PROJECT BESZÁMOLÓ

Bevezetés a Természetes Nyelvű Szövegfeldolgozásba

Tartalom

[Project bemutatása 2](#_Toc68182587)

[Csapat tagok 2](#_Toc68182588)

[Fejlesztési idő 2](#_Toc68182589)

[Adathalmaz 2](#_Toc68182590)

[ELőfeldolgozás 2](#_Toc68182591)

[Modell 2](#_Toc68182592)

[Kiértékelés 3](#_Toc68182593)

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

# Project bemutatása

*Példa project neve és rövid bemutatása.*

# Csapat tagok

* *Név 1*
* *Név 2*
* *Név 3*

# Fejlesztési idő

*82 óra*

# Adathalmaz

* *A példa project adathalmaza 567000 rekordot tartalmaz.*
* *A rekordokhoz tartozó mezők:* 
  + *Text: Maximum 5000 karakter hosszan a cikk teljes szövege*
  + *Title: A cikk címe*
  + *Label: A Cikkhez tartozó kategória*
  + *CreateTime: A cikk létrehozásának időpontja.*
* *Az adatok címkézettek*
* *A címkék többes osztályúak*

# ELőfeldolgozás

* *Az adatok címkei szövegesek, ezért azokat decimális számmá kell konvertálni*
* *A szövegből el kellett távolítani a helyesírási hibákat*
* *Stopword-öket eltávolítottuk. Azért, mert..., az a hipotézisünk, hogy...*
* *Stemming alkalmazása. Azért, mert..., az a hipotézisünk, hogy...*
* *Lematizáció alkalmazása. Azért, mert..., az a hipotézisünk, hogy....*
* *Sok html tag-et találtunk. Eltávolító eljárást készítettünk hozzá.*
* *A szöveg túl hosszúnak bizonyult a rendelkezésre álló erőforrásainkhoz ezért csökkentettük annak hosszát.*
* *A szöveget tokenizáltuk xy technikával.*
* *A szótárunkat felépítettük 15000 szó mérettel mert....*
* *Gyakorisági adatokat nyertünk a nyers szöveg mellé....*
* *Az adathalmazt felbontottuk Train/val/test, 60/20/20 arányban tanításhoz és kiértékeléshez.*

# Modell

* *A modellt több neurális hálózat összekapcsolásával építettük fel.*
  + *Konvolúciós neurális hálót használtunk mert*
  + *LSTM-et használtunk mert*
* *Az alábbi aktivációs függvényeket használtuk, mert.*
  + *Act1*
  + *Act2*
* *Loss függvényünk SparseCategorcialCrossEntropy volt mert.*
* *A pontosságot Accuracy-val mértük*
* *A tanítást 10 cikluson keresztül végeztük mert...*
* *A batch méret 100 volt mert*
* *Az alábbi paramétereket próbáltuk ki*
  + *Paraméterek halmaza 1....*
  + *Paraméterek halmaza 2....*

# Kiértékelés

* *A tanuló adathalmazon a pontosság 75%-volt*
* *A validációs adathalmazon a legjobb paraméter beállítás pontossága: 69%-volt*
* *A teszt adathalmazon a pontosság: 71%-volt*

# További fejlesztési lehetősége, Tapasztalatok

*Azt tapasztaltuk, hogy minden egyes újabb preprocessing lépés bevezetésével jelentősen javult a hatékonyság, ezért az adathalmaz vizsgálata további statisztikai eszközökkel indokolt lehet azért, hogy újabb tisztítási lehetőségeket találjunk a további javulás érdekében. A modell változtatása jelentősen nem javította a teljesítményt, ezért annak további hyperparameter optimalizálása nem indokolt. Esetleg más modell kipróbálása, mint például Naiv-Bayes osztályozó indokolt lehet amennyiben a statisztikai analízis erős valószínűségi függőséget mutat az egyes rekordok között.*