ćemo izračunati njihov F measure, te će aritmetička sredina biti rezultat
- Micro F measure - napraviće se nova matrica, gde će vrednosti predstavljati zbir po njihovim vrednostima

MATRICE KONFUZIJE

- F measure za jednu matricu konfuzije se računa kao harmonijska sredina preciznosti i odziva, gde imamo i dodatni parametar beta, koji nam daje odnos važnosti između te dve vrednosti
- Preciznost se izračunava $P=TP/(TP+FP)$
- Odziv se izračunava kao $R=TP/(TP+FN)$

KRAJ