PROJEKT BESZÁMOLÓ

Bevezetés a Természetes Nyelvű Szövegfeldolgozásba

Tartalom

[Projekt bemutatása 2](#_Toc68182587)

[Csapattagok 2](#_Toc68182588)

[Fejlesztési idő 2](#_Toc68182589)

[Adathalmaz 2](#_Toc68182590)

[Előfeldolgozás 2](#_Toc68182591)

[Modell 2](#_Toc68182592)

[Kiértékelés 3](#_Toc68182593)

[További fejlesztési lehetőségek, Tapasztalatok 3](#_Toc68182594)

# Projekt bemutatása

*Multiklasszifikáció olyan adathalmazon, mely hotel értékeléseket tartalmaz címkézve.*

*Adathalmaz forrása: https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews*

# Fejlesztési idő

*~15 óra*

# Adathalmaz

* *A példa projekt adathalmaza 20491 rekordot tartalmaz.*
* *A rekordokhoz tartozó mezők:* 
  + *Review: a hotel szöveges értékelései*
  + *Rating: az egyes értékelésekhez tartozó szám (1-5)*
  + *Rating\_name: a Rating szöveges megfelelője*
* *Az adatok címkézettek*
* *A címkék többes osztályúak*

# Előfeldolgozás

* *A Rating\_name oszlop felvétele szkript segítségével*
* Az adathalmazt felbontottuk (train/val/test) 60/20/20 arányban egy szkript segítségével
* A szöveg tokenizálásához a Bert model előtanított DistilBert tokenizálóját használtuk
* Az általunk használt szótár mérete 30523

# Modell

* *A modellt több neurális hálózat összekapcsolásával építettük fel:*
  + *Embedding neurális hálót használtunk, mely az encode-olt szótárat vektorokká alakítja*
  + Conv1D réteget használtunk, mert szekvenciális adatokkal dolgozunk.
  + *Flatten réteget használtunk, ami a konvolúciós réteg kimenetéből egyetlen vektort készít*
  + Három Dense réteget használtunk a modellben
* *Softmax aktivációs függvényt használtunk, mert multiklasszifikálást hajtunk végre*
* *Loss függvényünk SparseCategorcialCrossEntropy volt, mert multiklasszifikációt végeztünk*
* *A pontosságot Accuracy-val mértük*
* *A tanítást 5 cikluson keresztül végeztük, mert több ciklus esetén nem történt pozitív változás, az eredmény stagnált*
* *A batch méret 32 volt, kisebb érték használata esetén nem tapasztaltunk javulást*

# Kiértékelés

* *A tanuló adathalmazon a pontosság ~97%-volt*
* *A validációs adathalmazon a legjobb paraméter beállítás pontossága: ~55%-volt*
* *A teszt adathalmazon a pontosság: ~55%-volt*

# További fejlesztési lehetőségek, Tapasztalatok

*Azért ezt a modellt alkalmaztuk, mert a Bert modell előnye, hogy előre felépített szótárral rendelkezik. Más modell alkalmazása esetén nem tapasztaltunk javulást.*

*Multiklasszifikációt végeztünk öt különböző címke szerint. Mivel sok próbálkozás után sem tapasztaltunk pozitív változást, ezért az a sejtésünk, hogy az általunk felhasznált adathalmaz nem ideális, vagy az elképzelésünk rossz.*