Text Analysis of the Annual Report Narratives

This quick analysis shows some of the potential ways we can perform quantitative analysis on the text in the annual narrative reports. I think this type of analysis could be useful as a summary product of the reports and as a tool to assist in the review of the reports (i.e., additional analysis to consider while reading a particular state’s report). It may be especially interesting in future years to see how the reports change over time. The type of text analysis I have explored so far or plan to explore include:

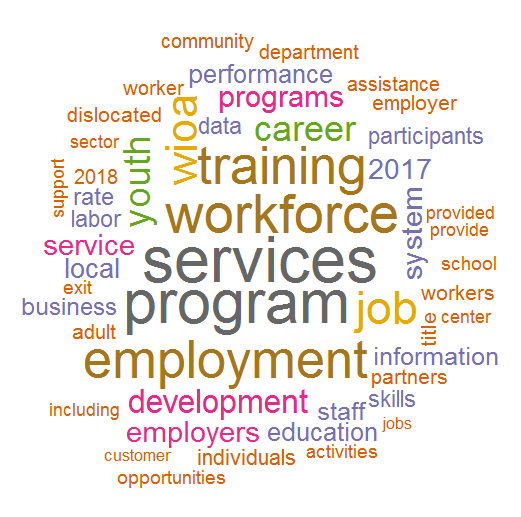
* [Word cloud report and cloud comparisons](#_Report_Word_Cloud)
* [Unique words of states](#_Unique_Words_of)
* [Sentiment of reports](#_Sentiment_in_Reports)

# Word Cloud of the Reports

In this analysis, I created a word cloud to examine the words that are used the most in the reports. I created two word clouds for all the states combined (i.e., the states we have so far): one without the stop words (e.g., the, this, and, etc.) and one without the stop words and other very common words (e.g., program, services, etc.). The second word cloud is an attempt to see if there is any interesting words that pop out with the common/expected words removed. These can’t be finalized until we receive all the reports.

In addition, I also created comparison word clouds that see how a state compares to nation as a whole. This can be done for all states in both comparing against all words and against the cloud with the common words removed. However, it also can’t be completed until we receive reports from all the states, so I just give a couple example word clouds below.

**All States and All Words**



**All States without Most Common Words**



# Unique Words of States

For this analysis I plan seeing what words are uniquely used by states in their reports. In addition to removing stop words (e.g., the, and, or, etc.), I will removed the words that were common amongst the states (e.g., program, services, etc.). I will likely do multiple versions with one with it drilled down until there is only one unique word per state and one where it isn’t quite so drilled down so there may be multiple states that have the same word. The later one would be a way to potentially find aspects were some states are similar. For a better data visual, I plan to try to display the words in a map of the USA.

# Sentiment in Reports

In this analysis I was interested in seeing the sentiment expressed in the reports. It is basically quantifying the number of positive and negative words used. The first chart, *Sentiment of Narrative Reports,* shows the counts of the positive and negative words and has a line that represents the overall sentiment (i.e., the net of the positive and negative).

Since there is almost certainly a correlation between the number of sentiment words and the length of the report and since all states have an overall sentiment score that is positive (i.e., the line), I also created a *percent positive* variable to account for differences that are simply due to the size of the report. This second chart, *Scatterplot: Sentiment Score Vs Percent Positive,* attempts to examine the meaning of the sentiment score after accounting for the size of the report. The scatterplot shows that there is a positive correlation between *sentiment score* and *percent positive*, so states with a higher sentiment score tend to have a higher percentage of positive words in their reports.

