Text Analytics using Tableau

The following tutorial will show you how to perform text analysis in Tableau 8.2. To get started, you will need the following:

- Tableau version 8.2
- Data: Political Speeches.xlsx

Part 1: Loading the Dataset

First, load the dataset into Tableau.

- Connect to Data > In a file > Microsoft Excel > Select the Political Speeches.xlsx
- You may see the dialogue window that says “Errors occurred while reloading the data source.” If so, click OK. Then, click Go to Worksheet. This should bring up the Tableau a new worksheet. We can now begin.

Part 2: Learning about the Data: Word Length and Phrase Length

First, let’s learn a little bit about the data. The data is text from speeches by the four major presidential candidates (Hillary Clinton, John McCain, Barack Obama, and Mitt Romney) in 2008. The data has column names as follows:

- **Candidate**: The candidate
- **Party**: Republican or Democrat
- **Phrase**: These are two word phrases in the speech
- **Position**: This shows where the phrases and words occur in the speech. For example, for John McCain, the first phrase is “thank you,” which occurs at position 1, or the beginning of the speech.
- **Word**: These are the actual single words in the speech (not two word phrases)
- **Phrase_length**: This is how long the phrases are measured by characters. For example, the phrase “thank you” is 9 characters long
- **Word_length**: this is how long the corresponding words are measured by characters. For example, the word “thank” is 5 characters long.

I. Word Length and Word Length Filter

Let’s begin by working with the data and displaying the words in the speech by candidate, and filtering the speech by word length.

a. First Drag Position from Measures to Dimensions
b. Drag Position into Rows
c. Drag Candidate into Columns

Note: the tutorial data came from the example posted here - http://www.tableausoftware.com/blog/text-analysis-election08-stump-speeches
d. Drag Word into the text button

e. Drag Word Length into Filters > Select Maximum > Select ‘Range of Values’ > Apply > Click OK. This will allow you to filter the data by the length of the word.

f. Now let’s add a Quick Filter.
   i. Go to Analysis > Quick Filter > select Max Word Length. A Max(Word Length) filter should appear on the right side. Click and Drag this under the Marks Card.
   ii. Slide the filter to only show words with 5 or more characters:

   iii. Notice now that you can quickly filter the words in each speech by word length.

iv. Save this Worksheet as “Speech Word Filter by Candidate”

v. Your worksheet should look like this:
II. **Phrase Length and Phrase Length Filter**

Let’s continue by working with the data and displaying the phrases in the speech by candidate, and filtering the speech by phrase length.

a. Right click on the “Speech Word Filter by Candidate” worksheet. Click “Duplicate Sheet”
b. Rename this sheet to “Speech Phrase Filter by Candidate”
c. Remove the Word from the Marks Card.
d. Remove Max(Word Length) from the Filters.
e. Drag Phrase into the Text button on the Marks Card
f. Drag Phrase Length into Filters.
   i. Set it up just like last time, as a range, and add it as a Quick Filter.
   ii. Drag the Max(Phrase Length) Quick Filter under the Marks Card

g. Filter the Max(Phrase Length) so only show phrases from 12 to 27 characters in length.
h. Please notice now that you can now filter the two word phrases in the speeches by their length in characters.
i. Your Report should look like this:

Note: the tutorial data came from the example posted here - http://www.tableausoftware.com/blog/text-analysis-election08-stump-speeches
Part 3: Top Phrases (Bar Chart)

I. Top Phrases by Party

II. Now that we have an understanding of the data we are working with, let’s do some analytics on the text. Let’s begin with the top phrases by political party.

a. Drag Phrase into Rows
   a. Select Filter and then Add all members.

b. Drag Phrase into Columns. Click on the down arrow that appears when you hover over Phrase (still in Columns) and change the Measure of Phrase to Count.

c. Change the graph to order the phrases from most to least. (Hover over the worksheet data and a little icon appears. Click it and this will rearrange the data.)

d. Drag Phrase into Label.

e. Notice that the most popular phrase is “of the” followed by “in the” “to the,” etc. This doesn’t really help much. Remember, SO WHAT?? Let’s filter this to show some more meaningful phrases.

f. Drag Party into Filters. Check Democrat. Click Apply.

g. Drag Phrase Length into Filters.
   i. Set it up just like last time, as a range, and add it as a Quick Filter.
   ii. Drag the Max(Phrase Length) Quick Filter under the Marks Card

h. Change the phrase length to show only those that are 12 to 26 characters long.

i. Now, we get some more meaningful phrases, and we can tell a better story. Notice that we can see the most used two word phrase by Democrats is “the American”, followed by “American people” and “new leadership.”

j. Save this chart as “Top Phrases by Party”

k. Your worksheet should look like this:
II. Top Phrases in All Speeches
Let’s get a high-level look at the top phrases in all speeches.

