Text Analysis of Jeopardy!

Mikayla J. Scott

Syracuse University
IST 687

Project Overview

This data analysis was done on a Jeopardy! dataset of 200,000 thousand questions obtained from https://www.kaggle.com/tunguz/200000-jeopardy-questions.

Target Questions

There were three main target questions that the analysis worked to provide answers to. Those questions are listed below:

1. What Jeopardy! categories are the most likely to occur (sports, literature, pop culture, etc.)?
2. Is there a correlation between words and the value of questions?
3. Is there a way to predict the round of the question based on wording?

Technical Specifications

Throughout the duration of the project there were several applications and languages that contributed to the completion of the analysis. Microsoft Excel was used as a data management platform, comma separated values (CSVs) were generated to import and export data in and out of R. The main language used for this project was R. R was used with a handful of different libraries and packages that helped to create graphs, clean text, and perform simple statistics. Python was used in the very last portion of the project to run a complex algorithm to determine which categories are most likely to occur.

Agile

This project didn't follow the typical agile method because there was only one person working on the project. The project was broken out into phases, those phases are reflective of different stages in the analysis process. Those stages and descriptions can be found outlined in the section below.

Phase 1: Data Collection and Cleaning

Phase one was the primary data collection and cleaning phase. For the data cleaning process there were several things that needed to happen to do text processing. This included making all the questions and categories lowercase. It was critical that this was done early on because later in the project everything from that point on was extremely case sensitive.

Once the category and question columns were all lowercase, then punctuation was removed. Once the words were clear of any punctuation, symbols, and abstract characters. The questions and categories were tokenized, and stop words were removed. After the data went through the processes described above then it was ready for further analysis.

Phase 2: Data Processing and Graph Creation

There were several different sub-phases of Phase 2. Most of the work was done within this Phase. The first initiative of this phase was to determine the words within both the Category and Question dataset that had the highest frequencies overall. Word clouds were generated to display the results.

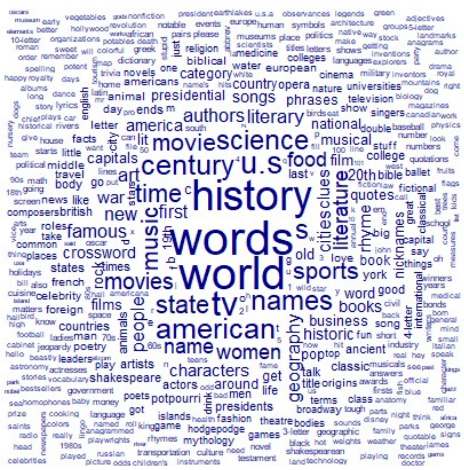

Figure 1: Category Word Cloud

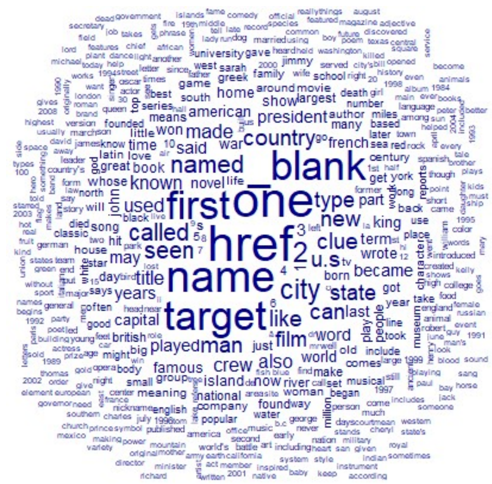

Figure 2: Question Word Cloud

After the word clouds were made, there was a definite difference between the Category word cloud and the Question word cloud that needed to be addressed. After further research and understanding of the data, it became apparent that there were some words included in the question category that did not make any logical sense for a Jeopardy! Question. Words such as href, one, name, target, _blank, were all words that were associated with Questions that had media attached to them as links. This includes but is not limited to pictures, audio files and videos. In effort to show that a large fraction of the data set has pictures, those words were not omitted from the dataset.

Words that were omitted from the datasets and added to the stop word list after interpreting the results can be found in the appendix of this document.

The top 50 words in each dataset were recorded as lists with their corresponding frequencies and stored for future comparison. These lists can be found within the Final Results PowerPoint associated with this project.

Following the most frequent words across all questions and categories, the next sub phase required the data to be grouped by the Round column. There are three different rounds in Jeopardy!, (Jeopardy!, Double Jeopardy!, Final Jeopardy!). This grouping resulted in three data frames for analysis. Working in parallel with the Round grouping, grouping by the Value followed identical procedures. Each question had an associated value with it, in the game of Jeopardy there are eight different values. These values range from $400-$2000, and include a

value called None. It was important to not remove the None value because this was the value associated with the Final Jeopardy! question. To understand the most frequent word in each of the value dataset, the dataset had to be grouped by value, then evaluated separately. This action resulted in eight different data frames.

From those data frames there was a combined total of eleven graphs created to depict the top thirty words with the highest frequencies among those specific grouped by data sets. Examples of two of those graphs can be seen below. To see all the graphs, please refer to the Final Results PowerPoint.

