Technologie pro velká data, 2. domácí úkol

Termín odevzdání: 31. 12. 2019

Forma odevzdání: zaslat e-mailem (jan.hucin@profinit.eu) zdrojový kód (PySpark) k oběma krokům zadání a tabulku průměrných počtů unikátních slov po žánrech.

Domácí úkol obsahuje tyto hodnocené kroky:

- vytvoření sloupce v DataFrame (1 bod)
- naplnění správnými hodnotami (3 body)
- analytický dotaz (průměrné počty unikátních slov po žánrech) (1 bod)

Za neúplné nebo jen částečně správné řešení se uděluje přiměřeně nižší počet bodů s granularitou půl bodu.

V supercvičení Spark (https://github.com/stameser/BDT/blob/master/cviceni/05_SPARK_SQL/spark_sql_cviceni.ipynb) proveďte kroky až do bodu 2.3. Následně proveďte tyto dva kroky:

1. Přidejte do DataFrame sloupec *slova_poc_unik* obsahující počet všech unikátních slov písně (pokud se slovo vyskytuje v textu víckrát, počítá se jen jednou). Slova jsou v textu oddělena mezerami. Zkontrolujte, zda pro písně s prázdným polem *text* je počet unikátních slov 0, a pokud to tak není, opravte v takových případech počet unikátních slov na 0.
2. Pro každý žánr spočítejte průměrný počet unikátních slov na píseň. Z výpočtu předem vyřaďte všechny písně s prázdným textem.

Pro výpočet počtu unikátních slov můžete s výhodou použít uživatelskou funkci (udf), ale možný je i jiný postup.