PROJECT BESZÁMOLÓ

Bevezetés a Természetes Nyelvű Szövegfeldolgozásba

# Project bemutatása

Kindle review–k osztályozása -- Github repó

# Fejlesztési idő

70 óra

# Adathalmaz

* [Adathalmaz forrása](https://www.kaggle.com/bharadwaj6/kindle-reviews)
* A project adathalmaza 61.934 egyedi rekordot tartalmaz.
* A rekordokhoz tartozó mezők:
  + **#:** aposzt azonosítója
  + **asin:** a termék azonosítója
  + **helpful:** az adott review értékelése segítőkészség szempontjából
  + **overall:** A termék értékelése a felhasználótól
  + **reviewText:** A termék szöveges értékelése
  + **reviewTime:** A review posztolásának dátuma
  + **reviewerID:** A review azonosítója
  + **reviewerName**: A poszt írójának azonosítója (felhasználóneve)
  + **summary:** A poszt témája
  + **unixReviewTime:** 1970. jan 1.-e óra eltelt másodpercek a posztolásig
* Az adatok címkézettek
* A címkék többes osztályúak

# Előfeldolgozás

* Az adatok címkéi 1-5-ig terjednek, ezért átcsoportosítjuk őket *negatív, semleges* és *pozitív* csoportokká. Ezek jelölése az *overall* mező szerkesztése után: 1, 2, 3.
* A szövegből eltávolítottuk a nem használt oszlopokat, majd a *reviewText* mezőt megtisztítottuk azoktól a karakterektől és elemektől, amiket a *Stopwords* korábbi tapasztalatok alapján elkerült.
* *Stopword*-öket eltávolítottuk, mert feleslegesek számunkra, valamint félrevezethetik az értékelést.
* Tokenizációhoz és a szótár felépítéséhez *WordPiece* technikát alkalmaztunk, mert úgy véltük, hogy ez illeszkedik legjobban az adatbázisunkra és a megoldandó problémára.
* 3 csoportra osztottuk a review-okat a következőképpen: 1-2 negatív, 3 semleges, 4-5 pozitív.
* Az *overall* skálázása
* Az adathalmazt felbontása *Train(60%), Validation(10%)* és *Test(30%)* adatbázisra tanításra és kiértékelésre
* Az osztály kiértékelésének ábrázolása
* Szótár létrehozása

# Modell

**Simple rnn**

* A szótár méretek: 30.523 szó
* Az aktivációs függvény a „softmax”, a loss a "sparse\_categorical\_crossentropy",  
  pontossághoz pedig az „accuracy”.
* 5 tanítási cikluson keresztül a batch\_size az 32
* az 5. epoch végén a validációs adathalmazon a pontosság 72,6%.

**Lstm summary**

* A szótár méretek: 30.523 szó
* Az aktivációs függvény a „softmax”, a loss a "sparse\_categorical\_crossentropy",  
  pontossághoz pedig az „accuracy”.
* 4 tanítási cikluson keresztül a batch\_size az 64
* az 4. epoch végén a teszt adathalmazon a pontosság: 71%

# További fejlesztési lehetősége, Tapasztalatok

Másféle, hasonló termékek review-jaival összevetni, hogy meglássuk, azonos számbeli értékeléseket kapott termékek esetén a szöveges jellemzés mennyire tér el a jelenlegi adatbázisunkban látottaktól, ezzel megfigyelve a különbséget egy kultúrtermék (olvasásra szánt eszköz) és egy ettől a céltól teljesen eltérő eszköz között

Az űrlap teteje

Az űrlap teteje