Zeneszövegek érzelemelemzése alapján való hangulat meghatározó

Készítette: Hajdú Patrik Zsolt, RP329D - Dátum: 2025.11.07.

Kivonat:

Ez a projekt egy NLPn alapuló rendszert mutat be, amely zeneszövegek alapján érzelmi profilt készít az adott zenéről majd zenei ajánlást generál hasonló zenékről.

A rendszer két fő komponensből áll:
1 Egy érzelem-klasszifikációs modellből, amely be lett tanítva egy érezlemekkel címkézett Twitter-üzeneteket tartalmazó adathalmazból.
2 Egy ajánló motorból, amely a betanított alkalmazza egy közel 67+ ezer dalszöveget tartalmazó Spotify adatbázison. A rendszer a felhasználó által megadott dal érzelmi kategóriája és dalszövege alapján javasol új, hangulatban és témában hasonló zenéket.

Kulcsszavak:

NLP, dalszöveg, érzelem, Kaggle, scikit-learn, TF-IDF, koszinusz-hasonlóság

Tartalomjegyzék

1. A feladat rövid ismertetése

A házi feladat célja egy olyan NLP-alapú rendszer megvalósítása volt, amely képes zeneszövegeket elemezni és érzelmi alapon kategorizálni.

A feladat két fő részből állt:

- **Érzelem-elemzés:** Egy adott (előadó, dalcím) páros alapján a rendszer megkeresi a dal szövegét, és egy betanított gépi tanulási modell segítségével meghatározza a hat alapérzelem (öröm, bánat, szeretet, harag, félelem, meglepetés) közül melyik jellemzi a megadott zenét.
- **Zenei Ajánlás:** Az első lépésben meghatározott érzelmi profil és a dalszöveg ismeretében a rendszer az adatbázisból olyan dalokat ajánl, amelyek mind érzelmileg, mind szövegileg hasonlítanak a bemeneti dalhoz.

2. Adatok (törölje, ha nem releváns)

A rendszer két, a Kaggle platformról származó adathalmazra épül:

1. Érzelem Adathalmaz (Emotions Dataset)
    - Forrás: https://www.kaggle.com/datasets/nelgiriyewithana/emotions/data
    - Tartalma: kb. 417+ ezer címkézett angol nyelvű Twitter-üzenet. A text oszlop tartalmazza az üzenetet, a label oszlop pedig a hozzá tartozó érzelmi kategóriát az alábbiak szerint: 0 (sadness), 1 (joy), 2 (love), 3 (anger), 4 (fear), 5 (surprise).
2. Dalszöveg Adathalmaz (Spotify Million Song Dataset)
    - Forrás: https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset
    - Tartalma: kb. 57+ ezer dal adatait tartalmazza, beleértve az artist (előadó), song (dal címe) és text (dalszöveg) oszlopokat. Az adathalmaz erőssége, hogy a dalszövegek több soros, formázott szövegek.

3. A megoldás ismertetése (ezt a címet átírhatja, ha szeretné)

A rendszer egy többlépcsős NLP folyamatot (pipeline) valósít meg Python nyelven, a scikit-learn, pandas és nltk könyvtárak felhasználásával.

[Adatok Betöltése] ---> [Modell Tanítása (Emotions)] --->
[Dalszövegek Elemzése (Spotify)] ---> [Hasonlósági Mátrix Építése] ---> [Interaktív Ajánlás]

Érzelem-klasszifikációs Modell Tanítása

Mivel a dalszöveg-adatbázis nem tartalmazott érzelmi címkéket, szükség volt egy külső modell tanítására az "Emotions" adathalmazon.

Szöveg-előfeldolgozás: A modell tanítása előtt az "Emotions" adathalmaz szövegeit normalizáltam.
Ez a lépés kulcsfontosságú a modell pontosságának növelése érdekében.

A lépései:
- Kisbetűsítés (.lower()).
- Tokenizálás (NLTK word_tokenize).
- Angol stop-szavak (stopwords.words('english')) és írásjelek (string.punctuation) eltávolítása.
- A tiszta tokenek visszaillesztése egy stringgé.

Vektorizálás:
Egy scikit-learn Pipeline-t használtam. A preprocess_text függvény által tisztított szövegeket TF-IDF vektorokká alakítja. Ez a módszer súlyozza a szavakat az alapján, hogy milyen gyakoriak egy dokumentumban, de milyen ritkák a teljes adathalmazban, így kiemelve a jelentéssel bíró kulcsszavakat.

