Jules Capaldo, Peter Christie, Megan Crawford, Clarinsa Djaja

25 April 2022

Unstructured Data Analytics

***Jeopardy!* Question Bank Analysis:**

Using Unsupervised Methods to Explore *Jeopardy!*'s Unstructured Data

Introduction and Motivation

For our final project, we utilized the *Jeopardy!* game show question bank to employ unsupervised learning techniques to improve player strategy-making and gameplay. In order to accomplish this, we examined most common questions and answers, which would help players make a better guess when in a pinch. We also made connections between the categories of the most common answers to add context. TF-IDF scores grouped by the year of the air date allowed us to examine the five most important terms for both questions and answers, giving the player clues about what they should study when preparing to appear on *Jeopardy*. Finally, we considered limitations and next steps for developing gameplay strategy, helping players maximize their success on the show.

Dataset and Data Preparation

Beginning with our data cleaning process, we used the *Jeopardy!* dataset, which had 216,931 rows and 8 columns: index, Category, Value, Question, Answer, Round, Show Number, and Air Date. We realized that using all of these rows would be quite time consuming for analysis, so we removed all of the rows that had NA values, that had <ref in the question, that had "None" as the value, and that were part of the Final Jeopardy rounds (since they had varying dollar amounts associated with them and because we wanted to focus our analysis on the Jeopardy and Double Jeopardy rounds). We also converted the Air Date column to a DateTime data type. Due to the size of the dataset, we split it into two datasets so that they could be uploaded to GitHub and later concatenated them into the clean jeopardy dataframe in Python. After removing the index column, our clean dataframe had 202,535 rows and 7 categorical columns. 102,301 rows belonged to the Jeopardy round, whereas 100,234 belonged to the Double Jeopardy round.

Data Exploration

I. Top Categories

Once we cleaned our dataset, we found the top categories across both rounds, as well as in the Jeopardy round and in the Double Jeopardy round so that we could learn which categories are recurring, allowing a player to focus their studies on these (and similar) categories. Overall, the top three categories were “Before & After”, “Literature”, and “Science”; these were the same top three categories for the Double Jeopardy round. On the other hand, for the Jeopardy round, the top three categories were “Sports”, “Potpourri”, and “Stupid Answers”. It is important to note that categories were more likely to repeat themselves in the Double Jeopardy rounds, where the top three categories appeared 448, 377, and 264 times, respectively. See Appendix Figure 1a, 1b, and 1c for the top ten categories in each round. It is interesting to note the differences in the types of questions which appear in the Jeopardy! round versus the Double Jeopardy! round. The most common categories in the Jeopardy! round tend to be associated more with common knowledge and are more accessible to the standard viewer: Television, Sports, U.S. cities, and Potpourri (a random assortment of questions). On the other hand, the common categories in the Double Jeopardy round tend to be oriented towards someone who has a higher level of education or cultural refinement: Literature, Opera, Ballet, and Shakespeare.

II. Filtering on Most Common Categories

After determining the top ten categories for each round, we found the running total for all categories and plotted them, see Appendix Figure 2. Based on this plot, we can see that relatively few categories account for a disproportionate number of questions. For example, 50,000 questions are taken from only about 575 different categories. Given that there were over 25,000 unique categories, we reduced our dataset to only the categories that appear the most frequently, choosing the top 100 categories. The 100th most common category contained 133 questions, indicating that all of the top 100 categories have appeared at least 27 times over the course of the series since there are five questions per category. The reduced dataset had 22,638 rows, making it more manageable to work with for strategy development.

Insights

I. Most Common Terms in Common Categories and Answers

Based on our reduced dataset, we created a corpus based on the Question column and calculated the term frequencies for questions in the top 100 categories. After realizing that calculating term frequency on our reduced dataframe produced a large, sparse matrix, we removed English stopwords (as well as " `` ", "--", " ' '", "...", " 's ") to create a new corpus that we tokenized. Based on the distribution plot, found in Appendix Figure 3, the most common terms in the questions of the top categories included “name”, “first”, “city”, “one”, “country”, and “capital”.

We can draw three conclusions about *Jeopardy!* questions based on these frequencies: questions tend to be identification (based on “name”), they tend to be about the “first” of something, or they are commonly location questions. Knowing these question types would be beneficial for contestants preparing to compete on Jeopardy because it would inform them to focus their studies on historical events and on geographical locations.

