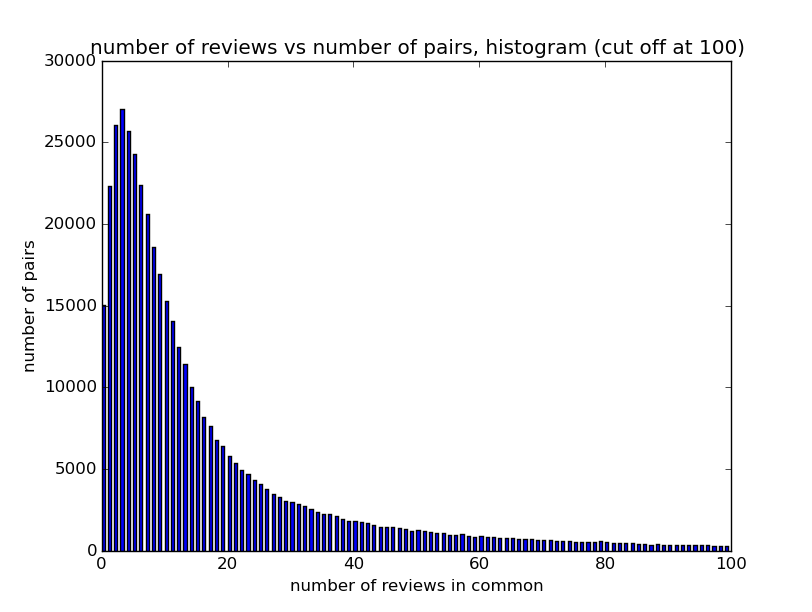
**Problem 1a)**

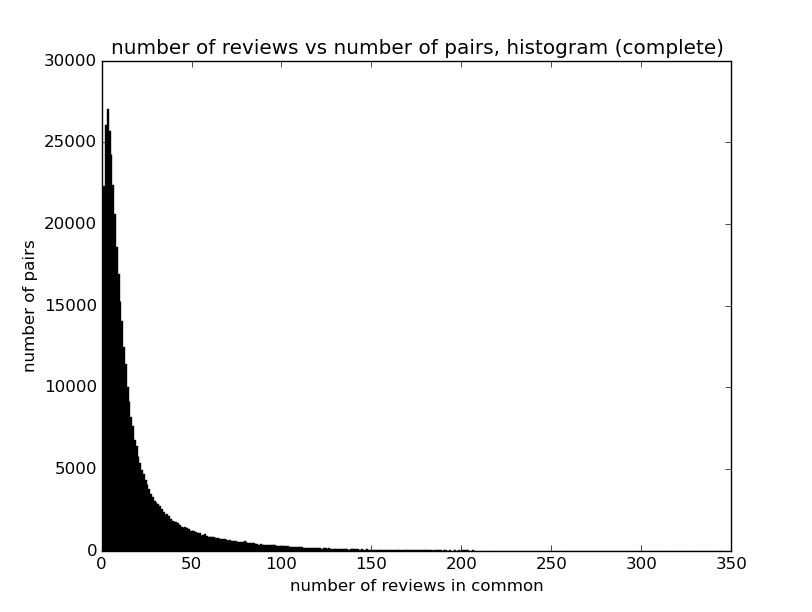
-mean of number of reviews in common: 18.8079220449

-median of number of reviews in common: 10.0

-the entire graph extends well beyond 300 reviews in common (~340) on the horizontal axis, however, that number of overlap between different users is very rare, thus, to make the general distribution more “visible”, I also show one graph zoomed in on the section between 0 and 100 movies reviewed in common, Figure 1. The entire graph is shown in Figure 2.



Figure



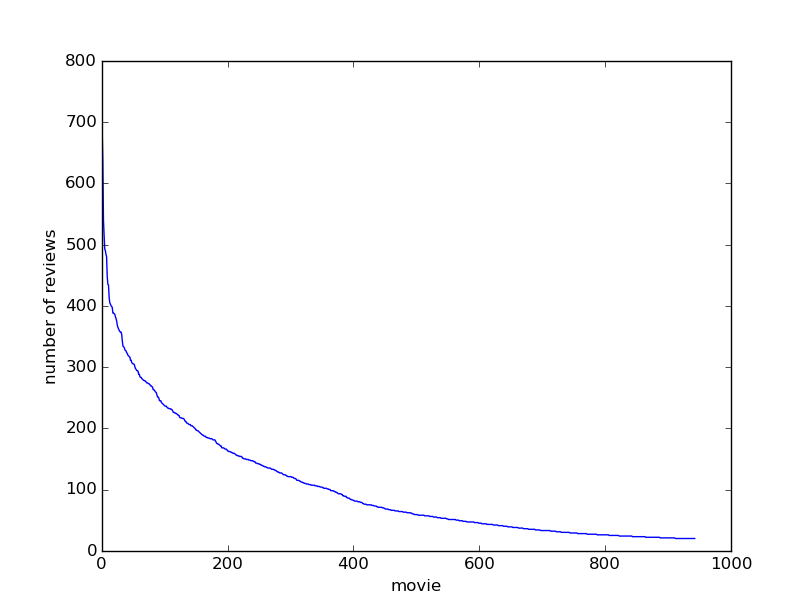
Figure

**Problem 1b)**

Maximum review count: 737 found at index: [404]

Minimum review count: 20 found at index(indices):

[18, 33, 35, 92, 142, 146, 165, 201, 241, 299, 308, 363, 417, 440, 474, 557, 570, 571, 595, 630, 635, 684, 731, 739, 808, 811, 823, 865, 872, 887, 894, 925]



Figure

Figure 3 shows this “line”. This is similar to most of the examples listed on the Wikipedia page. Thus, I say yes, it does follow Zipf’s law.

**Problem 2a)**

I think approach A is better. This way, we won’t skew the results by inserting a value once we do the kNN evaluation for the movie rating. In addition, once we do the kNN, we can treat a 0 as a flag for data that we should not include when averaging. Furthermore,

**Problem 2b)**

xxxxxx