Data set source:

This repo utilizes data obtained from the following reddit subsite:https://www.reddit.com/r/datasets/comments/1uyd0t/200000\_jeopardy\_questions\_in\_a\_json\_file/ The original data is a json file containing 216,930 Jeopardy questions, answers and other data obtained by crawling www.j-archive.com. I have formatted these data as a csv and compressed them into a zip for convenience. According to j-archive, the total number of Jeopardy! questions over the show's span (at the time of collection) is 252,583 - so this is approximately 83% of them. In particular, around the last two years of game play are missing.

Data set description:

This data set contains 216,930 rows (34MB) of Jeopardy questions covering decades of the show’s airings. I have included the code to conduct a basic analysis of this text in my repo. This data set contains a mix of alphanumeric values, described by the below schema.

The schema for this data set is below:

|  |  |
| --- | --- |
| COLUMN NAME | DATA REPRESENTED |
| ‘SHOW NUMBER’ | NUMBER OF SHOW WHERE QUESTION APPEARED |
| ‘AIR DATE’ | DATE SHOW AIRED IN YYYY-MM-DD |
| ‘ROUND’ | ROUND OF GAME WHEN QUESTION APPEARED |
| ‘CATEGORY’ | QUESTION CATEGORY |
| ‘VALUE’ | DOLLAR VALUE OF QUESTION |
| ‘QUESTION’ | TEXT OF QUESTION |
| ‘ANSWER’ | TEXT OF ANSWER |

There are a few things to consider when analyzing the text of this file. First, rows are unordered (in terms of date/show airing) which may impact analysis depending on objective. The code I used to analyze text statistics does NOT order the questions by date. Second, not all of the ‘question’ text is actually text. There are a number of questions that utilize pictures or videos to ask the question during the show. These non-text items are typically a-tagged href objects such as hyperlinks. I have not crawled these hyperlinks but the code from earlier convention assignments and regex should handle this issue. Lastly, if you would like to output this data into another format, just be aware that the body of the file is comma delimited but the headers are not.

Descriptive statistics:

These data are presented in a comma-delimited text document. After case-folding splitting by comma, and removing stop words, there were 1,282,185 tokens. Of these tokens, 63,201 are unique. The clean (case-folded/alphanumeric/excluding stop words) data demonstrate a lexical diversity of .049 with an average token length of 5.88. The Jeopardy\_Analysis notebook included in this repository shows my steps for cleaning and analysis.

Interesting question:

I think this data set could be really interesting for analyzing lexical diversity over a time-span and for pattern analysis. It would be interesting to see if Jeopardy uses key words in a way that is consistent with something like Google N-grams and to see if there is a pattern to the way Jeopardy assigns weight (dollar value) to certain categories or terms. It might also be interesting to see which categories receive the most attention and if the lexical diversity of questions becomes more diverse (complex) as the value of the question increases.

Next steps:

My next steps would be:

1 Arrange the questions by date/show number.

2 Use regex and href code to gather the non-text question values (or eliminate these, if a small proportion).

3 Conduct an analysis of lexical diversity by category and value, value by category, and compare answer text n-grams to a measure of common vernacular like N-grams.