After checking the term frequencies for Questions, we did the same for Answers. Returning to the entire data, all of the top 20 answers are proper noun locations. The top 3 are China, Australia, and Japan. Appendix Figure 4 contains the list of the top 20. The first non-location to appear was George Washington. Predominantly, these are countries, but cities and a few states are present in the top twenty answers. While one may assume that the locations are strictly geography questions, we found that they can also be related to history, culture, literature, and others. Additionally, we checked the most common categories associated with three common answers: China, Cleopatra, and Chicago. Results can be seen in Appendix Figure 5. This revealed that category titles have become more niche/specific and also are more likely to feature puns, which has become a development in more recent years of the show. The dispersion plot for the most common categories also reveals this conclusion: that early shows tended to reuse the same categories fairly frequently, as can be seen Appendix Figure 6, while the more recent ones had more variety or different names for specified categories. One hypothesis for this shift away from these originally common categories is the change towards similar categories which go by a slightly different name. For example, the history category has appeared less frequently in more recent shows. However, other history adjacent categories or categories which would be classified under the topic of history have become more popular.

II. TF-IDF Analysis

Finally, we used TF-IDF to determine the most important terms appearing in both Questions and Answers. For our first attempt, we used each question as the documents and had the corpus be all of the questions from all of the shows. However, we found it more beneficial to instead group all of the questions from a given year and use the year’s questions as a document. Then, all of the questions from each show were combined into one cell per year. The same process was used for the Answer TF-IDF calculations.

The TF-IDF scores of Questions indicated that with time, the most important words have shifted to the previous five years; for example, in 2008, the top five TF-IDF terms were 2008, 2007, 2006, 2005, and 2004. We recommend, then, knowing general knowledge of current events and culture from the last five years while studying for *Jeopardy!* to ensure the greatest success for answering these questions. Additionally, fill-in-the-blank questions have become increasingly important based on TF-IDF. The TF-IDF scores of Answers, on the other hand, show that towards later years, there was a shift to more notable figures and names present among the most important terms. For example, in 2011, both “Barack” and “Obama” were both included in the top 5. See Appendix Figures 7a and 7b for the top 5 most important words for Questions and Answers from selected years.

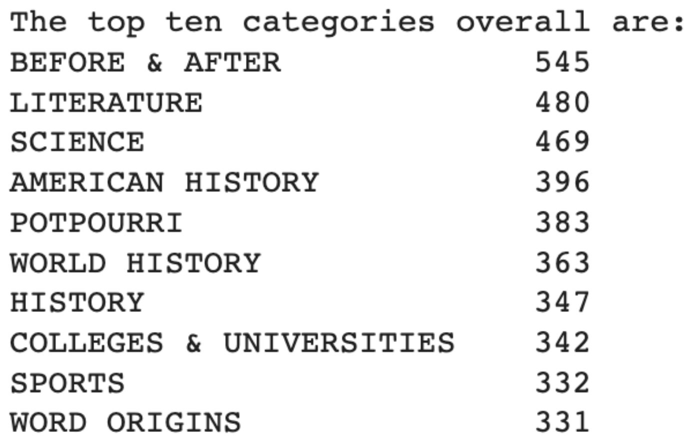
Limitations and Improvements

When completing our unsupervised analysis for *Jeopardy!* strategy development, we faced some limitations. First, we only examined trends in the Jeopardy! and Double Jeopardy! rounds; perhaps we would have encountered different conclusions if we had included Final Jeopardy! in our analysis. Second, we solely focused on unsupervised learning techniques but did not pursue semantic similarity. We could have gained additional strategic insights had we chosen to work with word embeddings to try and cluster questions, categories, or answers. Finally, due to the number of and the variety of questions which could be asked on the show, it was hard to identify important terms since many terms only appear once.

Based on these limitations, some next steps could be conducting supervised learning, such as predicting the dollar value of a question based on its difficulty. We also could use our insights to create a rules-based classifier to classify the questions. Another potential next step is to utilize topic modeling to identify terms which commonly show up together to see common question sentence patterns.

