

Project presentation

Machine Learning and Big Data Processing(ELEC–Y591)

Cédric HANNOTIER Mathieu PETITJEAN Hasan Can YILDIRIM

June 13, 2018

Outline

- 1 State of the art
- 2 Approach
- 3 Results

State of the art

Outline

- 1 State of the art
- 2 Approach
- 3 Results

The Netflix dataset

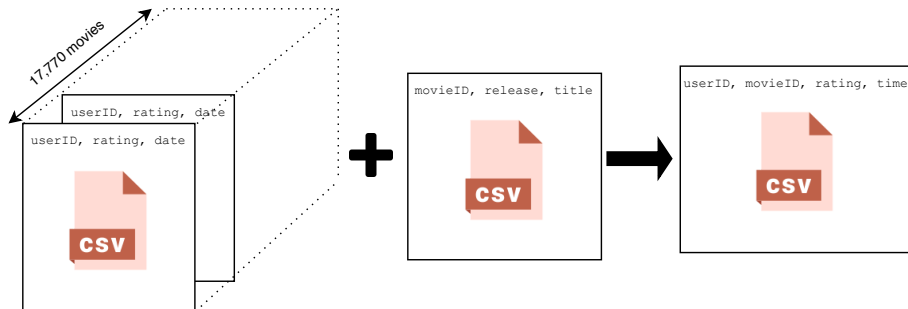
- 5.5 GB of data
- 17,770 movies
- 480,000 users
- \simeq 100 million ratings

against

The MovieLens dataset

- 875.6 MB of data
- 27,278 movies
- 138,493 users
- \simeq 20 million ratings

Netflix data reshaping



- 1 Discard MovieLens entries based on timestamps.
- 2 Discard all movies not present on both datasets. A movie was uniquely identified by its title and release date.
⚠ "Lord of the Rings, The" and "The Lord of the Rings (2001)"
- 3 Recast timestamps: from YYYY-MM-DD and elapsed seconds to common reference.
- 4 Rounded MovieLens ratings.

⇒ 5800 common movies, 52,875 users
remaining in ML and 478,756 in Netflix.

Outline

- 1 State of the art
- 2 Approach
- 3 Results**

Results