豆瓣读书数据集(Douban Dushu Dataset)

Dateset Description:

DouBan DuShu is a Chinese website where users can share their reviews about various kinds of books. Most of the users in this website are unprofessional book reviewers. Therefore, the comments are usually spoken Chinese or even Internet slang.

In addition to the comments, users can mark the books from one star to 5 stars according to the quality of the books. We have collected more than 37 million short comments from about 18 thousand books with 1 million users. The great number of users provide diversity of the language styles, from moderate formal to informal

Data Preprocessing:

1. convert full width symbols to half width symbols
2. remove some special symbols
3. convert traditional Chinese to simplified Chinese

Terms of Use:

1. Respect the privacy of personal information of the original source
2. The original copyright of all the data belongs to writers of the reviews and DouBan
3. The dataset is **only** for study and research purposes. Without permission, it may not be used for any commercial purposes
4. Redistribution is **NOT** allowed
5. Some items must be deleted if the copyright owners claim
6. If you want to use the dataset for depth study, please cite this paper: TBD

Application Form

Name: _____
Affiliation : _____
Address : _____
Zip code : _____
Phone : _____
Email : _____

I have read the above terms and conditions, and are willing to accept the terms. (The following part will be signed by the person in charge of the research group the applicant in. Under the terms).

Name: _____ Position: _____
Signature: _____ Date: _____