Novemer, 3, 2017

Social Network Analytics, Empirical Exercise #4
Due on Friday, November 17, 2017 at 8:00am

Similarity of feature film production companies and their films

1. There is a .zip folder on canvas containing the file "keywords_films_producers.csv", that contains information on the films a producer has made and the plot keywords these films are tagged with. In the columns of this file, "pindex" is a numerical identifier for each film, "title" is the film's title, "year" is the year the film was made, "keyindex" is an identifier for keywords, "keyword" is the plot keywords that appear in the film, "pcindex" is an identifier for production companies, and "prod_company" is the name of the production companies.

   Use the keywords shared between two producers to determine how many producers are maximally similar according to cosine similarity. Is this number the same for Jaccard similarity? Perform this calculation directly on the matrices—the original "keyindex_producers.csv" file has been manipuated into another file, "production_keyword_matrix_1985.csv", that contains an incidence matrix of films that were produced in 1985 and the keywords they contain. The `unmatrix()` function in the gdata package may be helpful for examining the distance result. If you like, you can try to perform the manipulation of the original data on your own.

2. ~~There is also a file with box office data for some of the films in the file, "box_office_films.csv", also on Canvas. Do producers seem to change their strategy after low box office results? That is, do producers tend to make films that are more different from their older films after a film does not perform as well?~~

2. Instead of the above, focus on the following: there is also a file with box office data for some of the films in the file, "box_office_films.csv", also on Canvas. In the columns of this file, "total_box" is the box office revenue earned by that film, "budget" is the dollar amount, where it is available (0 is unavailable), "release coverage" is the total number of screens the film was broadcast on at its peak release, divided by the total number of screens in the U.S., "pindex" is a film identifier consistent with the keywords file, and "title" is the film's title.

   It's possible to use the "keyindex_producers.csv" file and these box office results to figure out if the films that are close or distant from one another in terms of their shared keywords tend to also bring home similar or different box office revenues. Make the distance comparison among firms produced in the focal year, and the two years prior to that year—the original "keyindex_producers.csv" has been manipulated into another file, "film_distance_3year_window.csv", that contains pairwise Jaccard distances between films for this time comparison. Produce a plot showing this relationship. It is likely helpful to go through setting up these kind of grouped lag structures, so you can try to perform the manipulation of the original data on your own. The `dist()` function in the proxy package may be helpful for making the distance calculation if you do this on your own.

3. Large and small film producers compete over how to position their films in topic space and compete for box office revenues against each other. We can define a large producer as one that is at the 75^{th} percentile or above for box office revenues that year. Of the top 250 keywords, in terms of box office revenue earned, that were used over the last 10 years, how many of these appeared primarily (i.e., modally) in films made by large companies? By small companies? By collaborations between the two? Illustrate this on a network plot that links together keywords that appear in the same film. On the nodes, provide a visual indication for which keywords appear more often, and, on the edges, for which keywords appear more often with one another. What does this plot suggest about coproduction relationships between production companies? The "data table testing and usage.R" file may be helpful for setting up the attributes of the network.

Extra challenge problem: Producers can choose to make films that are more specialized, or produce a portfolio that covers a wider range of topics and features. Use the keywords to generate

a yearly measure of feature coverage for each producer for the years the producer has produced at least one film. Take this measure as the average Jaccard distance between each pair of keywords in the set of keywords a producer uses in its films that year. Do producers that make more specialized films, in terms of film topics, experience higher box office revenues per film? What does this suggest about audience tastes for films? Note that this calculation is intensive, and does not have to be performed directly on the matrices. Here, some looping sequencing that recycles the ojbects and refreshes the memory may be helpful. The `t(combn)` family can also be useful.