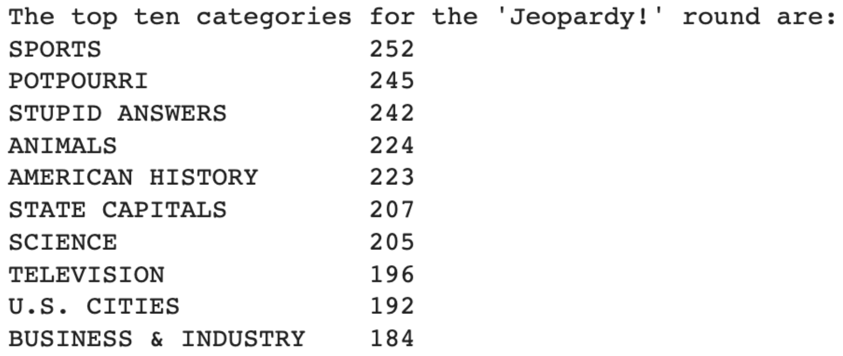
APPENDIX

Figure 1a.



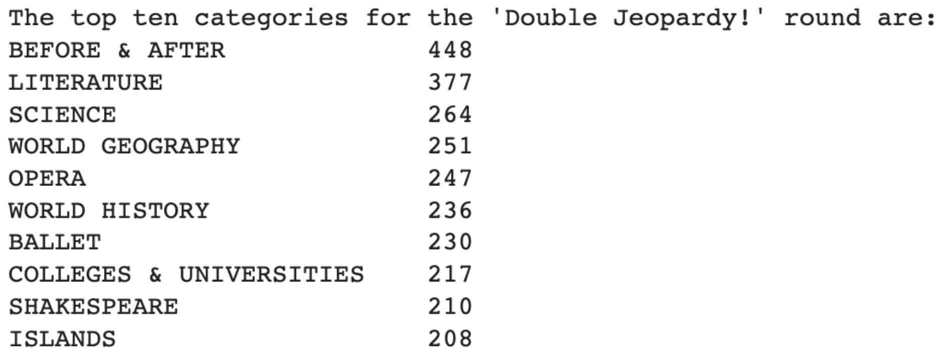
```
The top ten categories overall are:
BEFORE & AFTER              545
LITERATURE                  480
SCIENCE                     469
AMERICAN HISTORY            396
POTPOURRI                   383
WORLD HISTORY               363
HISTORY                     347
COLLEGES & UNIVERSITIES     342
SPORTS                      332
WORD ORIGINS                331
```

Figure 1b.



```
The top ten categories for the 'Jeopardy!' round are:
SPORTS                  252
POTPOURRI               245
STUPID ANSWERS          242
ANIMALS                 224
AMERICAN HISTORY        223
STATE CAPITALS          207
SCIENCE                 205
TELEVISION              196
U.S. CITIES             192
BUSINESS & INDUSTRY     184
```

Figure 1c.



```
The top ten categories for the 'Double Jeopardy!' round are:
BEFORE & AFTER              448
LITERATURE                  377
SCIENCE                     264
WORLD GEOGRAPHY             251
OPERA                       247
WORLD HISTORY               236
BALLET                      230
COLLEGES & UNIVERSITIES     217
SHAKESPEARE                 210
ISLANDS                     208
```

Figure 2.

Note 1: X-axis = number of unique categories, Y-axis = cumulative total of number of questions

Note 2: Area A → Category appeared multiple times

Area B → Category appeared once, asked five questions

Area C → Category appeared once, asked fewer than five questions.

