

Félévi NagyHF specifikáció

Üzleti Intelligencia Labor

[Félév: 2025 tavasz]

Autók árának előrejelzése, Netflix és Youtube nézettségek, vélemények

Papp Inez Anna - (XX125Q)

Inez.papp.2000@gmail.com

Bemutató

A feladatom folyamán 3 adathalmazt használok, az egyikben autók eladási árai szerepelnek, a másikban országonkénti youtube nézettségek két különböző hónapban, a harmadikban pedig Netflixes sorozat értékelések. Az autós adathalmaz idősoros adatokat tartalmaz, így a LSTM alapú predikcióhoz fogom alkalmazni, a másik két adathalmazon pedig ETL jobokat fogok végrehajtani és reportokat készíteni belőlük. A Netflixesnél a KPI, hogy melyiknek a legjobb az értékelése, a youtubeosnál, hogy melyik ország nézi a legtöbbet, az autósánál pedig, hogy az ár emelkedik vagy csökken különböző tényezők figyelembevételével. Az ETL jobokkal megtisztítom az adatokat, majd átalakítom őket az elvárt formátumokba. A reportoknál különböző megjelenítéseket készítek, melyik interaktív, és az adatokról mutatnak be különböző nézeteket, ezzel valamilyen új információt bemutatva.

Főbb funkciók

Az alkalmazás funkcióinak bemutatása részletesen. Készülhet áttekintő ábra a funkciókról. (opcionális)

- Általam választott adatforrások:
 - o Autók eladási árai
 - o Országonkénti youtube nézettség
 - o Netflix sorozatok és értékelések
- Adattárolás
 - o Raw réteg: az adatok .csv fileokban vannak tárolva
 - o Stage réteg: megtörténik az adattisztítás, hibás adatok kiszűrése
 - o Data Warehouse réteg: dimenzió bevezetése
- Megvalósítandó ETL jobok
 - o Transzformáció 1: Hibás adatokat keres és kilógolja azokat a youtube-os adatforráson
 - o Job 1: Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot
 - o Transzformáció 2: Adattisztítás, hibás értékek kiszűrése a Netflixes adatforráson
 - o Job 2: Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot
 - o Transzformáció 3: Netflix értékelések alapján kategorizálja, hogy ez egy Jó, Közepes, vagy Rossz film, ehhez sávok vannak megadva, a jókat kimentti egy json fileba
 - o Job 3: a jó filmek feltöltése webszerverre
 - o Transzformáció 4: Youtube nézettség csoportosítás kontinensenként, majd aggregáció: átlagosnézettség a megadott két hónapban
 - o Job 4: Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot
- Az adatokból olyan adattípus készül, ami alkalmas arra, hogy a PowerBI fel tudja használni
- Reportok
 - o Report 1: Kördiagrammon ábrázolom a 3 legtöbb youtube felhasználóval rendelkező kontinenst, mellette egy táblázatban a kontinensek nevét és az átlagos nézettségüket a két megadott hónapban. A másik oldalon clustered bar charttal ábrázolom az országokat, és nézettségüket, és lefűrást használok.

- Report 2: Csatlakoztatom a youtube és netflix táblákat a Netflix nézettség ország és a filmet készítő ország alapján. Egy táblázatban megjelenítem az által a top 2 ország által készített filmeket, amely két országban a legmagasabb a youtube nézettség, az értékelésük szerint csökkenő sorrendben.
- Report 3: 3 táblázatban ábrázolom a Jó, Rossz és Közepes értékelésű Netflix filmeket, majd az x tengelyen évek szerint megjelenítem a kategóriákat egy scatter charton
- Report 4: Kördiagramon ábrázolom a Netflixes filmekben a műfajok arányát, egy másikon a film felfedezésének arányát
- Data Science
 - LSTM hálózatot használva az autó árakról szóló idősoros adatok alapján készítek predikciót a jövő adatokról

Választott technológiák

- Pentaho az ETL transzformációkhoz és jobokhoz
- Power BI a reportokhoz
- Python a predikcióhoz