Intro

The expansion of internet and social media such as Facebook and recommendation sites like Rotten Tomatoes and Yelp significantly enlarge the scope of customer-to-customer (C2C) interactions(Libai et al., 2010). People are encouraged to share thoughts and hands-on experience on products or services they have purchased. Therefore, numerous subjective opinionated reviews are produced. Such reviews are significant for both merchants and readers. Merchants could retrieve customers’ feedback in order to further improve their products and customers are able to make informed decisions pertaining to day-to-day tasks and purchase. Hence, review mining is employed to “extract positive opinions, negative opinions and object features for review sentences”.

At present, most papers on the treatment of film criticism simply analyze the emotion of film criticism, identify and determine whether the specific words in the review are positive or negative emotion

Recommendation System

We build the recommendation system in following steps:

Firstly, preprocess the raw data.The processing mainly includes the following steps:

(1) Delete stop words and punctuation

(2) Case conversion (turn all letters to lowercase).

(3) As for directors and actors, the processing data requires all first and last names to be combined together into one word. To achieve this, we remove the spaces between the first name and last name after case conversion. For example, two directors Chris Evans and Chris Hemsworth will have different values by connecting the first and last names, however if we treated first and last name as different tokens, the same value will be 50%.

(4) Match the optimized audience ratings through rotten\_tomatoes\_link.

After preprocessing, we can get the data as shown in the figure. 翻至下一页再返回

Now we get five types of data, including movie plots, genres, directors, cast and optimized audience ratings.

Secondly, feature construction. Use the RaKe function to extract keywords from the entire sentence in the "movie\_info" column to build a new column ‘Key\_words’.

Thirdly is data encoding and quantification. We encode character data and standardize the encoded data to eliminate the dimensional relationship between variables, so as to make the data comparable.

Finally, calculation of similarity. We combine genres, directors, cast, adjusted\_scores and Key\_words, and extract the word bag that builds this information called ‘Bag\_of\_words’. Then we use the CountVectorizer method and cosine similarity to calculate the similarity among movies. Finally, we can obtain the movie recommendation based on the similarity.