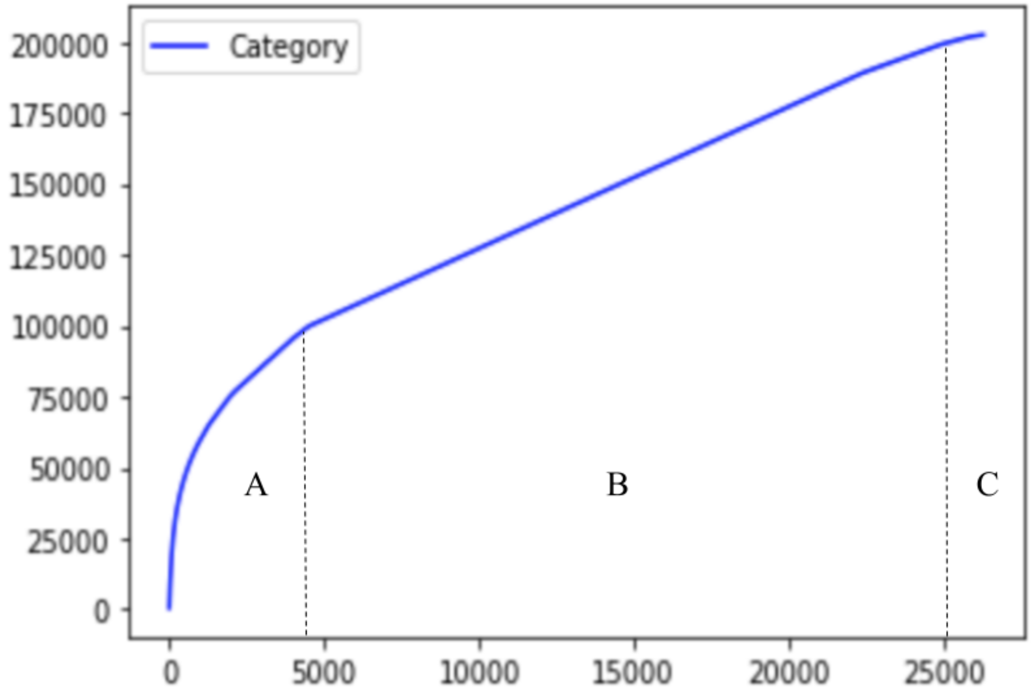


Figure 3. Most common terms in questions from the top 100 categories.

