Exploring Organic Clusters with Clustering Algorithms

Jay Chu
Jericho McLeod
CSI-777 Proposal

The game Magic: The Gathering is a trading card game (TCG) that was patented by Richard Garfield and Wizards of the Coast in 1994 (Citation: USPTO). In this game, players compete against one another using collections of cards of typically at least, but no less than, 60 cards, with an additional collection of 15 cards from which substitutions may be made between matches. Tournament-styled events are held with some frequency and the results of these events are made publicly accessible on websites such as www.mtgtop8.com, along with the list of cards making up each competitor's collection. Tournaments require that each deck have a name, and players using very similar collections of cards will often use the same deck name. Decks which diverge from an existing named collection are sometimes given a new name. This pattern of retaining a deck name for similar, but not necessarily identical, collections of cards, and attaching a new name when the deck has diverged to some degree, is a naturally occurring clustering behavior. How effective is this behavior at clustering collections of cards among publicly reported tournament entrants?

Considering individual cards as dimensions, quantities of cards as dimensional measurements, and card collections as objects, we will algorithmically cluster decks using the K-Means algorithm. Then, we will utilize K-Nearest-Neighbors to assign names to the clusters. The problem space is bound $1<K<N$, where K is the number of clusters we will identify and N is the number of decks in the sample. We will maximize the number of clusters being utilized while minimizing the number of clusters with the same name. The relationship of these two parameters will then be used to describe the precision with which the organically emerging clusters have segregated collections of cards.

The dataset has already been obtained by scraping the previously mentioned tournament results website, www.mtgtop8.com. Approximately 23,000 tournaments are listed on this site, with entrants ranging from 2 to 8 with a currently unknown distribution. We will assume it is normal, though tournaments often have greater than 8 players, meaning a true normal distribution will appear skewed right in our sample. Given that the rules of the game require 60 cards and a normal distribution, we expect a minimum of 6,900,000 cards to be considered. Given the computational expense of the algorithms we intend to employ, we may sample the data at random, study a specific time period, or study some subset of tournament types.

The obtained data was scraped using Python scripts, the BeautifulSoup and HTTPRequests libraries, and AWS instances for parallelization of requests. The data has been stored in JSON format, and subsequent data cleaning, analysis, and reporting will be conducted in Python. Libraries to be used include Scikit-Learn, Pandas, Numpy, and Matplotlib. Given that the data is relatively clean in the scraped format, we expect minimal effort will be spent in data preparation.

We expect broadly defined clusters to be relatively accurate, but with diminishing precision along edge cases. Specifically, the point at which a specific collection leaves one cluster and becomes a member of another may change between different sets of clusters. This is due to expected variances among behaviors for assigning a new name to a collection. However, it is noted that subsetting data temporally or by some other attribute may further impact the veracity of our analysis of clustering behavior, and this will be acknowledged in our final report.

Works Cited:

"United States Patent: 5,662,332." Patent Full Text and Image Database, United States Patent and Trademark Office, 2 Sept. 1997, http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&d=PALL&s1=5662332.PN.