Dalszövegek Elemzése

A 1. lépésben kapott emotion_pipeline modellt alkalmaztam a teljes (mintavételezett) Spotify adathalmaz text oszlopára. A predikció futtatása után minden dal kap egy numerikus érzelmi címkét (0-5), amit aztán a EMOTION_MAP szótár segítségével visszaalakítottam szöveges címkévé (egy új emotion_name oszlopban).

Ajánlórendszer

Az ajánlás három pilléren nyugszik:
- *Felhasználói Bevitel Kezelése*: A main_cli függvény fogadja a bevitelt, és eltávolítja a felesleges szóközöket az előadó és a dalcím elől/mögül.
- *Érzelmi Szűrés*: Megkeresi a bemeneti dalt (kisbetű-érzéketlenül), és azonosítja annak kiszámított érzelmi kategóriáját (pl. joy). Az ajánlások kizárólag az ebbe a kategóriába tartozó dalok közül kerülhetnek ki.
- *Szöveges Hasonlóság*: A csak érzelmi alapon történő szűrés nem elegendő. A rendszer a dalszövegek tartalmi hasonlóságát is méri. Ehhez a teljes Spotify dalszöveg-adatbázison felépítettem új, globális TF-IDF mátrixot. Amikor a felhasználó dalt választ, a rendszer kikeresi annak TF-IDF vektorát, és Kosinusz-hasonlóság segítségével megméri a távolságát az összes többi dal vektorától.

Végül pedig a végső ajánlási lista ezen hasonlósági pontszámok szerint van csökkenő sorrendbe rendezve.

4. Futási példák

A program terminálból való indítás következtében az alábbi konzol kimenetet jeleníti meg. Mint az feltűnhet mindössze 12+ ezer zene van a minta vételezés után, mivel ennek mértéke állítható a programban. Ez a program indításának gyorsítása érdekében lehetséges.

```
Loading datasets...
Loaded: 416809 emotion samples.
Loaded: 11530 lyrics (after sampling).

Training emotion classification model...

Model training complete. Duration: 138.88 sec.

--- Model Evaluation (For the Report) ---
              precision    recall  f1-score   support

     sadness       0.93      0.94      0.94     24238
         joy       0.91      0.93      0.92     28214
        love       0.80      0.75      0.78      6911
       anger       0.90      0.89      0.90     11463
        fear       0.85      0.84      0.84      9542
    surprise       0.77      0.70      0.73      2994

    accuracy                           0.90     83362
   macro avg       0.86      0.84      0.85     83362
weighted avg       0.90      0.90      0.90     83362

----------------------------------------------------

Analyzing lyrics database with the emotion model...
This might take a few minutes depending on the sample size...
Lyrics analysis complete. Duration: 15.78 sec.

Emotion distribution in the lyrics database (Top 5):
emotion_name
joy        5920
sadness    3705
love        707
anger       672
fear        459
Name: count, dtype: int64

Building similarity matrix (TF-IDF) from lyrics...
Similarity matrix ready. Shape: (11530, 35533)
```

Példa 1: Sikeres ajánlás

```
Enter artist name: abba
Enter song title: cassandra

--- Analysis: abba - cassandra ---
Determined emotion: sadness

Recommendations (based on similar emotion and text):
  1. Conway Twitty - Don't Tell Me You're Sorry (Similarity: 0.28)
  2. The Temptations - Sorry Is A Sorry Word (Similarity: 0.26)
  3. Gordon Lightfoot - Remember Me (Similarity: 0.18)
  4. Hanson - Being Me (Similarity: 0.17)
  5. Religious Music - Angels Among Us (Similarity: 0.15)
```

Példa 2: Dal nem található

```
Enter artist name: Travis Scott
Enter song title: goosebumps

Sorry, the song 'Travis Scott - goosebumps' was not found in the database.
Tip: Try to enter the name exactly as it appears in the Kaggle dataset.
```

5. A megoldás értékelése

Amit sikerült megvalósítani:

- A rendszer sikeresen betanít egy 90%-os pontosságú szöveg-klasszifikációs modellt.
- Képes a felhasználói bevitel (kis/nagybetűk, szóközök) robusztus kezelésére.
- Megvalósítja a feladatkiírásban szereplő kettős célt: először érzelmet azonosít, majd az azonos érzelmű dalokon belül szöveges hasonlóság alapján ajánl.

```
--- Model Evaluation (For the Report) ---
              precision    recall  f1-score   support

     sadness       0.93      0.94      0.94     24238
         joy       0.91      0.93      0.92     28214
        love       0.80      0.75      0.78      6911
       anger       0.90      0.89      0.90     11463
        fear       0.85      0.84      0.84      9542
    surprise       0.77      0.70      0.73      2994

    accuracy                           0.90     83362
   macro avg       0.86      0.84      0.85     83362
weighted avg       0.90      0.90      0.90     83362

----------------------------------------------------
```