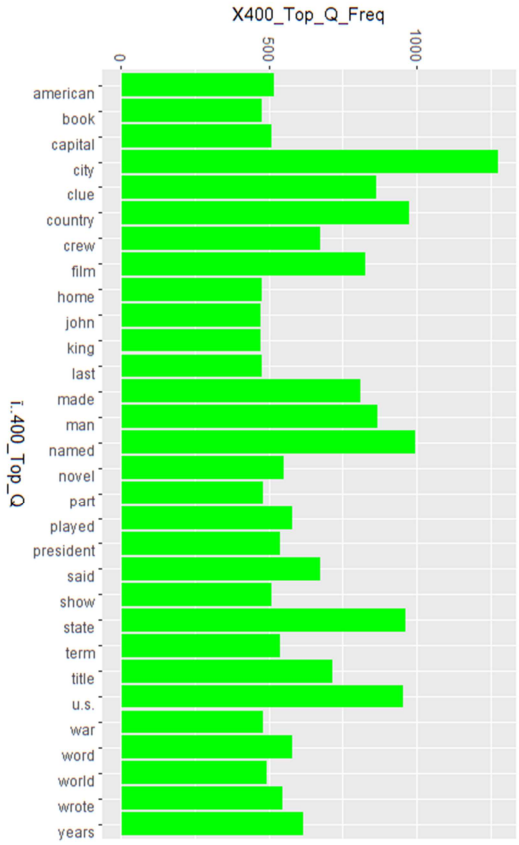

Figure 3: $400 Top Word Freq

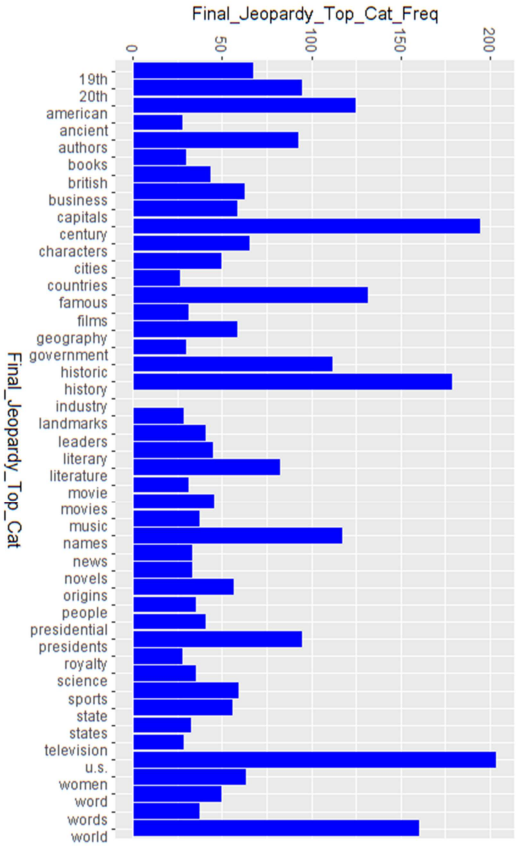

Figure 4: Final Jeopardy Top Word Freq

Phase 3: Associated Word List Creation and Category Classification Algorithm

This phase was primarily aimed at answering the first question. All the work done in this phase was written in Python for efficiency purposes. The code that is discussed in this section can be found in the 687_Jeopardy Jupyter Notebook file.

To answer the first question a specific algorithm was designed to work off associated word lists for each broad category that a question could fall into. These broad categories are as follows: sports, history, pop culture, geography, science/math, literature. The associated word lists were

lists of terms that were closely related to or associated with the broad category. Each associated word list was comprised of anywhere from 250 – 600 words closely related to the subject matter.

The algorithm takes in a list of strings as a parameter. Inside the function there are six different hit counters defined. These counters keep track of when there is a match between a word in a question and in an associated word list. From there the results are stored into a list which is then transformed into a data frame. The algorithm is made to go through the entire dataset and calculate the number of matches per each category associated word list. An example of this data frame can be seen below.

	sports	history	literature	pop culture	geography	science/math
0	0	3	0	2	1	3
1	3	0	0	1	1	1
2	2	2	0	1	1	1
3	0	1	1	1	2	1
4	0	2	0	0	0	0
...	...	...	...	...	...	...
216925	0	0	2	0	0	0
216926	0	1	0	0	0	2
216927	0	0	0	0	0	0
216928	2	2	0	0	2	0
216929	0	2	0	1	1	0

Figure 5: Matching Hits per Broad Category

To determine which category had the overall most matches, the columns were totaled, and their results recorded. Those results can be seen below:

```
Total Combined Question Concentrations
Sports: 129207
History: 142075
Literature: 58316
Pop Culture: 62649
Geography: 94365
Science and Math: 77793
```

Figure 6: Summed Columns

With the completion of this function and the results, Phase 3 was completed.

Phase 4: Results

The results phase was purely dedicated to interpreting the results of the analysis and answering the questions that were asked at the beginning of the project. The answers to those questions can be found in this section:

1. What Jeopardy! categories are the most likely to occur (sports, literature, pop culture, etc.)?

    a. History, Sports and Geography are the most likely to occur as broad categories. This result is based on the output and summation of the topic_classification function.

2. Is there a correlation between double jeopardy questions?
    a. There is no correlation between words and the value of questions, in fact; a majority of the top 30 words in each of the values closely match those of the other value questions.
3. Is there a way to predict the round of the question based on wording?
    a. There is not a way with the current analysis to determine a strong predictive trend for the round in which a question might fall. More in depth technical analysis needs to be done.

In addition to answering the target questions, phase 4 was also dedicated to document, and PowerPoint drafting and finalization.

Conclusion

This project was based on the premise of answering three primary questions. To answer those questions a series of data manipulations, natural language processing and calculations were performed on it. These actions listed previously were carried out within three phases with the last phase (Phase 4) dedicated to interpreting and translating the results.

There are many avenues left to go down to determine concrete answers to the questions that were asked at the beginning of the project. The avenue that was focused in this analysis was word frequency, and what that looked like among different groups of data. To have a comprehensive understanding of the correlations between all the datasets there needs to be a more in-depth analysis on the remaining attributes.

Appendix

seen	last	since
called	became	may
like	1	since
2	one	get
new	last	now
also	called	just
known	whose	great
used	4	make
can	best	
3	may	
many		