**PROJECT BESZÁMOLÓ**

**Project neve**

*Amazon reviews*

**Fejlesztési idő**

*10 óra*

**Adathalmaz**

* *A példa project adathalmaza 1102318 rekordot tartalmaz.*
* *A rekordokhoz tartozó mezők:* 
  + *ReviewText: Értékelések szövege*
  + *Rating: Az értékelés pontszáma*
* *Az adatok címkézettek*

**ELőfeldolgozás**

* *Stopword-öket eltávolítottuk. Azért, mert ezzel eltávolítjuk a mondatokból a leggyakrabban használt szavakat, az a hipotézisünk hogy mivel ember által írt értékelésekről van szó, ezért sok stopwordot tartalmaznak, amelyek rontják a modell működését.*
* *Stemming alkalmazása. Azért, mert ezzel eltávolítjuk, a mondatok szavainak toldalékait, az a hipotézisünk, hogy ezzel javulna a modell teljesítménye és pontossága.*
* *Lemmatizáció alkalmazása. Azért, mert ezzel a szavak szótöveit határozhatjuk meg, az a hipotézisünk, hogy jelentősen javulna a modell hatékonysága és általánosítása ennek alkalmazásával.*
* *Html tag-hez eltávolító eljárást készítettünk.*
* *A szöveget tokenizáltuk WordLevel technikával.*
* *A szótárunkat felépítettük 15000 szó mérettel mert nem volt érezhető javulás nagyobb mérettel, illetve erőforrás takarékosság miatt.*
* *Gyakorisági adatokat nyertünk a nyers szöveg mellé.*
* *Az adathalmazt felbontottuk Train/val/test, 60/20/20 arányban tanításhoz és kiértékeléshez.*
* Az adathalmazból a nem használt oszlopok eltávolításra kerültek.
* Helyesírási hibák eltávolításra kerültek.

**Modell**

* *A modellt Logisztikus regresszióval építettük fel.*
* *Az alábbi aktivációs függvényt használtuk, mert 0 és 1 közötti értéket biztosít bármilyen bementre, ezáltal két csoportba osztható az adathalmaz elemei logisztikus regresszióval.*
  + *Sigmoid*
* *A pontosságot Accuracy-val mértük*
* *A tanítást 3 cikluson keresztül végeztük mert nem volt érezhető javulás.*
* *Az alábbi paramétereket próbáltuk ki*
  + *random\_state =0*
  + *max\_iter=2000*

**Kiértékelés**

* *A tanuló adathalmazon a pontosság 69.2%-volt*
* *A validációs adathalmazon a legjobb paraméter beállítás pontossága: 69%-volt*
* *A teszt adathalmazon a pontosság: 69%-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.*