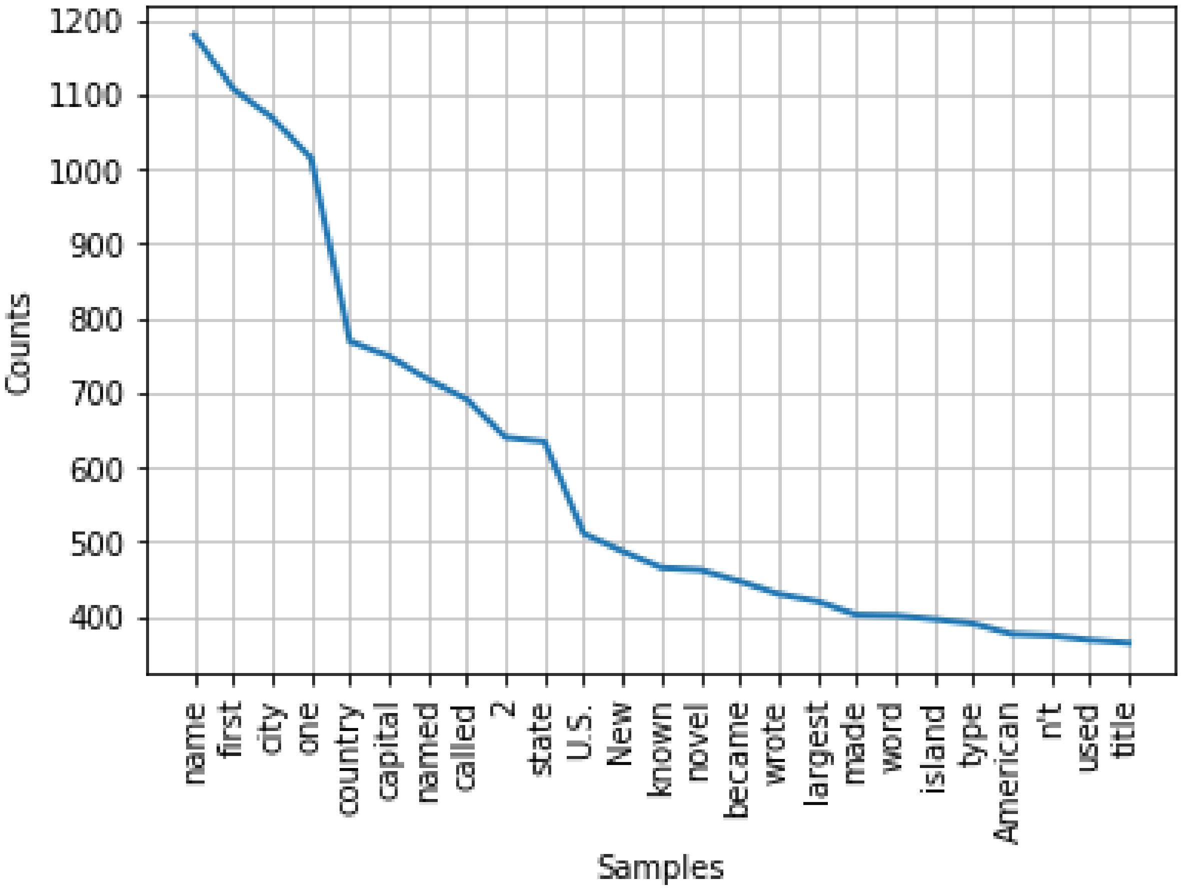
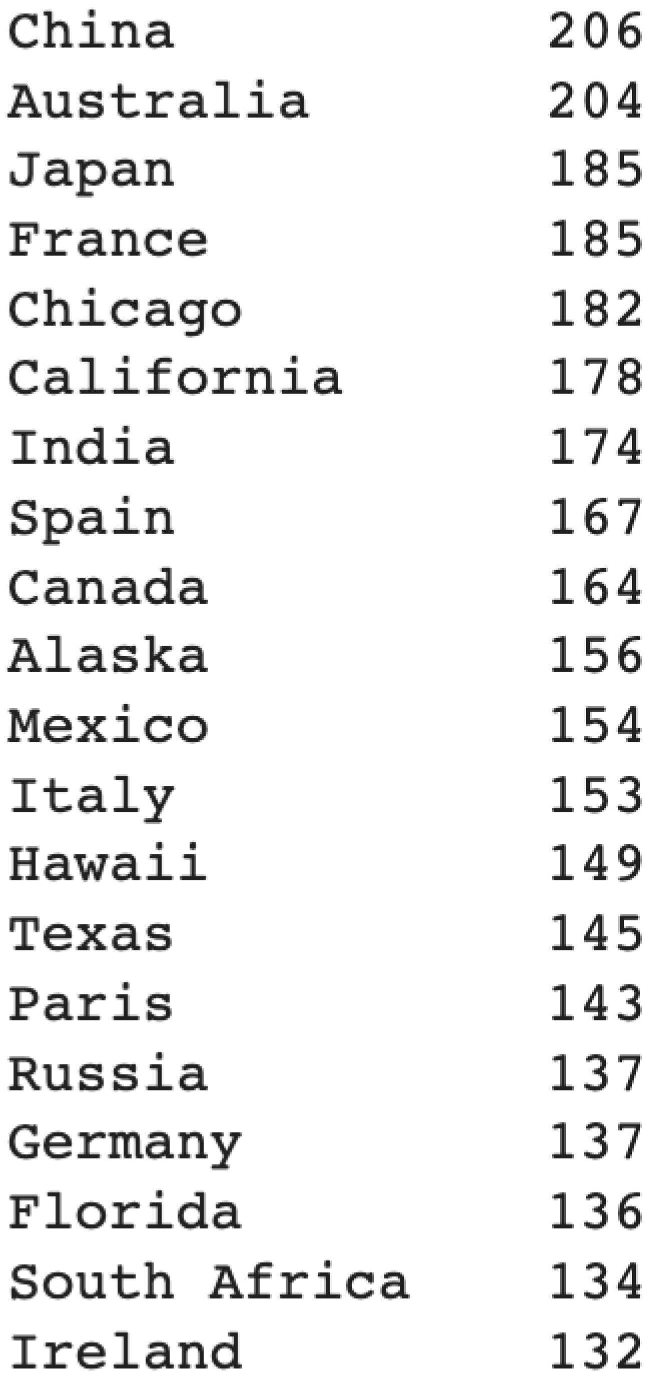
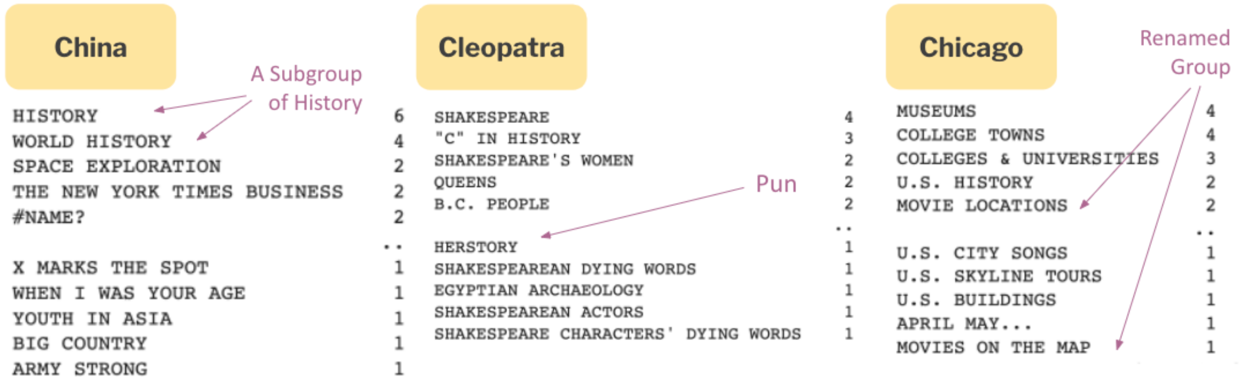


