Seattle Library's Most Controversial:

Mainstream is boring! Is it possible to use the information in the library database to determine which items are the most controversial? Can we implicitly discover which books or DVDs are strongly loved by few, but detested by many? Is it possible to discover subcultures and gauge people's varying passion for different topics?

I attempt to quantify the above concept by examining variations in durations of checkout, popularity, and page length to determine a 'controversy score' for each item. Then, we can plot which subjects are the most controversial, and by inspecting individual item names we can discover which items have the highest variance of interest, and are thus the most titillating items in the library.

Duration can be easily taken as "cin - cout". I use UNIX_TIMESTAMP() to simplify working with dates. For 'popularity', we can simply do a join with the 'popularity' table. However, page number is a lot trickier. Actually, I got pretty far into this project before I realized the 'pages' column in the extras table was garbage. However, after quickly inspecting the library website, it's pretty clear a web crawler can snatch the information that's needed with a small amount of effort. After some investigation, it seems like the bib number used by the URL always has a '030' and a redundant code appended to it... strange. Are they trying to keep someone out? Or create the impression of security? At any rate, I will fetch the page counts at a later date if possible and add it to the project as described below

Part of the challenge in creating a 'controversy' score is determining what constants to use to weight the inputs (checkout quantity, duration, and page length). I began by determining average checkout duration for all books in the library, and the corresponding variance, min, and max.

Duration of checkout (days)	
Average	228
Stddev	1669
Min	0.000694 (1 minute)
Max	16075 (44 years)

The average checkout time seemed quite high, given the loan period is only 21 days. I thus limited the query by ignoring books that were offensively delinquent (checked out more than 90 days). This hopefully removes the outlier effect.

Duration of checkout (days), delinquents removed	
Average	20.23
Var	16.7
Min	0.000694 (1 minute)
Max	89.99

Looks much more reasonable now.

Now, we need to understand the popularity of each bibitem, so I repeat the above process looking at popularity rather than checkout duration.

Popularity Statistics

Average	80.7826
Var	286.4692
Min	1
Max	18552

A reasonable approach to measuring controversy is to look at items that are quite popular, but have a high variance of checkout times, indicating that some users loved the book (and thus may have kept it past due) while others could not make it past the first few pages and returned it immediately. We can normalize checkout duration by page length when it becomes available..

$$controversy_{bibitem} = \alpha * \frac{popularity_{bibitem}}{popularity_{max}} + \beta * \frac{Var(duration_{bibitem})}{duration_{max}}$$

This formula will reward items that are checked out often and have a highly variable checkout time (normalized by book length). Alpha and Beta can be used to tweak the scoring down the road (for now I use a value of 1). The page counts aren't available yet, but we can add that variable as a quotient on the right.

Because of the large size of the database, I limited the range of the query to just the philosophy and theory of religion (210). The results and query are included in separate files (the final query in the list). Interestingly enough, the controversial movie 'Religulous' showed up as the highest ranking result.

There are a few problems with this approach, which essentially stems from a lack of information. We do not know if people will return their book immediately if they do not like it - they may just leave it on the shelf until the last minute. Given that it seems like the average checkout time is much longer than the actual loan period, there could be quite a bit of noise in the results, even though we removed most of the truly delinquent records. Additionally, different people have different reading speeds, and there's no way to account for this.

A possible improvement is to award even higher values to books that were returned in the first 24 hours. We can also run the same kind of analysis on a subject level to get even more interesting results.