Big Data Analysis

Cilj i zadatak projekta je da se prikupe podaci, obrade statički i kroz streaming, analiziraju i vizualizuju. Za izradu projekta se koristi Databricks platforma i Spark (Python), uz mogućnost korišćenja dodatnih biblioteka. Notebook je potrebno detaljno dokumentovati, objasniti primenjene transformacije i dobijene rezultate analize u okviru markdown ćelija.

Projekat se radi **samostalno**, nosi 50 bodova i sastoji se iz dve celine (transformacija i analiza). **Izbor teme projekta je ostavljen studentu.** Za svaku celinu je potrebno kreirati poseban Notebook.

1. Prikupljanje i transformacija podataka

Prvi deo projekta uključuje pronalaženje odgovarajućih datasetova za obradu što podrazumeva pretragu javno dostupnih podataka sa odgovarajućih izvora i zatim učitavanje istih u Databricks Notebook. Takođe je potrebno izvršiti transformacije tih podataka, te ih prilagoditi za analizu.

Potrebno je koristiti minimalno 2 dataset-a i izvršiti minimalno 15 transformacija**.**

Postoji dosta izvora javnih podataka. Neki koje možete da pretražite:

- Kaggle
- data.gov.rs
- data.gov (US)
- EU open data portal
- Github
- Public APIs
- FiveThirtyEight

Mogu se koristiti i fajlovi iz DBFS-a sve dok se ispoštuju ostala ograničenja projekta.

Prilikom učitavanja dataset-a u Notebook, potrebno je napisati sadržaj (imena kolona, opis) i izvor.

Obrađene podatke je potrebno sačuvati u okviru DBFS-a.

2. Analiza podataka

Transformisani podaci se učitavaju u nov Notebook za potrebe analize istih.

Pre svega, potrebno je kroz minimalno 10 chart-ova vizualizovati podatke (pre ispitivanja hipoteza). Za vizualizaciju koristiti smislene upite koji nisu korisceni prilikom transformacija i minimalno tri različita display type-a.

Zatim je potrebno definisati minimalno 7 hipoteza i analizom ih potvrditi ili opovrgnuti. Ispitivanje hipoteza se radi relevantnim upitima, vizualizacijom dobijenih rezultata, a zatim i donošenjem zaključaka. Umesto postavljanja hipoteza, dozvoljeno je analizirati podatke primenom odgovarajućeg ML algoritma (ukoliko imate iskustva sa tim). Preporuka je u ovom slučaju koristiti Spark MLlib, ali nije nužno i ograničenje.

Nakon ispitivanja hipoteza, potrebno je odraditi structured streaming. Za streaming koristiti jedan od transformisanih datasetova. Prilikom streaming-a koristiti jedan isti upit više puta, kako bi se vizuelno prikazao protok podataka (kao što je odrađeno na vežbama).

Obratiti pažnju na optimizaciju koda, koristiti keširanje tamo gde je potrebno, po završetku structured streaming-a ugasiti sve što je potrebno i sl.

Analiza podataka se neće bodovati ukoliko pre toga nisu odrađeni prikupljanje i transformacije.

Kod mora biti dokumentovan (propraćen odgovarajućim komentarima u markdown ćelijama) inače vredi 0 poena.

Bodovanje i predaja

Projekat vredi 50 bodova, **uslov** za polaganje je 25 i boduje se na sledeći način:

	Poeni
Prikupljanje i Transformacija	15
Vizualizacija	5
Analiza / ML	20
Structured Streaming*	10

U slučaju da je neka od stavki implementirana parcijalno, biće dodeljeni parcijalni poeni.

* Umesto Structured Streaming-a, može se raditi Batch Processing koje vredi maksimalno 4 poena.

Predaja i odbrana:

- Rok za predaju ukoliko polažete u februarskom ispitnom roku: **18.02. 23:59**
- Potrebno je kreirati folder na drive-u i u njega ubaciti korišćene datasetove kao i notebook-ove u **ipynb** i **html** formatu (notebook-ove exportujete nakon izvršavanja svih ćelija)

- Link ka drive foldera se šalje na email adresu mbakic@raf.rs
- Subject mail-a mora biti: BDA Februar Ime Prezime <Broj indeksa>
    - npr: “BDA Februar Student Studentic RN 1/12
- Zakasneli projekti ili projekti koji nisu poslati po prethodnom uputstvu vrede 0 poena

- **Odbrana projekta je obavezna i biće održana u terminu ispita (21.02). O tačnom rasporedu bićete obavešteni.** Ako je student iz bilo kog razloga sprečen da prisustvuje odbrani, obavezno to najavite što pre, kako bi se organizovao vanredni termin za odbranu.