**PROJECT BESZÁMOLÓ**

**Project neve**

*Amazon Food Review Predicter*

**Csapattagok**

* Barna Benjámin
* Kalló György Dávid
* Takács Balázs
* Tóth Ferenc

**Fejlesztési idő**

*68 óra*

**Adathalmaz**

* *A példa project adathalmaza 29769 rekordot tartalmaz.*
* *A rekordokhoz tartozó mezők:* 
  + Id: Az értékelés azonosítója
  + ProductId: A termékhez tartozó azonosító
  + UserId: A megrendelő azonosítója
  + ProfileName: A felhasználó felhasználóneve
  + HelpfulnessNumerator: Az értékelés segítőkészsége, hogy mennyire tartották jónak az értékelést a többi vásárló
  + HelpfulnessDenominator:
  + *Score: Az értékelés, amit majd megpróbálunk predictelni.*
  + Time: Az értékelés írásának időpontja
  + Summary: Az értékelés összegzése
  + Text: Az értékeléshez tartozó szöveg.
* *Az adatok címkézettek*
* *A címkék többes osztályúak*

**Előfeldolgozás**

* Először felosztjuk az adathalmazt train, val és test halmazokra, egyenlően elosztva bennük az 1, 2, 3, 4 és 5-ös értékelésű vélemények számát.
* *Az adatok címkei számok, ezért azokat nem kellett átalakítani.*
* *Megkevertük az adathalmazokat, hogy jobb tanulást biztosítsunk.*
* *Stopword-öket eltávolítottuk. Azért, mert nincs semmi jelentőségük és mindenhol megtalálhatóak a szövegben, ez megnehezítheti a tanulást.*
* *Stemming alkalmazása. Azért, hogy eltávolítsuk a ragokat a szavakról.*
* *Lematizáció alkalmazása. Azért, hogy kinyerjük a szótöveket, ezzel jóval csökkentve a különböző szavak számát.*
* *A szöveget tokenizáltuk BPE technikával.*
* *A szótárunkat felépítettük 10000 szó mérettel*
* *Gyakorisági adatokat nyertünk a nyers szöveg mellé....*
* *Az adathalmaz Train/val/test, 50/33/17 arányban van felosztva a tanításhoz és kiértékeléshez.*

**Modell**

* *A modellbe két konvolúciós réteget, maxpooling1D, GlobalMaxPooling1D és egy Dense réteget használtunk.*
* Az Embedding réteg, egy előre feltanított 50 dimenziós GloVe modell-t használ.
* *Az alábbi aktivációs függvényeket használtuk*
  + *relu*
  + *softmax*
* *Loss függvényünk SparseCategorcialCrossEntropy volt, mert a címkénink nem One-Hot kódoltak.*
* *A pontosságot Accuracy-val mértük*
* *A tanítást 5 cikluson keresztül végeztük mert megfigyeléseink szerint eddig tudott tanulni overfit nélkül*
* *A batch méret 32 volt.*

**Kiértékelés**

* *A tanuló adathalmazon a pontosság 80%-volt*
* *A validációs adathalmazon a legjobb paraméter beállítás pontossága: 51%-volt*
* *A teszt adathalmazon a pontosság: 52%-volt*

**További fejlesztési lehetősége, Tapasztalatok**

*Azt tapasztaltuk, hogy minden egyes újabb preprocessing lépés bevezetésével javult a hatékonyság. A modellnél rengeteg paraméter-kombinációt kipróbáltunk, többféle réteggel is megpróbáltuk, megpróbáltunk dropout réteget is berakni, de egyikkel se sikerült javítani az eredményen.*