Thomas Hervey
MAT 259
1-14-16

Assignment 1

Assessing the Temporal Popularity of New Titles

Introduction

Seattle Washington, a progressive growing area with heavy technological and artistic influences is one of the American cities at the center of the information age. Because of this, it makes sense that the Seattle public library system is both large and diverse with many frequent users. The 21st century has seen a surge in consumer information, which usually takes many forms. Every year, records are being broken for box office purchases, number of video rentals, and book series publishing. However, many suggest that the growth rate of the world wide web has made other media and information sources obsolete. While there may be a larger growth in interest for ebooks and online media, I predict that physical media will too become more popular over time.

Interest Areas & Questions

> Based on the evolution of the data, I seek to answer the following questions:
> -Does a title's age affect its checkout frequency? *(ie. same book in 2006 vs 2015)*
> -Do checkout frequencies increase with new media? *(ie. 2006 release vs 2015 release)*
> -Has there been an increase in young adult interest? *(interest in hyped-up genre)*

The below results are preliminary and have further tweaking based on visualization needs. For example, to look at new media popularity, it must be determined when an item first becomes available for checkout. After this information has been generated, I will likely incorporate several statistical and data analysis tools to generate regression results, and possibly some basic machine learning.

The query currently pulls from three tables to synthesize unique identification information, all checkout and checkin information, several important dates (ie. publishing year), and genre. Several 'quality' columns (ie. pubNewYear, cOutCount) were generated to further help separate data during follow up queries. *Please see comments within the query for further details.*

Analysis

After running into a few issues, I found out that my SQL query has a runtime error, preventing it from terminating. Therefore, I limited the results to 10000 which took ***4.082 seconds to complete***. Preliminary results suggest that some of my hypotheses are correct. Looking at

several particular items such as Herper Lee's popular new *Go Set a Watchman*, frequency is very high initially with a dip after a few weeks. Other titles follow similar trends and generally taper off as time goes on. Further investigation needs to be made into my second two questions comparing newer and older media. In addition, I would like to modify the web scraping script to help eliminate publishing year errors from the x_splOrgWebScraping table.

```
SELECT
   rawcIn.itemNumber AS id, -- to uniquely identify
   barcode AS barcode, -- to uniquely identify
   checkOutCount AS cOutCount,
   bibNumber AS bib, -- to identify multiple activity of same content
   title AS title, -- sanity check
   itemType AS itemType, -- genre for further 'young adult' analysis
   type AS pubYear, -- published year
   checkOut AS cOut, -- observe checkout frequencies
   MONTH(DATE(checkOut)) AS cOutMonth, -- observe checkout frequencies
   checkIn AS cIn, -- observe checkout frequencies
   MONTH(DATE(checkIn)) AS cInMonth, -- observe checkout frequencies
   -- if published since library opening & checkouts began
   CASE
      WHEN type >= 2006 THEN 'yes'
      ELSE NULL
   END AS pubYearNew
   -- TODO: Determine arrival criteria
         -- CASE CURDATE() -- (-) arrival
                  -- END as ageAtLibrary
                  -- must determine arrival!!!
FROM
   spl3._rawXmlDataCheckIns AS rawcIn,
   spl3.x_splOrgWebScraping AS scrape,
   spl3.x_checkOutCountItem AS cOutCountItem
WHERE
   rawcIn.bibNumber = scrape.bib
      AND rawcIn.itemNumber = cOutCountItem.itemNumber
ORDER BY id
```

Sketches, Low-Fidelity Mockups & Predictions