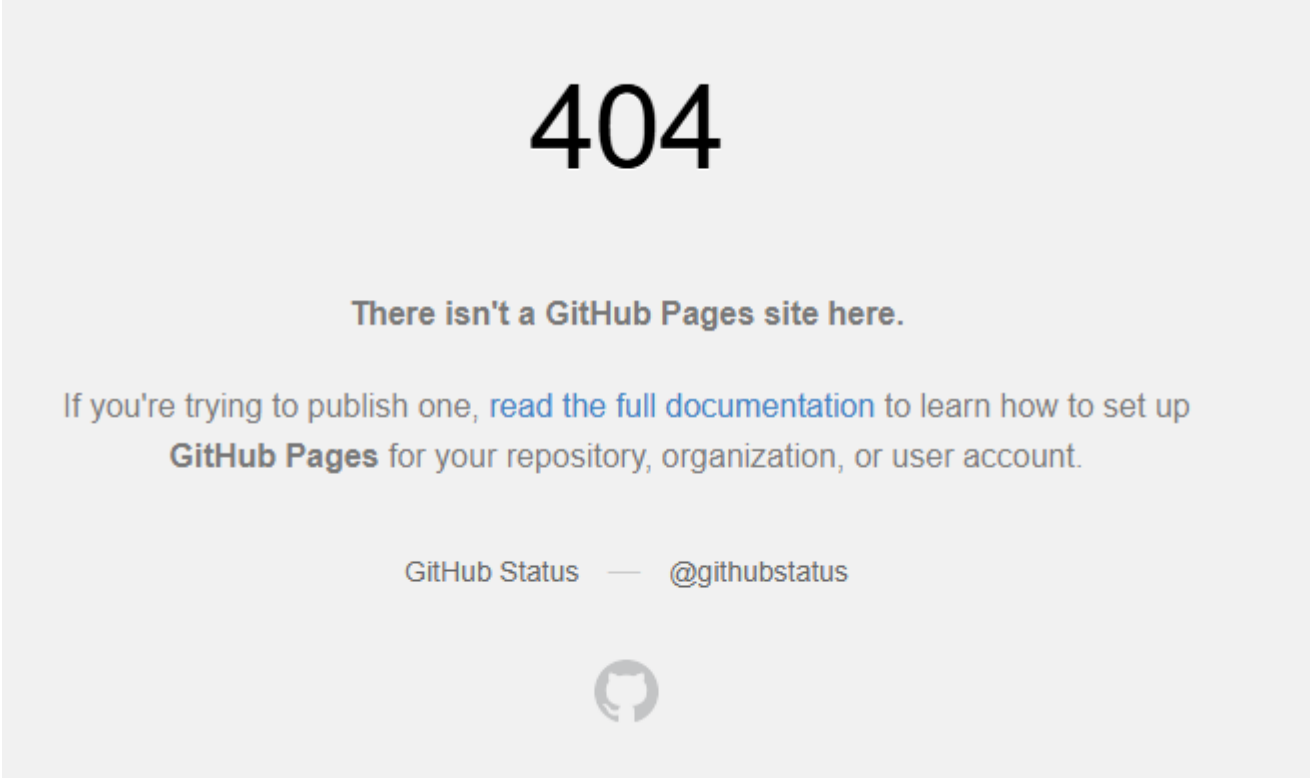
6. Hivatkozások

Adathalmazok:

- https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset
- https://www.kaggle.com/datasets/nelgiriyewithana/emotions/data

Könyvtárak:

- Pandas (adatkezelés). https://pandas.pydata.org/docs/
- NLTK (szöveg-előfeldolgozás). https://www.nltk.org/
- Scikit-learn (gépi tanulás, TF-IDF, koszinusz-hasonlóság). Mivel a hivatalos weboldaluk nem elérhető https://scikit-learn.org/, ezért ChatGPT és Gemini-t használtam minden a könyvtár függvényeinek használatáról való kérdésem megválaszolásához.



Szakirodalom:

- Dr. Mészáros Tamás: "NLP_e03_stat_modellek.pdf"
- Dr. Mészáros Tamás: "NLP_e05_tanult_modellek.pdf"

Mesterséges Intelligencia:

- A programot a Visual Studio Code-ban fordítottam, futtattam és írtam, amelyben köztudottan már megtalálható az AI alapú GitHub Copilot kód kiegészítés.
- A program készítés során olykor használtam az automatikus kód kiegészítés funkciót.