

Magyar MI Diákolimpia

Epikus Emojik 🤖 🎉 💬 🔥

Feladatleírás

2025. május 24.

1. Epikus Emojik

Emoji Emőke egy önjelölt filmesztéta, aki megszállottan rajong a mozi világáért és az emojikért. Arra tette fel az életét, hogy minden létező filmcímet egyetlen jól megválasztott **emoji-szekvenciával** (4-5 emoji) foglaljon össze.

Élet filozófiája:

„Ha egy film nem írható le három emojival, nem is érdemli meg, hogy megnézzem!”.

Emőkének sok időbe telik kitalálni az emojikat minden film címhez, így nem jut idő arra amit igazán szeret csinálni: **AI diákolimpiára készülni.**

A ti feladatotok: Segítsetek Emőkének automatizálni a filmek emojik nyelvére való fordítását. Mind ezt egy előre tanított modell és egy adatbázis segítségével, amit megkaptok. A különböző részfeladatok bevezetnek Titeket a nyelvi modellek tanítási ciklusába, és a végén egy modell áll majd Emőke rendelkezésére, ami nagyban segít majd neki a munkájában. A részfeladatokat egy Jupyter Notebookban oldhatjátok meg (Google Colab ajánlott, de lokális környezetben is elfogadott csak lassú a tanítás).

Példák:

- **Lion King** $\Rightarrow$ 🦁 🏰 🌅
- **The Godfather** $\Rightarrow$ 🧔 🗡️ 💰 🧑
- **The Lord of the Rings: The Two Towers** $\Rightarrow$ 🧝 📿 ⚔️ 🌲 🏰

2. A feladat pontozása

- | | |
|--|-----------|
| 1. Adat felfedezése | [10 pont] |
| 2. Adat tisztítása | [5 pont] |
| 3. Modell betöltése | [5 pont] |
| 4. Adat előfeldolgozása | [20 pont] |
| 5. Baseline értékelés, nem finomhangolt modell | [10 pont] |
| 6. Modell tanítása | [25 pont] |

- | | |
|---|-----------|
| 7. Kiértékelés a tanító adatokon | [3 pont] |
| 8. Vizualizáció és teljesítmény összehasonlítás | [7 pont] |
| 9. Figyelem Mechanizmus Vizualizációja | [15 pont] |

Maximális pontszám: 100 pont

3. Technikai tudnivalók

A megoldáshoz a Google Colab felülete javasolt, a videokártya használata miatt, ahol gyorsabb a tanítás körülbelül 60-szor, mint CPU-n. Lokálisan is megoldható a feladat, de a max pontszámot elérni processzoron tanítással szinte lehetetlen a feladat megoldási ideje alatt.

A feladathoz egy Huggingface platformon megtalálható modellt kell használni, ami előre volt tanítva szöveg-emoji adathalmazon, de nem filmcímeken. Ez fontos, mivel így egy kis adathalmazon tudjuk finomhangolni a modellt a mi saját céljainkra. Minden szükséges információ és a szükséges könyvtárak megtalálhatóak a notebook elején.

Az adathalmaz az IMDB 1000 legjobbra értékelt filmjét tartalmazza, és elérhető a következő linken:

 Film cím - Emoji Adathalmaz

A feladat megoldása után, függetlenül attól, hogy Google Colabban vagy lokálisan dolgoztál, a `.ipynb` fájlt kell feltöltened a CMS rendszerbe a megfelelő feladathoz. Más fájl feltöltésére nincs szükség.

4. Hasznos források

 HuggingFace Transformers - Dokumentáció

 HuggingFace Datasets - Dokumentáció

 T5 modell - alapelvek

 BLEU score - értelmezés és példa

 Python Regex - Emojik szűrése