Figure 4. Most Common Answers across the entire data set.



```
China               206
Australia           204
Japan               185
France              185
Chicago             182
California          178
India               174
Spain               167
Canada              164
Alaska              156
Mexico              154
Italy               153
Hawaii              149
Texas               145
Paris               143
Russia              137
Germany             137
Florida             136
South Africa        134
Ireland             132
```

Figure 5. Categories associated with a specific answer.



China

```
HISTORY                          6
WORLD HISTORY                    4
SPACE EXPLORATION                2
THE NEW YORK TIMES BUSINESS      2
#NAME?                           2
                                ..
X MARKS THE SPOT                 1
WHEN I WAS YOUR AGE              1
YOUTH IN ASIA                    1
BIG COUNTRY                      1
ARMY STRONG                      1
```

Cleopatra

```
SHAKESPEARE                          4
"C" IN HISTORY                       3
SHAKESPEARE'S WOMEN                  2
QUEENS                               2
B.C. PEOPLE                          2
                                    ..
HERSTORY                             1
SHAKESPEAREAN DYING WORDS            1
EGYPTIAN ARCHAEOLOGY                 1
SHAKESPEAREAN ACTORS                 1
SHAKESPEARE CHARACTERS' DYING WORDS  1
```

Chicago

```
MUSEUMS                     4
COLLEGE TOWNS               4
COLLEGES & UNIVERSITIES     3
U.S. HISTORY                2
MOVIE LOCATIONS             2
                           ..
U.S. CITY SONGS             1
U.S. SKYLINE TOURS          1
U.S. BUILDINGS              1
APRIL MAY...                1
MOVIES ON THE MAP           1
```

Figure 6.

Note: The X-axis roughly corresponds to the show's air date.

