**PROJECT BESZÁMOLÓ**

**Project neve**

*GPT-2 comment generator*

**Fejlesztési idő**

*20 óra*

**Adathalmaz**

* *A példa project adathalmaza 372848 és 434084**rekordot tartalmaz. (De ebből jóval kevesebbet használunk a modell generálási ideje miatt.)*
* *A rekordokhoz tartozó mezők:* 
  + *Videó azonosító*
  + *Comment*
  + *Likes*
  + *Replies (száma)*
* *Az adatok címkézettek*
* *A címkék többes osztályúak*

**ELőfeldolgozás**

* *Az adatokat soronként beolvasva, „ , ” karakterenként tagolva szétválasztjuk, majd kinyerjük a második oszlop (Comments) tartalmát. Ez még ’ ” ’ között van, amit eltávolítunk az input tisztasága érdekében.*

**Modell**

* *A modell a GPT-2 szöveggeneráló transzformer, amely az adott inputot feldolgozva generál egy relatíve releváns választ.*

**Kiértékelés**

* *A háló pontosságát nagyon nehéz (szinte lehetetlen) meghatározni, van hogy teljesen relevánsan válaszol, van hogy teljesen értelmetlen outputot ad. A modell nagyságától függően változik az output szöveg pontossága is (és hogy mennyire használjuk (--top\_k)).*

**További fejlesztési lehetősége, Tapasztalatok**

*Az inputként átadott „példa párbeszéd” kiterjesztésével valószínűleg még jobb eredményeket tudnánk elérni. Mivel egy nagyon fejlett neurális hálóról beszélünk, ezért a továbbfejlesztések lehetősége szinte végtelen, rengeteg mindenre lehet használni (pl. elő commentekre válaszolás, stb…). A tapasztalatok szerint, a háló bőven nem tökéletes, de nagyon impresszív válaszokat tudott generálni. További finom hangolással valószínűleg még ennél is jobb eredményeket lehetne elérni(--top\_k, input mérete, minősége). Illetve a hálóhoz tartozik több féle nagyságú (azaz pontosságú) modell, amikkel pontosabb eredményeket kaptunk (774M-ig próbáltuk), de a generálási idő is jelentősen megnőtt.*