a. Right Click on the Top Phrases by Party worksheet, click Duplicate Sheet
b. Rename this worksheet to “Top Phrases in All Speeches”
c. Remove the Party field from Filters
d. Remove Phrase from the Marks Card
e. Drag Candidate into Color
f. We can now see a quick look at the top phrases used by Clinton, McCain, Obama, and Romney
g. Your worksheet should look like this:

III. Top Phrases by Candidate
Let’s go into some top phrases by specific candidates.

a. Right Click on the “Top Phrases in All Speeches” worksheet
b. Click Duplicate Worksheet
c. Name this worksheet as “Top Phrases: Obama (Bar)”
d. Drag Candidate into Filters
   a. Filter only by Obama. Click Apply and Ok.
e. Remove Candidate from the Color in the Marks Card
f. Add Phrase to the Labels
g. We can now see a quick look at the top phrases used by Obama
h. Your worksheet should now look like this:

Note: the tutorial data came from the example posted here - http://www.tableausoftware.com/blog/text-analysis-election08-stump-speeches
**Part 4: Top Phrases (Word Cloud)**

**I. Top Phrases: Obama (Text)**

Now let’s create a word cloud using the top phrases by President Obama.

a. Drag Position into Rows
   i. Change the measure of Position to Average so it looks like AVG(Position).
   ii. Also add a quick table calculation to AVG(Position). Specifically, add the ‘percent of total’ quick table calculation.

b. Drag Word Length into Columns. Change the measure to Average.

c. In the drop-down menu of the Marks Card, change it to say “Text”

d. Drag Phrase into the Text button

e. Drag Phrase Length into Color. Change it to AVG(Phrase Length).
   i. Change the Color to Orange-White-Blue Diverging. Also click “Reversed” Click Apply and click OK.

f. Now the graph looks a little cluttered. Let’s filter some things.

g. Drag Candidate into Filters. Filter it by Obama.

h. Drag Phrase Length into Filters.
   i. Set it up just like last time, as a range, and add it as a Quick Filter.
   ii. Drag the Max(Phrase Length) Quick Filter under the Marks Card
   iii. Filter everything so that it shows the phrase lengths from 14 to 26.
i. Now we still have a lot of phrases showing. Let’s filter it down by a certain condition. Let’s assume that the most emphasized and meaningful phrases are those said more than once. So, let’s filter the data to show only the phrases said more than one time.
   i. Drag Phrase into Filters.
   ii. Double click Phrase.
   iii. Under the General tab, go down to the Summary area. You should see a field that says Condition: **None**. Click on the word **None**. This will bring up a Filter window.
   iv. In the Filter window, select By field:
   v. Using the drop-down arrow, select Phrase. Select Count.
   vi. Select the greater than symbol and put a 1 in the text box.
   vii. Click Apply and Click OK.
   viii. You should now see the phrases said more than once.

j. Let’s make the sizes of the words bigger.
   i. Drag Phrase Length into the Size Button.
   ii. Change the measure to Average.
   iii. Now use the Size button to Resize the words.

k. We now have a quasi-world cloud for the top phrases. Let’s format the cloud so it looks cleaner.

l. Right click on one axis. De-select “Show Header”. Do the same for the other axis.

m. Let’s remove the Grid Lines.
   i. Right click on the graph. Select Format.
   ii. Click on the tiny paintbrush icon under the Format window.
   iii. Next to Grid Lines, change the drop down to say None.

n. Success!

o. Rename your worksheet “Top Phrases: Obama (Text)”

p. Your worksheet should look something like this:
II. Top Phrases: All other Candidates

Now that you have the skills to make a word cloud for Obama, make three additional ones for the remaining candidates (Romney, Clinton, and McCain). Save these as “Top Phrases: CANDIDATE_NAME (Text)”

Part 5: Dashboard & Write-Up (On Your Own)

1. Using what we have just learned, create an interactive dashboard using the worksheets we just made. Save the Dashboard as “Text Analysis_Candidate Speeches”. Please include at least three graphs. Please incorporate the filters and actions as well.

2. Write-up a one page, double-spaced document about the scenario bellow:

You are a newly hired analyst for a political consulting firm. Your boss knows that you learned some text analytics in college. So, your boss wants you to write him a one-page document that describes what text analytics is and what it can be used for. You decide to use the Text Analysis_Candidate Speeches dashboard that you created to show your boss an example of text analytics. In the report, please describe text analytics, and how your dashboard could help gain insights into future presidential speeches for potential presidential candidates. Basically, why would you even bother with text analytics? What could it do in the realm of political speeches? How could knowing the most used keywords, etc. help future candidates?

Save your file as Text_Analysis_Write_Up.docx.

Please upload everything to the course website.