MAT 259: Final Project

Grant McKenzie | March 2014 | grant.mckenzie@geog.ucsb.edu

Concept

From a research perspective, one of the most fascinating aspects of a dataset such as the Seattle Public Library check-outs is exploring the similarities between items. By similarity, I mean the ways in which two items are similar or dissimilar. For example, do the authors of two books use similar language? Is the subject matter similar? Do the topics or themes discussed in the books flow along the same lines? Given this concept of topic or theme similarity, I've always thought it would be very interesting to visualize data points (books in this case) in thematic space. In this case the placement of an item in a three-dimensional space would be based on its similarity to other items in the same set. For example, two children books on the topic of dogs would be placed very closely together in space, while a book on Cold War politics would be placed much farther away.

One statistical method for approaching this idea is Multi-Dimensional Scaling (MDS). MDS takes a series of attributes for each item in a dataset and uses these attributes to compute a location in N-dimensional space. Items that are more similar are clustered together while those dissimilar are placed further apart. The number of dimensions "N" is up to the user to determine and in this case I have chosen to represent the data in three dimensions.

The basis for MDS will be a series of attributes related to each item. These attributes will be based on descriptive data pertaining to each item. Given textual descriptions of each book, for example, a latent Dirichlet allocation (LDA) model can be run resulting in a finite set of topics. LDA is an unsupervised, generative topic model that approaches text as a bag-of-words. The co-occurrence of words in a specific document and across documents produces a set of topics from which each original book, in this case) can be defined. In essence, each book will be given a unique distribution of topic probabilities. This distribution of topics can then be compared to each other distribution of topics (through a Euclidean distance measure for example) and a resulting matrix of similarity values is produced.

Data

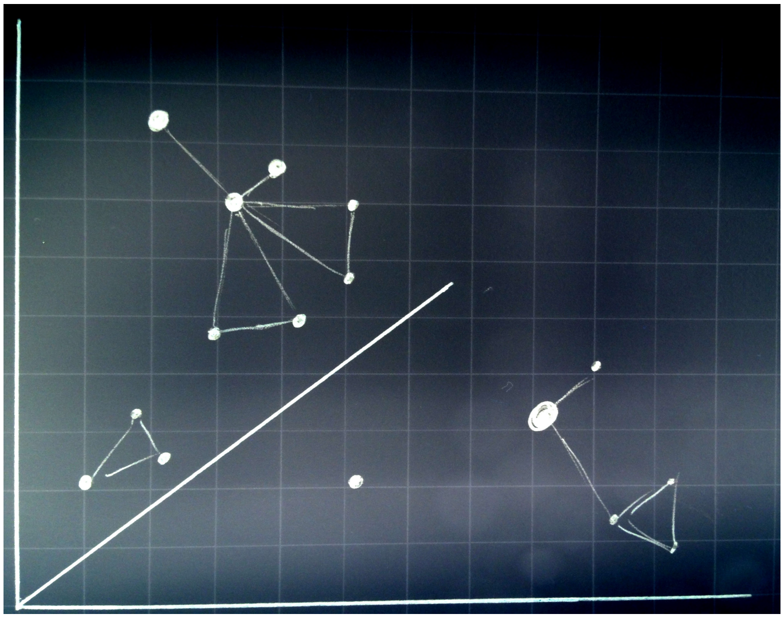
The first step in collecting data for this project is to determine the number and type of items that should be explored. Given the vast amount of data it is sensible to reduce the study dataset. For this project I chose to take the top 5,000 most check-out *books* from the Seattle Public Library. The query to get the *titles*, *bibNumbers* and *check-out counts* for these books is listed below.

```
SELECT bibNumber, title, count(*) as ct
FROM public.spl2
WHERE substring(title,3,2) == 'bk'
GROUP BY bibNumber, title
ORDER BY ct DESC
LIMIT 5000;
```

After exporting these results in CSV format, the next step is to ascertain more descriptive content about the specific books. In order to do this I chose to explore the *Google Books API.* Limited to 1,000 requests a day, the title of each book is queried against the *Google Books Database* and metadata about each book is returned. The following information can be accessed and downloaded:

- Direct Google Books Item link
- Title
- Authors
- Publisher
- Published Date
- Description
- ISBN
- Page Count
- Categories
- Thumbnail
- Language
- Text Snippet

Doodle



Design Decisions

The results of MDS will be 3 numeric values (bounded between 0 and 1) for each Book in the dataset. These 3 values will represent the X, Y and Z parameters of the three dimensional space. Given the similarity value matrix that shows a similarity value for each book to each other book, I will draw lines between books that have similarity values in the top 30%. This will add more structure to the visualization and show clusters of books that are considered similar. Additionally, nodes (books) in the visualization will be labeled with the book title styled based on the Category of book. One would expect that books of the same category will be clustered together, but this may not be the case. The size of the node (volume of the sphere) will be based on the total number of checkouts of the book in question. This should lead to a very interesting visualization combining location, node size, node color and strength of relation indicated by a line.

More to come as I play around with the different parameters.