NAGY NYELVI MODELLEK HASZNÁLATA (CHATGPT...)

3. forduló

A kategória támogatója: Cambridge Mobile Telematics

Ismertető a feladatlaphoz

Kérjük, hogy a feladatlap indítása előtt mindenképp olvasd el az alábbi útmutatót:

> Amennyiben olyan kategóriában játszol, ahol van csatolmány, de hibába ütközöl a letöltésnél, ott valószínűleg a vírusirtó korlátoz, annak ideiglenes kikapcsolása megoldhatja a problémát. (Körülbelül minden 3000. letöltésnél fordul ez elő.)

> Helyezéseket a 4. forduló után mutatunk, százalékos formában: adott kategóriában a TOP 20-40-60%-hoz tartozol.
>
> A feltűnően rövid idő alatt megoldott feladatlapok kizárást vonnak maguk után, bármilyen más gyanús esetben fenntartjuk a jogot a forduló érvénytelenítésére!

Üdvözöllek a Nagy Nyelvi Modellek kategória 3. fordulóján!

A következő feladatsorban megismerkedünk az OpenAI platformjával, és egy HuggingFace eszközzel. Regisztrálni, bejelentkezni egyik helyre sem kell, viszont a HuggingFace példát, érdemes lehet előre betölteni, hogy ezzel se teljen az idő a verseny közben.

Jó versenyzést!

1. feladat 2 pont

Az OpenAI kutatói 2018-ban tették közzé az első Generative Pretrained Transformer (GPT) modelljüket, mely alapfelépítésén azóta sem sokat változtattak, és amin a ChatGPT-t hajtó modell is alapul. A jelenleg legnépszerűbb megközelítés szöveg generálására az, hogy a mondatot szóról-szóra építik fel a nyelvi modellek, mindig a következő token* valószínűségét becslik. Minden egyes token generálásakor egy valószínűségi eloszlást kapunk a modell által látott tokenek felett. Ezen tokenek közül választható stratégia, ami meghatározza a modell hogyan válasszon. Milyen beállítási lehetőségeink vannak az OpenAI modellel?

---

token - szavak helyett pontosabb, ha *token*-ekről beszélünk. A token-ek a szövegben található gyakori karaktersorozatok, vagyis a szó-részletek. Gondoljunk csak bele, mennyivel hatékonyabb szó-részeket, szótöveket és ragokat tárolni ahelyett, hogy minden szó minden ragozott formáját külön memorizálnák és számolnánk erre a valószínűségeket. A folyamatot, ami a folyó szövegből token-eket generál tokenizálásnak nevezzük.

Válaszok

- [ ] Mindig a legvalószínűbb token választása
- [ ] A legvalószínűbb N token közül véletlenszerűen (N értéke megadható)
- [ ] Azon tokenek közül választ, amik valószínűségének összege elér egy P határértéket (P értéke megadható)
- [ ] A valószínűségek mentén D mélységű keresőfát épít, majd a szimulált valószínűségek közül választ egyet (D értéke megadható)
- [ ] A valószínűségi eloszlást egy T hőmérsékleti változóval állítjuk, majd a kapott eloszlásból választunk (T értéke megadható)

2. feladat 1 pont

A GPT-3 megjelenése egy új paradigmát is behozott a nyelvi modellek világába (általában innentől kezdve beszélünk nagy nyelvi modellekről): few-shot tanulás. Ez mit takar?

Válasz

- ( ) A nyelvi modell rövid leírás és 3 vagy több példa alapján, súlyainak átállítása nélkül, képes számos feladat megoldására
- ( ) A nyelvi modell rövid leírás és 3 vagy több példa alapján, súlyainak átállítása után, képes számos feladat megoldására

A nyelvi modellt 3 vagy több példával elég finomhangolni (fine-tune) a specializált feladat elvégzéséhez

3. feladat 0 pont

A ChatGPT használható webes interfészen és API-n keresztül is. Az OpenAI a generált token-ek mentén számláz (jelenleg $0.0015 a költsége 1000 token generálásának). A token-ek a szövegben található gyakori karaktersorozatok, vagyis a szó-részletek. A folyamatot, ami a folyó szövegből token-eket generál tokenizálásnak nevezzük, amely megközelítésenként eltérhet, a lényegük azonban ugyanaz: olyan token-ek előállítása, melyek a szavak építőelemei. Az OpenAI modellek (amilyen a ChatGPT is) által használt tokenizáló online is elérhető. A következő szöveggel próbáljuk ki a tokenizálást az online eszközzel, mi lenne a költsége ezt legenerálni a ChatGPT API-val?

```
The biggest lesson that can be read from 70 years of AI research is that general m
```

A választ számként kell beírni dollár ($) jel nélkül.

Válaszok

4. feladat 2 pont

Saját tokenizálót is csinálhatunk, vagy használhatjuk a HuggingFace valamely előre tanított tokenizálóját. A HuggingFace leginkább a természetes nyelvi feldolgozó alkalmazásokhoz épített gépi tanulási modellek és adathalmazok megosztását lehetővé tevő platformjáról nevezetes. Bár a kutatói közösség az elsődleges felhasználói a felületnek, újabban lehetőség van ún. demo-k megosztására ahol ki lehet próbálni például a nyelvi modelleket és tokenizálókat is.

Tekintsük az alábbi tokenizálót, a következő szöveget hány token-né bontja szét?

```
BERT (Bidirectional Encoder Representations from Transformers) language model.
```

Tipp: minden HuggingFace példánál megnézheted annak forrását is. Figyelj rá, hogy tokenizálásnál a mondat elejét / végét egy-egy speciális tokennel szokás jelölni (ez a forrásban is látszik).

Válasz

5. feladat 1 pont

Most kérjük meg a ChatGPT-t, hogy exportáljon ki adatokat számunkra Excel táblázatba. Első lépésben generáljunk valamilyen adatot, például a következő promt használatával:

Generate a list that consists of 10 different fruits native in Hungary.

Ezt követően kérd meg az eszközt, hogy ezt exportálja ki Excelbe, például így:

Export the list above into a spreadsheet.

Mit tapasztalsz?

Válasz

- A válasz tartalmaz egy linket, amivel kiexportálhatom a listát.
- A modell VBA kódot generált, amit beillesztve Excelbe megkapom a listát
- A ChatGPT nem képes adat exportálására Excel táblázatba, erről ad leírást.
- A megosztás gomb segítségével egy letölthető dokumentum jön létre.

Megoldások beküldése