Text Analysis of the State of the Union Addresses

Data Visualization and Design | CUNY Graduate Center | Summer 2019

Goals

- Explore ways to visualize Text
- Explore ways to visualize metadata about a text

Data

This Tableau workbook that uses this dataset

Dataset

Dataset Small

Originally from here

Premise

We have a few questions:

1. Have there been any trends in the length, lexical density, or significance of State of the Union Addresses?
2. Did George W. bush or Barack Obama have an overall different tone to their addresses?

Getting Started

There are really 2 datasets here: the metrics, and the first 500 words of each address. These are on 2 sheets. The metrics give an overall, distant description of the datasets. All of these metrics could reasonably be created in Excel or a text editor. More advanced analysis techniques are beyond the scope of this course, but check out my text analytics lab for more.

Merge Data

Merge fields & Pivot

1. Join Metrics sheet to words sheet
2. Select from w1 to w499 (shift+click)
3. Select 'Pivot' from the drop down menu
4. Create Year from the first 4 characters of Date

Trends: A Line Chart

1. Drag CALC: Year to the Measures Pane and then to the Columns
2. Drag Wordcount to the Rows
3. From the Analytics pane, drag Trend Line into the view, and then drop it on the Polynomial model type.
4. What has happened over time? Have speeches gotten longer or shorter?
5. Annotate some key speeches. Right click on the data point
6. Select 'Annotate Point' and write a short description in the box.
7. You can drag and drop this box - make sure no lines cross and it doesn't obscure any points or lines on your chart.

Stacked bar Chart of Total words and Unique Words

We want to make a bar chart that visually illustrates the difference between the total number of words and the unique number of words that a president uses. This offers a glimpse into the linguistic complexity of the speech. You might expect a shorter speech to have a greater ratio of unique words to total words if it were to have the same amount of content.

1. Drag President and CALC: Year to Columns.
2. Drag Measure Names to Color on the Marks card.
3. On Color, right-click Measure Names, select Filter, select the check boxes for 'Unique Words' and 'Word Count', and then click OK.
4. From the Measures pane, drag Measure Values to Rows.
5. On the Marks card, change the mark type from Automatic to Bar. For more information, see Bar Mark.
6. Clean up your axis labels, etc.
7. Sort President by Median of Year
8. Change the Alias of the Legend (Right Click)

Sentences

We've also broken our speeches down by sentence. If we take sentences as a proxy for written or formal variety versus more colloquial forms, this may give us some insight into the tone of a president's speeches.

1. Drag President and CALC: Year to the Columns
2. Drag SentenceCount to the Rows - change the Aggregation to Median to give an overall picture of each president.
3. You may want to create a new field that is the number of words divided by the number of sentences to give a more accurate picture of the speaking style of a given president.

Word Clouds - a first look into the topics

1. Pick 2 presidents to compare - you will make different word clouds for each, but it's too much to use all the presidents.
2. Drag Single Words to Text on the Marks card.
3. Drag Single Words to Size on the Marks card.
4. Right-click Single Words on the Size card and select Measure > Count.
5. If necessary, change the Mark type from Automatic to Text.
6. Select just the top 100 by dragging 'Single Words' to Filter and select 'Top' and Choose 7. To remove Stopwords, Command+Click on all the words that seem meaningless

No Stops By Year

1. Drag President & YEAR to Columns
2. Change the Marks to Circle
3. From the Analytics menu, select