**PROJECT BESZÁMOLÓ**

**Project neve**

Steam értékelések szentiment analízise wordlevel tokenizációval

**Fejlesztési idő**

9-10 óra

**Adathalmaz**

* A példa project adathalmaza 260000 rekordot tartalmaz.
* A rekordokhoz tartozó mezők:
  + Review: Felhasználók által adott vélemény
  + Label: Jelzi, hogy a felhasználó ajánja vagy nem a játékot
* A címkék többes bináris osztályúak

**ELőfeldolgozás**

* A review értéke Recommended vagy Not recomended lehet ezért azokat átalakítottuk számmá.
* Stopword-öket eltávolítottuk.
* A szöveg túl hosszúnak bizonyult a rendelkezésre álló erőforrásainkhoz ezért csökkentettük annak hosszát.
* Lényegesen több recommended volt benne mint Not recommended ezért a előbbiből levágtuk a szükséges mennyiséget
* A szöveget tokenizáltuk Word tokenizálás technikával.
* A szótárunkat felépítettük 40000 szó mérettel
* Az adathalmazt felbontottuk Train/val/test, 60/20/20 arányban tanításhoz és kiértékeléshez.

**Modell**

* A modellt egy neurális hálózat segítségével építettük fel.
* Az alábbi aktivációs függvényeket használtuk:
  + Sigmoid, mivel két osztályba sorolunk be értékeket
* Loss függvényünk BinaryCrossEntropy volt mert két osztályba sorolunk be.
* *A tanítást 10 cikluson keresztül végeztük mert...*
* A batch méret 150 volt.
* Az alábbi paramétereket próbáltuk ki
  + Vocabulary = 10000, Epoch = 10
  + Vocabulary = 5000, Epoch = 6
  + Vocabulary = 40000, Epoch = 12

**Kiértékelés**

* A tanuló adathalmazon a pontosság 79%-volt
* A validációs adathalmazon a legjobb paraméter beállítás pontossága: 77%-volt
* A teszt adathalmazon a pontosság: 68%-volt

**További fejlesztési lehetősége, Tapasztalatok**

Az adat jellege révén review értéke elég kiszámíthatatlan és a válaszok nem mindig pontosak nehezebb rajta dolgozni, esetleg a jövőben szükség lehet még néhány preprocessing lépésre, amelyek kiszűrik a nem releváns szavakat, helyesírási hibákat és a néhány szavas mondatokat. A vocabulary 10000-ről való növelésén kívül modell változtatása jelentősen nem javított a teljesítményen. Más tokenizálás és neurális háló modell segítségével nagy eséllyel javítható a felismerés hatékonysága.