11/10 lecture

viernes, 11 de octubre de 2024 16:23

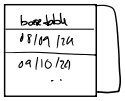
Trusted Zone

1) Combine All data

- IMDB → monthly, disjoint

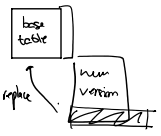
movies
actors
directors

- Netflix users → eventual, disjoint



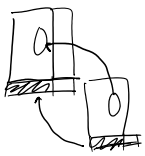
- Netflix users → daily, incremental

option 1:



+ simple
+ updated scores
We lose timestamps

option 2:



- logic

We keep timestamps

option 3:

↓ the table can get very big...
↑ we have all information

2) Check data quality: generic steps to avoid problems in data.

- Profiling: extract mean, value, IR, range, ... → quantitative
  nan, ... → qualitative
  } Analyse the data with which we are working. Detect outliers!
  - extreme
  - less extreme outliers that could be mistakes
- Removing duplicates: Some might have appeared in step A, combining data sources.
- Consistent format: remove special characters,...
- Data normalization: (categorical columns) "m1" "[a1, a2, a3]" → m1 a1 / m2 a2 / m3 a3
- Misspellings: typos!
  - Difficult to detect...
  - A) Easy way → use a spellchecker
  - B) more advanced way → count times that each category appears + Levenstein distance
  - C) Even ↑ tricky → Lemmatization/Stemming

- Outliers
- Missing values
- Scaling

} We don't handle this categories since the applied methods can add bias before training the model. We will handle them in the exploitation zone.

EXPLOITATION ZONE → Expose data to the analysts

Guidelines: how to do the exploitation zone! Some tips

- Explain in detail why and what you do!
- Separate entities (static info) vs actions (dynamic info)
- Reduce redundancy and increace query efficicency
- Prepare for more data, don't "overfit" to the data we have now.
- Build Ground truth (KPIs) --> We want to give to the analyst the more information, so we should generate KPIs for them to analyse

Actors = IMDB actors

opt 1

m1	a1
m1	a2
m2	a3

Option 2: the good one (do this)

One table for actors: a1 _
a2 _
a3_
Ona table for acts : a1 m1
a2 m1
a3 m2

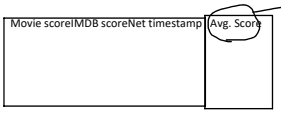
Movies= IMDB movies + Netflix movies

Join operation, big table containing all the information

Possible issues:
- Not easy join bc of different granularities
- Overlapping information (repeated) in the attributes-> Decide which one you take!

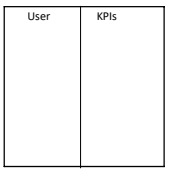
WE DON'T INCLUDE THE SCORES IN THE TABLE!,we will set them in another table, and do the same as in the actors configuration

Scores table will be :



The avg score could be, for example, an indication of a KPI

User views (3rd data source)



User movies