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

## Multi-label Text Classification with Keras: Azt fogjuk megnézni, hogyan kell -||- készíteni, egy féle kimenettel. Egy szövegosztályozó modellt fogunk fejleszteni, ami analizálja a szöveges kommenteket és eldönti, milyen címke érvényes rá. Ez a probléma részhalmaza a több kimenetelű modellnek. A különbség e között és egy multi-class classification probléma között:

* multi-class classification-nél egy példa vagy egy rekord egyetlen kimeneti osztályhoz tartozik a sok közül
* a szentiment analízis problémában, azt tanultuk, hogy egy szöveg lehet jó, rossz vagy átlagos, de egyszerre nem lehet több is
* a multi-label classification-nél egy példának lehet több kimenete egy időben, ebben a szövegben pl. toxic, obscene, insulting, stb..

# Fejlesztési idő

*10 óra*

# Adathalmaz

* *A példa project adathalmaza 159571 rekordot tartalmaz.*
* *A rekordokhoz tartozó mezők:* 
  + *id: azonosító*
  + *comment\_text: szöveg*
  + *toxic: 0/1 Az adott szöveg toxikus-e vagy sem.*
  + *severe\_toxic: 0/1 Az adott szöveg nagyon toxikus-e vagy sem.*
  + *obscene: 0/1 Az adott szöveg trágár-e vagy sem.*
  + *threat: 0/1 Az adott szöveg fenyegetés-e vagy sem.*
  + *insult: 0/1 Az adott szöveg sértő-e vagy sem.*
  + identity\_hate: 0/1 *Az adott szöveg személyes gyűlölet-e vagy sem.*
* *Az adatok címkézettek*
* *A címkék többes osztályúak*

# ELőfeldolgozás

* *Betöltjük az adattáblát a memóriába, szemlélteti az adathalmaz alakját és kiírja az adathalmaz címkéit*
* *Kivesszük a null értékű sorokat és kiíratunk egy random szöveget*
* *Megállapítjuk a szövegről, hogy melyik címke teljesül rá*
* *Kirajzoltatunk egy felcímkézett grafikont, ami szemlélteti melyik a leggyakrabban előforduló komment*

# Modell

* *A modellt több neurális hálózat összekapcsolásával építettük fel.*
  + *Konvolúciós neurális hálót használtunk mert*
  + *LSTM-et használtunk mert*
* *Loss függvényünk beágyazott vektorral volt megvalósítva.*
* *A pontosságot Accuracy-val mértük, melynek értéke: kb 99%*
* *A tanítást 5 cikluson keresztül végeztük, mert nagy volt a batch\_size.*
* *A batch méret 2048 volt, mert nem akartunk kifutni a memóriából, de túl kevés tesztet sem akartunk elvégezni.*
* *Az alábbi paramétereket próbáltuk ki*
  + *Paraméterek halmaza inputlayer,embedding,LSTM,dense száma:14,942,322*

# Kiértékelés

* *A tanuló adathalmazon a pontosság 98,13%-volt*
* *A validációs adathalmazon a legjobb paraméter beállítás pontossága: 98,07%-volt*
* *A teszt adathalmazon a pontosság: 98,10%-volt*

# További fejlesztési lehetősége, Tapasztalatok

*A* Multi-label Text Classification az egyik leggyakoribb szövegosztályozó probléma. Egy fajta mélytanuló megközelítést vizsgáltunk meg. Single dense output réteget használtunk több neuronnal, ahol minden neuron egy címkét reprezentál. Azt tanácsolnám, hogy az aktivációs függvényeket változtassuk meg és osszuk fel a tanítást, hogy leteszteljük kaphatunk-e jobb értékeket.