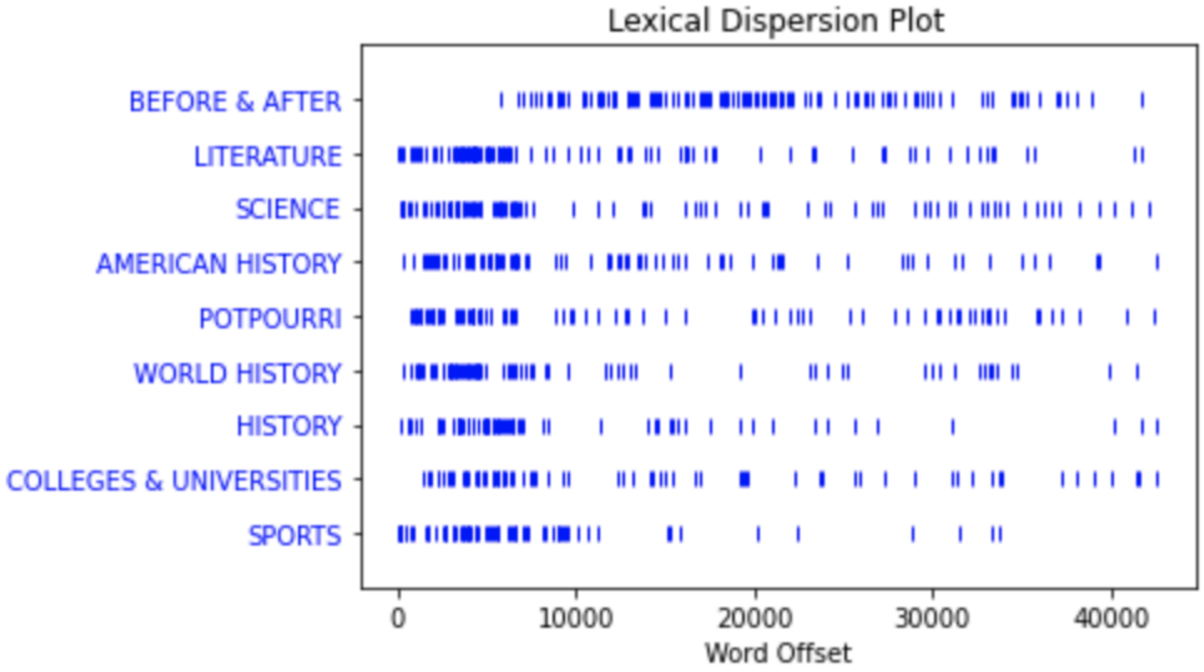
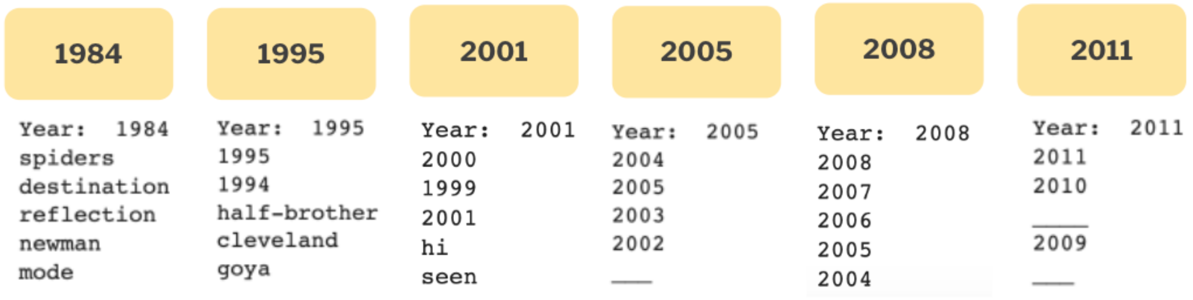


Figure 7a.



1984

```
Year:  1984
spiders
destination
reflection
newman
mode
```

1995

```
Year:  1995
1995
1994
half-brother
cleveland
goya
```

2001

```
Year:  2001
2000
1999
2001
hi
seen
```

2005

```
Year:  2005
2004
2005
2003
2002
___
```

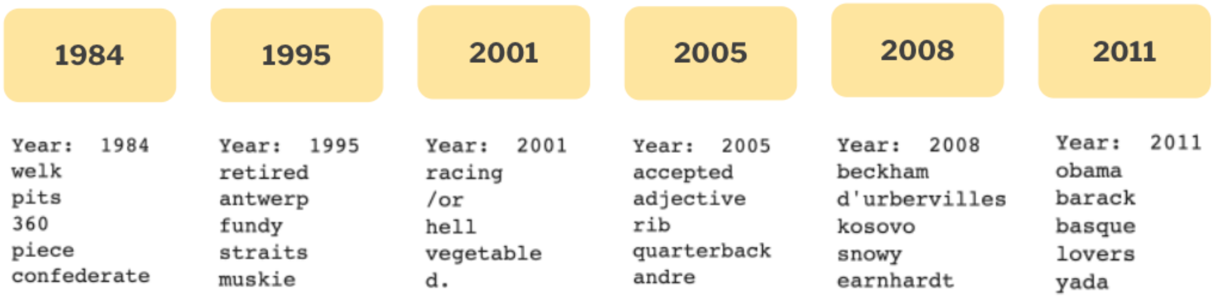
2008

```
Year:  2008
2008
2007
2006
2005
2004
```

2011

```
Year:  2011
2011
2010
____
2009
___
```

Figure 7b.



1984

```
Year:  1984
welk
pits
360
piece
confederate
```

1995

```
Year:  1995
retired
antwerp
fundy
straits
muskie
```

2001

```
Year:  2001
racing
/or
hell
vegetable
d.
```

2005

```
Year:  2005
accepted
adjective
rib
quarterback
andre
```

2008

```
Year:  2008
beckham
d'urbervilles
kosovo
snowy
earnhardt
```

2011

```
Year:  2011
obama
barack
basque
lovers
yada
```