Capstone 2 Proposal

Jessica Curley

Jeopardy

Motivations and Questions

Aspiring contestants on Jeopardy often focus on knowledge in geography, US state capitals and presidents, as well as American pop culture. Are these topics the most critical? I am also curious about how higher valued questions differ from the lower valued questions, and if there is a noticeable difference that an algorithm can identify, or if "difficult" really is subjective.

The questions I seek to answer are:

- Can Jeopardy answer-question pairs* be classified into meta-categories** using a machine learning algorithm? Can this inform focus areas to study.
- Can answer-question pairs be classified into high-value and low-value groups, suggesting there is some non-subjective nature that informs whether a question is higher-value, and therefore assumed "more difficult?"
- Are there topics that follow a measurable trend in occurrences over 35 seasons?

*answer-question pairs refer to the prompts given by the host, and the correct responses (given by the contestant).

** a meta-category would be a grouping of topics, like "history" that would include such categories form the show like "ON THIS DAY: JULY 26" and "US HISTORY".

The Data

This dataset is a .txt file downloaded from kaggle, and has 349,641 rows and 9 columns. Each row contains the information pertaining to a answer-question pair over 35 seasons of Jeopardy, up to the airdate of 7/26/2019.

- The columns are: `'round', 'value', 'daily_double', 'category', 'comments', 'answer', 'question', 'air_date', 'notes'.`
- The `'comments'` and `'notes'` columns do not have values for most rows, and in those instances, I will have to decide how to handle that information. Currently, the empty cells have a ' - ' in places of an NaN values. (`'comments'` are often comments said by Alex Trebek during the show about the category, and `'notes'` indicates whether a question was a Daily Double or part of a special tournament).
- The columns `'round'` and `'value'` have integer values. All other rows have strings. I will have to convert the `'air_date'` to a datetime object for any time series analysis.
- The phrase "What is…" sentence starter has been eliminated already (that wouldn't add much meaning as it is what every contestant says when they respond to a clue), as have any picture, video or sound clues.
- There are punctuation marks and stop-words that I will eliminate in order to do proper EDA and run an analysis on. There is also the challenge of word-number combinations (like "60-minutes") as well as n-grams >1 where multiple tokens represent one linguistic unit (like "North Dakota")

MVP

- Build a pipeline that reads in the file, removes stop-words and punctuation and handles the `'comments'` and `'notes'` columns, as well as any missing values.
- Construct and train an unsupervised machine learning model to classify which meta-category an answer/clue pair belongs in.

MVP + and ++

- Build and train a model that classifies questions as high-value or low-value
- Do a time-series analysis on specific meta-categories or words to predict the change of that topic occurring in an upcoming episode.