IMDb POLAR Dataset

This dataset is released for research that involves polarized items, substantially better data than previous benchmark datasets. The scope of research is typically where ratings are used for data analysis or to make predictions, such as in recommendation systems, in particular when studying the effect of algorithmic bias and iterated bias in recommender systems. This research is supported by the National Science Foundation INSPIRE grant IIS-1549981.

When using this dataset, please cite the following paper:

We constructed a balanced collection of movie ratings from IMDb, by crawling polarized movies based on their histograms. After pruning and annotating the dataset, we ended up with 612 polarized movies, each with at least 50 ratings; then we crawled all the reviews for these movies. We also crawled an almost equal number of randomly selected non-polarized movies from different genres and years. Similar to other movie rating datasets, the movie popularity distribution follows a log-normal distribution. In the interest of providing a benchmark for future work in this area, we are releasing this dataset to the public.

The IMDb POLAR dataset contains 1,340 movies and 427,074 ratings, with a size of 10 MB. The data was collected during the last week of March 2017. Each movie has a rating on a scale from 1 to 10.

We labeled each item as polarized or non-polarized using our annotation methodology as described in the paper, here.