# README

Dataset file name: book\_data\_1225.csv

In this new dataset, 1225 unique books are collected. A lot of features are added, but not too many that if you open the dataset with Excel it will explode.

As for now, when you feed the dataset into your Machine Learning algorithm, please use two groups of features:

**Group 1:** 3, 4, 5, 6, 7, 8, 10, 11

**Group 2:** Group 1 + 12 ~134

Both target on feature 1

Below is the data description:

We have 134 features in this dataset.

1. Rating: this is the target, as you might all know.
2. Title: as for now, we won’t include this feature into our machine learning algorithm testing. But the next dataset I’ll give you (in the next few days?) will have a bag-of-word transformation based on the titles.
3. Price: yeah, price.
4. Nook: whether the book comes with Nook (eBook) version.
5. Audio: whether the book comes with Audio version.
6. Hardcover: whether the book comes with Hardcover version. Note that originally all the values for features 4-6 are TRUE or FALSE. I changed them to 1 or 0.
7. Publisher: publisher of the book. I encoded them into numbers. Please find the “publisher\_encoding.txt” file for corresponding ids.
8. Published date: the date the book was published.
9. Until 1206: (added) how long has the book been published. I basically just computed the difference between the published date and 12/06/2014 (which is when the data description was written).
10. Pages: how many pages this book has.
11. Number of reviews: how many reviews this book has received on Barnes and Nobles so far.

12 ~ 134. Subjects: 122 subjects form a very long feature vector. If the book is classified as “history”, then under the “history” feature, it will have a value of 0. From feature 12 to 134, the data is extremely sparse. Also, note that I took out all of the subjects with less than 10 books.

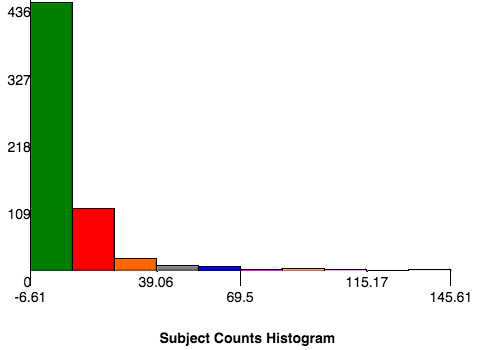
In total, we have 571 unique subjects

259 unique publishers

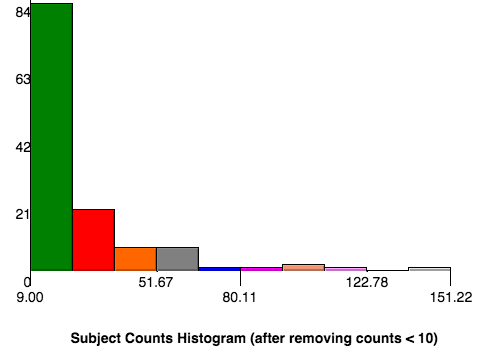
960 unique authors

This time we have books that are rated as one star. (The dataset covers all possible ratings, even though unbalanced)

Next page is just some of my sidenotes.



After deleting counts < 10



Then we have 122 subjects left.