Blog for 11-10-17 (Friday)

Tidy text

Gutenbergr

Bing

Sentiments

Continuing our work with tidy text, sentiment words

We are again working with gutenbergr

After loading all the packages we need we’ll download the novel Dracula from gutenbergr the download # is 345

Then we will get the Dracula words as a data frame Dracula\_words using the unnest tokens from tidy text

Then we’ll use bing to get the sentiment as either positive or negative. We’ll call this variable bing.

Then we need to use inner join to join together the dracua words & the sentiments from bing.

Now we’ll get rid of the column with the gutenbergr id number because we don’t need that. The ‘NULL’ deletes the column.

Now we’ll get all the positive sentiment words summarize & count them. Since there are lots of words we want the top ten. We played around with the numbers so we’ll use 66 so that we get the top ten words. We’ll save this into a variable Dracula\_pos.

Then we’ll factor out the Dracula positive words so we can graph them and save it as dracula\_pos$word.

Now we’ll use this to make a bar plot with ggplot.

We’ll repeat the process for the Dracula negative words. In order to get the top 10, we need the filter count to be 49. In addition the word ‘Miss” is included & since this is a ‘title’ we want to exclude it.

OOPS

We “lost” the sentiment attached to the words so we want to retain the sentiment. After summarize count we’ll add sentiment.

Now when we look at the variable-data frame Dracula\_pos we see the sentiment is still attached.

Again, we’ll repeat the process for the negative sentiments.

Now we don’t like that we have to hard code in a number 66 or 49 to get the top 10 words. We are going to use ‘top\_n(10,wt=count)’ instead. (we aren’t sure what the wt stands for (maybe weight?)

Now when we look at Dracula pos we have the same top 10 words.

Again, we’ll repeat for negative words.

Now, we’ll re-run our positive words plot.

We can put the two data frames positive & negative on top of one another.

We’ll use rbind to put the two together in a variable data frame called dracula\_comp.

Let’s look at our variable dracula\_comp. We can see the positive & negative sentiments are still attached.

Let’s plot these:

The two plots are stacked one on top of the other. It would be better if we could see these side-by-side. We can get ggplot to 'apply' something within the plot (no change to data) use facet-wrap (the sentiment column tells us what group it is in).

The graphs are now side by side, but still one on top of the other split on the y axis. We want to “free” the y axis.

Now our plots are side by side -Success!

Let’s go out to the web & get some HTML color codes. The website <http://htmlcolorcodes.com/color-picker/> is supposed to be interactive so you can choose the color & color code you need. We’ll add orange to the graphs.

It would be better if each graph was its own color so we’ll make the fill the sentiment.

Now let’s choose the colors we want by adding scale fill manual to the end.

So now we have orange & black. Now let’s add the outline colors with color

This may be our last text mining project