Group Name: Amazon Reviews
Group Members: Michelle Liang, Josef Klafka, Timmy Li, Jon Pekarek

We plan to use an Amazon review dataset that is 18 GB and sorted by user, with duplicate reviews already removed.
The dataset is a json file that contains 83.68 million reviews.
The link below contains accessible subsets of the big dataset, sorted by category of goods (eg. Books, Electronics, etc).
We have received permission from the owner of the big dataset to use it for our project.

Link to Dataset: http://jmcauley.ucsd.edu/data/amazon/links.html

Below are some questions we hope to answer by analyzing the data:

1.) Is there a tendency for a user to leave positive or negative reviews?
2.) Is there a correlation between the number of categorical goods reviews for a user and whether
their reviews are positive or negative?
3.) Do users tend to use the same positive or negative descriptive words?
4.) Correlation between the descriptive words and the helpfulness of the review. What kind of words do helpful reviews
use? What kind of words do unhelpful reviews use?
5.) We plan to do collaborative computing where we form all possible pairs between reviews to find the overlap of
descriptive words and give each pairing a similarity score. We want to then look at the pairs with very high similarity scores
and compare their helpfulness points to see if using certain descriptive words give reviewers a higher helpfulness score or
lower helpfulness score.
6.) We also want to find two reviews from two different reviewers that are very similar to each other and see whether that
means that the rest of their reviews are also very similar to each other.