MovieLens Data-analyse

Læringsmål

Ved afslutningen af denne case vil I kunne:

Sætte et miljø op med Apache Spark

Arbejde med PySpark

Udføre grundlæggende og avanceret data-analyser

Anvende data til at træffe data-drevne beslutninger

Case-beskrivelse

Zentropa, et dansk filmselskab, har samlet store mængder data om film, deres genre og hvordan de er blevet vurderet (ratings) gennem tiden.

Målet er at bruge denne data til at forudse, hvilke filmgenrer der generelt opnår de bedste anmeldelser, og som dermed er mest lovende at satse på for fremtidige filmproduktioner.

Det er jeres opgave at:

Analysere datasættet

Finde den genre, som i gennemsnit har de højeste ratings

Præsentere jeres resultater for Zentropa (dvs. klassen)

Oversigt

Analyser datasættet – Dan jer et overblik over datas indhold og opbygning

Find top-genren – Identificér hvilken filmgenre, der i gennemsnit har den højeste rating

Præsentation – Afslut med en kort præsentation af jeres undersøgelse og fremgangsmåde

Opsætning af miljø

Installer Apache Spark og PySpark på jeres computere

Find en installationsvejledning på nettet

Hent datasæt

Gå til *MovieLens* og hent det største datasæt, I kan finde. Note: Hvis I ønsker, kan I godt bruge et andet datasæt, så længe I er i stand til at udføre tilsvarende analyser

INDLÆS DATA I PYSPARK

Indlæs MovieLens-datasættet i PySpark

Få et overblik over datasættets struktur ved at analysere dets schema (kolonner, datatyper osv.)

Overvej: Hvordan er data opbygget?

BASAL DATAANALYSE

Hvor mange unikke film er der i datasættet?

Hvad er den gennemsnitlige rating for hver film samlet?

Lav en liste over de 10 film, der har den højeste gennemsnitsrating

AVANCERET DATAANALYSE

Konverter tidsdata: Lav om på timestamp (datetime), så det er let at læse (f.eks. til årstal)

Analyser ratings over tid: Undersøg, hvordan ratings for en bestemt film har ændret sig over en periode på 10 år

Forbind film med genrer: Match filmene med deres respektive genrer

Udregn ratings for hver genre: Find gennemsnitsrating for hver genre, og identificér den med højest gennemsnit

Ekstra opgave: Indsæt data i grafik

PRÆSENTATION

Når analysen er færdig, skal I forberede en kort fremlæggelse (5-10 minutter) for Zentropa (klassen).

Jeres præsentation skal inkludere:

En gennemgang af metode og fremgangsmåde

De vigtigste fund fra jeres analyse

Eventuelle anbefalinger baseret på jeres resultater

Tip: Brug gerne grafik (visualiseringer, grafer, PowerPoint med screenshots), der understøtter analysen.

VURDERINGSKRITERIER

Korrekthed: Er koden korrekt, og opfylder resultaterne opgavens krav?

Tydelighed: Er koden kommenteret (på engelsk) og let at følge?

Indsigt: Hvor dybdegående er analysen og konklusionerne?

Præsentation: Hvor godt formidler I resultaterne, og er visualiseringerne klare og meningsfulde?