Computational Methods for Document Analysis 01

Mirco Dietrich
Maurice-Roman Isele

18. April 2016

Exercise 2

The data file has a header (line 1-5), apparently containing information about the format of the following data. Assumably, the data is intended to be in a table format. The data contains amazon reviews about dvd's. Each review contains the following information:

productId, userId, profileName, helpfulness, score, time, summary, text

Apparently, those reviews were not only written by "normal" users, but also by some "experts" (e.g. the first two reviews). All reviews are refering to movies of various genres, which makes me guess that an online platform for movies collected those reviews (that would also explain why everyone has a profile name).

For the product with the productId "B00005CDCM" (movie 'Space Cowboys'), there are a total of 121 reviews. Most of the reviews are very positive with a score of 4 or 5, except for some like the review of "spacecadet007", who thinks the movie is technically very inaccurate and recommends you shouldn't even rent this movie. But only 3 out of 13 people agree with him (helpfulness rating).

There are reviews of every Harry Potter movie except HP Deathly Hallows Part 2, so the data probably was taken before the release date of HP DH 2