Classification Task

The classification task chosen to implement is a classifier for predicting categories. The models implemented and used for classification are variations of Multinomial Naive Bayes.

Curation of Dataset

In the game of Jeopardy, a category is unique to each game. However over time, as observed from this data which spans from the year of 1984 to 2011, there have been repetition of several categories and to evaluate the classifier, data points from 10 most frequent categories were collected. They are listed as followed:

American History	Before & After	College & Uni.	History	Literature	Potpourri	Science	Sports	Word Origins	World History
418	547	351	349	496	401	519	342	371	377

Preprocessing

For the preprocessing step, the capitalization of words from answers and questions was lowered and punctuation was removed. While the words were not lemmatized, any plural terms or letters occurring after apostrophes were removed. Only textual features, which were columns of answers and questions, were used for predicting categories. To identify the optimal combination of features to choose the best result, the model was trained on a feature set where the answer was concatenated to the question string in the beginning, as well another set where it was concatenated at the ending, and a feature set of only the question.

Examples of feature used for training:

Case 1 Question: ‘built in 312 bc to link rome the south of italy its still in use today'

Case 2 Answer + Question: ‘the apian way ‘ + ‘ built in 312 bc to link rome the south of italy its still in use today'

Case 3 Question + Answer: ‘built in 312 bc to link rome the south of italy its still in use today' + ‘ the apian way’

The sentences were vectorized to train the Naïve Bayes classifier as 1-gram, 2-grams, 1,2 – grams. All these combined with inclusion and removal of stop-words.

Results

A dummy classifier used for predictions as a baseline control outputted an F1-score of 0.107. This is expected as there are 10 classes with relatively similar sample sizes. The results of the Multinomial Naïve Bayes are listed below. The alpha hyperparameter was tuned between the values of 0.1 – 2, with a step-size of 0.1. The following are the results:

1. *Feature set: Only Questions (Optimal alpha = 0.6)*

N-grams, StopWords	1	2	1,2
True	0.602	0.359	0.625
False	0.627	0.555	**0.652**

2. *Feature set: Concatenation of Question and then Answer (Optimal alpha = 0.5)*

N-grams, StopWords	1	2	1,2
True	0.618	0.549	0.622
False	0.649	0.456	**0.659**

3. *Feature set: Concatenation of Answer and then Question (Optimal alpha = 0.6)*

N-grams, StopWords	1	2	1,2
True	0.618	0.445	0.624
False	0.650	0.545	**0.662**

The 1,2-gram model performs the best while 2-gram model performs the worst. While the number of features increase when using a 2-gram model, we also reduce the amount of data that is available. A 2-gram model increases predictive power when combination of words are more identifiable with one group and are common across its samples. This shows that in Jeopardy bigrams within categories are rare and as such questions within categories are also distinct. Furthermore the questions for each category selected are on average 15 words, which explains that each sample is less likely to contain a bigram that is present in its category. The results do show that while 1-gram is more informative than 2-grams, the bigram features do provide more context to improve predictive power.

The most successful combination of feature selection for this classification task is concatenation of answers and then questions. 'Questions' presented to players in Jeopardy are usually statements that describe the answer. Syntactically then it makes sense for the answer to be placed before the question so that it can be inferred as a noun being described. Identifying a correct syntactic flow is important because, as results indicate, existing bigrams can improve the classifiers performance in combination with 1-gram model and by improving syntactic flow, we increase the chance of acquiring bigrams that can be shared by other samples of the same category. In essence, answers preceding their questions make relatively sensible statements.

We also observe that removal of stop-words results in a drastic decrease in performance. As aforementioned, each question is 15 words which results in certain stop-words carrying heavy weight for particular categories. For instance, time indicative stop-words such as 'in' are most occurring in American History, History, and World History. The term 'we' also occurs frequently in American History, and predictably 'us' occurs the most in American History.

Conclusion

The optimal classifier is a (1,2) – gram model with no removal of stop words. Deducing from the results, I further arranged my features such that they were 'answer' + 'question' + 'answer'. This increased the F1-score from 0.662 to 0.664.